TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Mensch-Maschine-Kommunikation

Robust Methods for Content Analysis of Auditory Scenes

Jürgen Thomas Geiger

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. sc. techn. (ETH Zürich) G. Kramer
Prüfer der Dissertation:	1. Univ.-Prof. Dr.-Ing. habil. G. Rigoll
	2. Univ.-Prof. Dr.-Ing. W. Hemmert

Die Dissertation wurde am 24.04.2014 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 20.10.2014 angenommen.

Acknowledgment

This thesis would not have been possible without the support of many people. First of all, I would like to thank my supervisor Prof. Gerhard Rigoll for giving me the opportunity to work at the Institute for Human-Machine Communication at TUM and for the advices and support he gave me during my time at the institute. Furthermore, I would like to thank Prof. Werner Hemmert for doing the second review of this thesis. Over the years, fruitful discussions with colleagues contributed to the success of my work. Here I would like to mention Dr. Martin Hofmann, Prof. Björn Schuller, Felix Weninger, Florian Eyben, Dr. Martin Wöllmer, Erik Marchi, Zixing Zhang, Jun Deng, Dr. Tobias Rehrl, Dr. Alexander Bannat, Dr. Jürgen Blume, Prof. Frank Wallhoff, Nicolas Lehment, Dr. Moritz Kaiser, Daniel Merget, Philipp Tiefenbacher, Simon Schenk, Prof. Joachim Schenk, Maximilian Kneißl, and Mohamed Anouar Lakhal. Moreover, I would like to thank researchers from outside the institute: Dr. Martin Rothbucher, Dr. Ravichander Vipperla, Dr. Simon Bozonnet, Dr. Nicholas Evans, and Dr. Jort Gemmeke. For the technical and administrative support at the institute, I would like to thank Peter Brand, Heiner Hundhammer, Martina Römpp, Gertrud Günther, and Melitta Schubert. In addition, I am indebted to my family and my girlfriend for the encouragement during the last years. This work was partially supported by the project AAL-2009-2-049 "Adaptable Ambient Living Assistant" (ALIAS) co-funded by the European Commission and the German Federal Ministry of Education (BMBF) in the Ambient Assisted Living (AAL) programme and by the DFG excellence initiative research cluster "Cognition for Technical Systems" (CoTeSys).

Abstract

The increasing progress of audio analysis methods opens possibilities for more new applications. At the same time, recent improvements in these methods bring the established approaches constantly closer to their performance limits, which are defined by disturbing factors such as overlapping speech or noise and reverberation. This thesis presents progress in new possibilities and addressing disturbing factors, first, by proposing ideas for a system for the classification of acoustic scenes and a method for acoustic gait-based person identification. Both of them are two relatively new audio recognition tasks. Furthermore, improvements for two established methods (speaker diarization and robust speech recognition) are presented. To improve speaker diarization, different approaches to detect overlapping speech are proposed. To increase the robustness of a speech recognition system against noise and reverberation, an approach using memory-enhanced acoustic modelling is employed. Together, the proposed modules represent a complete system for auditory scene analysis. Starting from a coarse classification of the scene as a whole, persons can be identified using their step sounds or voice, followed by a transcription of the spoken contents. Experimental evaluations using publicly available databases or within public research challenges demonstrate the efficiency of the proposed methods.

Kurzfassung

Der zunehmende Fortschritt von Methoden zur Audioanalyse eröffnet immer mehr Möglichkeiten für neue Anwendungen. Jedoch bringen die jüngsten Verbesserungen die etablierten Methoden immer näher an ihre Grenzen, welche durch Störfaktoren wie beispielsweise überlappende Sprache oder Störgeräusche und Nachhall gegeben sind. Diese Arbeit präsentiert Fortschritte bei neuartigen Anwendungen und bei der Behandlung der erwähnten Störfaktoren. Zuerst werden ein System zur Klassifikation von akustischen Szenen und ein Verfahren zur akustischen Gang-basierten Erkennung von Personen präsentiert. Beides sind neuartige Audio-Erkennungsaufgaben. Anschließend werden Verbesserungen auf den zwei etablierten Arbeitsgebieten Speaker Diarization und robuste Spracherkennung vorgestellt. Um die Speaker Diarization zu verbessern, werden verschiedene Ansätze zur Detektion von überlappender Sprache vorgeschlagen. Um die Robustheit eines Spracherkennungssystems gegenüber Störgeräuschen und Nachhall zu erhöhen, wird ein Ansatz zur akustischen Modellierung mit erweiterter Kontextmodellierung verwendet. Gemeinsam stellen die vorgeschlagenen Module ein komplettes System zur akustischen Szenenanalyse dar: Ausgehend von einer groben Klassifikation der Szene als Ganzes können Personen mittels ihrer Schrittgeräusche oder der Stimme identifiziert werden, gefolgt von einer Transkription der gesprochenen Inhalte. Experimentelle Auswertungen mit Hilfe von frei verfügbaren Datenbanken oder im Rahmen öffentlicher Forschungswettbewerbe demonstrieren die Wirksamkeit der vorgeschlagenen Methoden.

Contents

1

Introduction

Auditory scene analysis characterises the capabilities of humans to understand complex auditory scenes [23]. In particular, the auditory scene analysis model describes how the human brain transforms the sensory perception of sounds into a mental representation. Different sounds are grouped (along time and frequency) to auditory streams which represent the sound sources. A well-known example which illustrates this method is the cocktail party effect [31]. This phenomenon describes the human ability to focus the attention on one sound source in a mixture of different sounds, for example one voice among several voices and other sounds. The field of computational auditory scene analysis, also referred to as *machine listening*, tries to reproduce the same capabilities with computers [228]. Originally, the focus of such systems is on audio source separation. When binaural recordings are available, audio localisation is also an important feature. In general, such systems perform low-level audio processing. After different audio sources have been detected and segmented, the goal is to classify or recognise these sources on a higher semantic level. This is where pattern recognition methods are deployed.

The present study addresses specific challenges in the field of audio pattern recognition. Methods are presented which aim at recognising the contents of auditory scenes. The focus is not on low-level processing such as source separation, but more on the classification of different sound sources using pattern recognition methods.

In general, there are many fields of audio recognition that have been the subject of research in the last decades. A large portion of audio analysis research is dedicated to human speech, and one of the oldest but still most prominent fields is automatic speech recognition (ASR) [176]. This field progressed over the last decades from speaker-dependent recognition of isolated syllables to highly performant recognition of conversational speech in adverse environments [178, 175, 115, 138]. Later on, researchers tried to identify the speaker [40, 26] or to further analyse the speaker's state or trait (e. g. emotion, age or health state), a field known as computational paralinguistics [197]. Another prominent field of research in audio recognition is automatic

music transcription [126], which consists of the recognition of notes (pitch) [24] and tempo [39].

Acoustic scene recognition is a newer research direction [166], and the detection and classification of acoustic events can also be included in this field [217]. Such systems have the goal of detecting and recognising any kind of sound (acoustic event) in an audio recording or of classifying an acoustic scene as a whole, consisting of different sound sources.

A related field is audio diarization, which aims at annotating an audio recording (e. g. of a meeting) with respect to the occurring sound sources [183]. Here, a special case is speaker diarization, where the focus is on human speech and the goal is to detect and identify all occurring speakers in an unsupervised manner [5]. Speaker diarization can also be seen as a variant of speaker recognition.

In all of the mentioned fields of audio recognition, pattern recognition methods are applied to a detection or classification problem. Each of these fields of research has different practical applications. Any computer system (e. g. mobile phones) can make use of ASR technology as an input device. Services for automated phone answering systems such as telephone banking use speaker recognition technology for access control. Home automation can employ many of the techniques to equip a *smart home* with the capabilities of analysing the ongoing activity.

Audio recognition methods are also used in robotics. For example, service robots with a multimodal dialogue system can improve their understanding of the environment with techniques for speech and audio recognition [10]. One example is the robotic companion Adaptable Ambient Living Assistant (ALIAS) [179] [54] [56], for which some of the techniques presented in this thesis have been developed. Especially a robotic system relies on techniques such as acoustic scene recognition and audio diarization, when it is deployed in a changing environment, where every day new sound sources are occurring.

An exemplary acoustic scene is displayed in Figure 1.1. In this scene, different sound sources contribute to the auditory sensation, e. g. different human activities or sounds from machines or cars. When these sounds are emitted simultaneously, they produce a complex mixture. The task of a listener is to recognise the individual sound sources, consisting of a temporal and spatial localisation as well as classification. Following these first processing steps, higher-level semantic meaning can be derived, e. g. about the activity in the scene or the spoken content. The focus of the present study is mainly on classifying different sound sources occurring in such a scene.

1.1 Objectives

While all of the discussed methods for audio recognition have made progress over the years and, in part, matured to a level which allows the application in practical

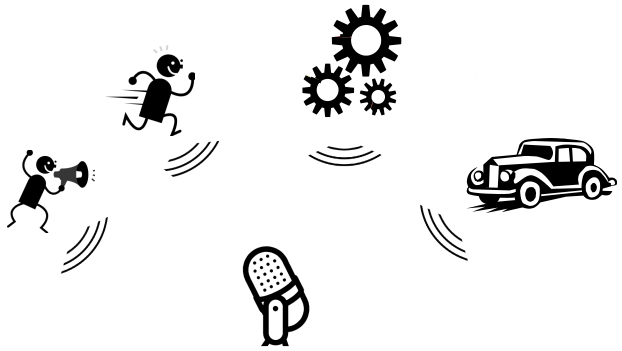


Figure 1.1: Sketch of an exemplary acoustic scene with different sound sources (a person speaking or running, a machine, or a car) and a microphone.

systems, many open problems are remaining. The goal of this thesis is to address these problems and to propose new solutions or improvements to existing solutions.

Since the field of audio signal processing is so extensive, first, a distinction is made, in order to define what is excluded from this thesis and was not focus of the underlying work. Most notably, all of the introduced methods work with monaural audio signals. Thus, methods for binaural audio source separation or localisation are not employed in this study. Although these two fields are the original dedicated problems of computational auditory scene analysis, the focus of the present study is on pattern recognition aspects. Different open problems of audio recognition tasks are addressed in this thesis.

(1) Recognition of acoustic scenes and events is a relatively new field of research. The goal of this field is to recognise acoustic scenes as a whole, where an acoustic scene is normally represented by a somewhat longer audio recording (several seconds to minutes). A system for acoustic scene classification should categorise the recording into one of a set of general classes, describing the location or the activity that is taking place. Possible classes are, for example, *office*, *street*, or *supermarket*. Such acoustic scenes are characterised by many diverse single acoustic events, potentially overlapping in time and frequency. This is why acoustic scene recognition is related to the detection and classification of acoustic events, where similar methodologies are applied. Since these fields are relatively new, there are no established methods. The first efforts consisted of adopting methods from other audio processing tasks to these problems. Strong methods are needed that are specifically tailored to these problems and which are capable of achieving high recognition rates. The goal of the present study is to investigate the application of different audio features and classifiers for the problem of acoustic scene recognition.

Furthermore, an open problem in the field of classification of individual acoustic events is how to learn new sound classes. The goal of this thesis is to propose a

system for acoustic event classification in an indoor environment, and to develop a method for learning new classes of acoustic events.

(2) Acoustic gait-based person identification denotes the idea of recognising humans by their step sounds. This is another new emerging topic in audio recognition. Human gait is considered as unique and it can therefore be used for biometric verification. The step sounds are a representation of human gait and convey characteristic information about the subject. First studies in this direction were already performed, in which the focus was mostly on the classification of presegmented footsteps from a small number of subjects. Methods are required to recognise a large number of subjects reliably, in order to establish this field of research. The objective of this thesis is to consider different methods for the problem of acoustic gait-based person identification. The suitability of static and dynamic classifiers, using different audio features, should be investigated. Presumably, dynamic methods that model the cyclic sequence of sounds during walking are expected to deliver better results.

(3) Speaker diarization methods have advanced to a level where small confounding factors lead to the majority of system errors. One of these problems is overlapping speech (i. e. when two or more speakers are speaking at the same time). Overlapping speech leads to impure speaker models and missed speaker errors. In order to further improve diarization systems, robust methods for the detection of overlapping speech are desired. To overcome this problem, one goal of this thesis is to develop methods for overlap detection.

Different approaches exploiting different characteristics of overlapping speech are thinkable. For example, source separation methods could be used to detect the contribution of individual speakers to a potentially overlapping speech signal. Established audio features that are used in other audio recognition tasks should be investigated for their suitability for overlap detection. It is also of interest whether methods that go beyond the pure acoustic signal are capable of detecting overlap, for example exploiting lexical information or temporal context. The influence of successful overlap detection on a speaker diarization system should be evaluated through methods for overlap handling.

Furthermore, most state-of-the-art diarization systems work offline (i. e. the whole audio recording has to be present), while some applications might require solutions capable of online processing. Therefore, another objective of this thesis is to propose a method for online speaker diarization.

(4) Robust speech recognition describes the problem of ASR in adverse environments. Such environments are for example characterised by additive noise or reverberation and can pose a major problem for conventional ASR systems, deteriorating recognition rates. Approaches for robustness can be categorised into three groups: robust features, speech/feature enhancement, and robust acoustic modelling.

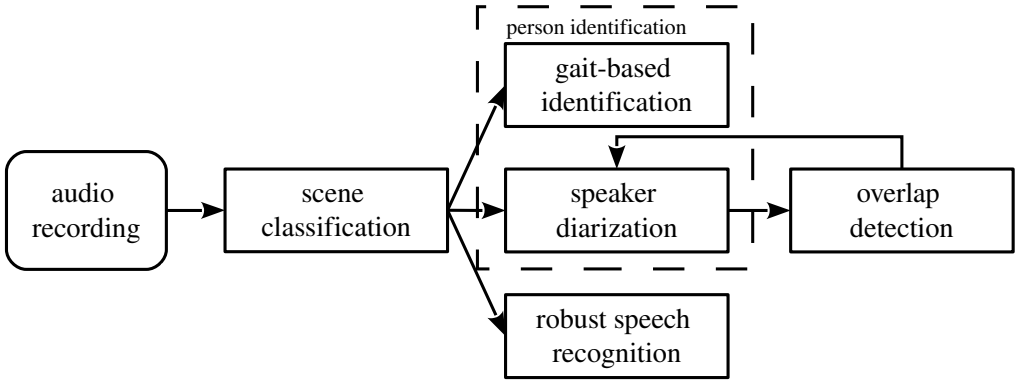


Figure 1.2: Structure of a system for audio analysis that includes all techniques proposed in this thesis.

The present study investigates methods from the third category, employing acoustic models that are robust towards the mentioned confounding factors. The main objective is to investigate how an acoustic model that is able to efficiently exploit context information performs in different configurations and conditions. This includes a comparison to and a combination with different other state-of-the-art acoustic models. The recognition performance should be studied for continuous speech recognition in highly non-stationary noise and reverberation.

1.2 Structure of this Thesis

This study presents contributions to each of the four mentioned topics, which directly leads to the structure of the thesis. Figure 1.2 shows the setup of a system that combines all techniques that are used in this thesis.

Such a system could be deployed in different application scenarios, for example for a robotic assistant or in a smart home. The first step of this system is to classify the acoustic scene as a whole, before forwarding the distinct acoustic events to the other processing modules. These modules contain a component for person identification, analysing either step sounds or human speech. Speech overlap detection helps to improve the speaker diarization module. Finally, a speech recognition module produces a transcription of the spoken content.

Chapter 2 proposes a system for acoustic scene classification. This system is capable of classifying recordings of acoustic scenes as a whole. A large set of audio features is considered, and classification is performed on time windows with a length of several seconds. Different classifiers are compared and the performance of different audio features is analysed in detail [57]. Experiments were performed within a public research challenge, where the proposed system was ranked in the upper third of all participants [58]. Additionally, a method is proposed for learning

new acoustic events within a system for acoustic event classification [70]. This method prevails over conventional learning methods.

Chapter 3 addresses the problem of acoustic gait-based person identification. Static classification is the first investigated method, employing a large candidate feature set. The system performance is analysed and improved using a method for feature analysis and selection [66]. In addition, it is shown how this system can improve a video-based system through multimodal fusion. Second, a dynamic classifier is employed. This system detects and models separate steps of the subjects. Modifications are proposed which improve the identification performance. All experiments in this chapter were performed with a large publicly available database that contains recordings of persons walking in a corridor.

Chapter 4 covers speaker diarization. Different methods are proposed to overcome the problem of overlapping speech in a diarization system. Several systems for overlap detection are presented and evaluated. Among them are methods employing source separation, different audio features, lexical information, and a classifier that exploits temporal context [68, 71, 63, 64]. These approaches improve the overlap detection performance compared to other state-of-the-art methods, which in turn leads to enhancements of a diarization system. Furthermore, a method for online speaker diarization is proposed [62]. All experiments were performed with publicly available databases.

Chapter 5 describes efforts towards robust speech recognition. The influence of highly non-stationary noise and high reverberation is tackled with a classifier that exploits long-range temporal context. This classifier represents a robust acoustic model, and it can be combined with conventional methods. Recognition systems based on this framework achieve large performance improvements in highly non-stationary noise [59, 69] and in reverberant environments [72]. Experiments were performed within the framework of public research challenges, where competitive results were obtained.

Chapter 6 summarises the results achieved in this thesis and draws some conclusions. In addition, some ideas for future research are presented.

2

Recognition of Acoustic Scenes and Events

One of the first tasks of a system that analyses the complete content of an acoustic scene is to classify the scene as a whole, i. e. to select a label from a coarse and diverse set of different types of acoustic scenes. This helps to provide a first insight into the contents of the scene. Additionally, a pre-classification of the sound sources into different types of acoustic events can be helpful. In further processing steps, different models suited for specific scenes can be applied, depending on the contents.

This chapter addresses these two topics: classification of acoustic scenes as a whole and classification of individual acoustic events. First, a system is proposed in Section 2.1 which classifies longer recordings of different acoustic scenes into one of a set of classes. Second, an acoustic event classification system is presented in Section 2.2. There, the specific problem of learning new acoustic events is addressed. The chapter concludes with a short summary in Section 2.3. Section 2.1 is based on the results published in [57, 58] and Section 2.2 builds upon the results published in [70].

2.1 Acoustic Scene Classification

This section describes a system for acoustic scene classification using large-scale audio feature extraction. From highly variable recordings, a large number of spectral, cepstral, energy, and voicing-related audio features are extracted. Using a sliding window approach, classification is performed on short windows. A support vector machine (SVM) classifier is used to recognise these short segments, and a majority voting scheme is employed to get a decision for longer recordings. SVMs are compared with a nearest neighbour classifier and an approach called latent perceptual indexing, where SVMs achieve the best results. A feature analysis using the t-statistic shows that mainly Mel spectra are among the most relevant features.

The proposed system was evaluated in the scene classification track of the challenge on detection and classification of acoustic scenes and events [75]. On the official development set of the challenge, an accuracy of 73 % was achieved in the best system configuration, while an accuracy of 69 % was obtained on the non-public test set, representing state-of-the-art results.

2.1.1 Introduction

Recognising the acoustic background is known as *acoustic scene classification* and belongs to the field of *computational auditory scene analysis* [228]. Typically, several different (overlapping) sound sources contribute to the scene, making it a complex interaction of different acoustic events. The goal of an appropriate system is to analyse an audio recording of such a scene with respect to the composition of acoustic events and to assign a coarse label from a set of classes. Distinguishing characteristics for such classes are outdoor vs. indoor, quiet vs. lively, or different sound sources (e. g. vehicles, machines, humans, animals), resulting in classes such as *street*, *restaurant*, or *office*.

There is little prior work in this field of research. Generally, the approaches found in the literature can be categorised into two different groups. The first method is to regard the whole acoustic scene as one object and to model the long-term distribution of spectral patterns, regardless of the exact order. One example is the 'bag-of-frames' approach that usually works with Mel-frequency cepstral coefficients (MFCCs) as audio features and uses Gaussian mixture models (GMMs) to model the distributions [7]. The second group of methods work with higher-level representations of the acoustic events happening at the scene. One previous study on acoustic scene classification, which can be counted among the second group of methods, investigated the application of various spectral, energy, and voicing-related features in combination with neural networks [142]. Effective features were selected and, together with this choice of classifier, were successful in discriminating among five types of television programs. In [166], a system for acoustic scene recognition is described and evaluated. This system uses a nearest neighbour classifier and various time-domain, frequency-domain, and cepstral audio features. The system's ability to recognise 17 different acoustic scenes has almost reached the level of human performance. The system for acoustic scene recognition described in [116] uses SVMs embedded in a hierarchical or parallel framework. High accuracy results were obtained for the task of classifying audio clips into one of five classes. One possible direct application of acoustic scene classification is a system as described in [193], which recognises cyclist's routes using scene recognition techniques.

In the scene classification track of the aforementioned challenge on detection and classification of acoustic scenes and events, different systems for acoustic scene recognition were evaluated and compared. The employed corpus (divided into a development set and a non-public test set) is categorised into ten different classes of

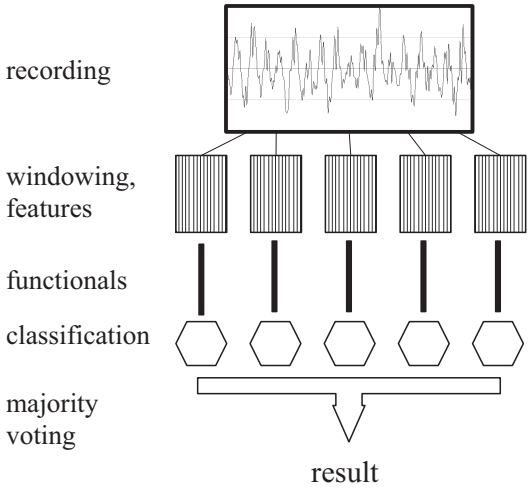


Figure 2.1: System overview for acoustic scene classification.

acoustic scenes. The present thesis describes a method for acoustic scene classification that is a contribution to this challenge. A sliding window approach is used to obtain statistical functionals of the low-level features for short segments of several seconds. SVMs are used for classification of these short segments, and a majority voting scheme is employed to get a decision for the whole recording. Furthermore, the proposed approach is compared to a method based on latent perceptual indexing [120] and to a nearest neighbour classifier as it was used in [166].

2.1.2 System Overview

Figure 2.1 shows the workflow of the proposed system for acoustic scene classification. An audio recording is first segmented into (potentially overlapping) windows with a length of several seconds. In this way, the system is designed to capture individual acoustic events in the scenes. Then, dynamic features are extracted on the frame level and summarised into one feature vector per window through the application of functionals. Subsequently, classification is performed with a static classifier on the window-level and the final result for the whole recording is obtained through majority voting. The system components are explained in more detail in the following.

2.1.3 Feature Extraction

Prior to feature extraction, the stereo recordings are mixed down to mono. This down-mixing causes a loss of information, while, on the other hand, it simulates realistic, simple conditions with devices like mobile phones that are equipped with only one microphone. For feature extraction, the open-source toolkit openSMILE [45]

Table 2.1: 57 cepstral, spectral, energy, and voicing-related acoustic low-level descriptors (LLDs) from the official openSMILE `emo_large.conf` feature set that are employed for acoustic scene classification.

Cepstral features (13)
MFCC 0 – 12
Spectral features (35)
Mel spectrum bins 1 – 26 (0 – 8 kHz), zero crossing rate, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, spectral flux, spectral centroid, relative position of spectral maximum and minimum
Energy features (6)
logarithmic energy, energy in bands from 0 – 250 Hz, 0 – 650 Hz, 250 – 650 Hz, 1 – 4 kHz, 3010 – 9123 Hz
Voicing-related features (3)
F0 (with cepstrum auto-correlation), F0 envelope, probability of voicing

is employed. Since the recordings of the acoustic scenes contain a high number of different sound sources of different nature, a large set of different audio features is extracted in order to acquire as much relevant information as possible. The employed feature set is the official openSMILE `emo_large.conf` feature set that is provided with the toolkit. This feature set was originally designed for speech processing, but it fits general audio analysis owing to its many spectral and further descriptors. In previous studies on the classification of acoustic scenes and events, such as [166, 215], similar audio features were used. All low-level descriptors (LLDs) are listed in Table 2.1, and each of them is extracted every 10 ms from 25 ms frames. The employed features can be grouped into cepstral, spectral, energy-related, and voicing features. In addition to MFCCs and Mel spectra, spectral roll-off points and other spectral features contribute to a comprehensive description of the spectrum. Furthermore, a number of energy-related features are computed. Since the recordings in the test data contain a considerable portion of speech and other voiced or harmonic sounds, the set also includes a small selection of voicing-related features. From the 57 low-level descriptors, 39 functionals are computed after adding delta and delta-delta coefficients, resulting in a total number of 6 669 features. The functionals are listed in Table 2.2 and include values such as mean, standard deviation, percentiles and quartiles, linear regression functionals, or local minima/maxima related functionals. In combination with the LLDs, the comprehensive set of functionals embodies an extensive characterisation of the audio signal. The whole set of 6 669 features

Table 2.2: 39 functionals in the official openSMILE `emo_large.conf` feature set.

Statistical functionals (21)
(positive) arithmetic mean, root quadratic mean, root quadratic, geometric, positive arithmetic mean of non-zero values, number of non-zero values , zero crossing rate, centroid, variance, standard deviation, skewness, kurtosis, quartiles and inter-quartile ranges, 95 %, 99 % percentile
Regression functionals (9)
linear regression slope, offset, corresponding approximation error (quadratic and linear), quadratic regression coefficients a, b, and c, corresponding approximation error (quadratic and linear)
Minima/maxima related functionals (9)
range, position of max/min, difference min/max to arithmetic mean, number of peaks, mean distance between peaks, arithmetic mean of peaks, difference mean of peaks - mean

contains a certain number of redundancies and thus serves as a good starting point for feature selection. Finally, all features are normalised, where the statistics of the training set are used to normalise the test set as well.

2.1.4 Window-based Classification

To better capture the non-stationary nature of the scenes, classification is performed on smaller windows. Each recording is split into (overlapping) windows with a length of several seconds, and the statistical functionals are computed for all LLDs in those segments. In two previous studies of acoustic scene recognition [166, 120], a window length of 1 s was proposed. However, the experiments performed for the present study showed that longer windows lead to better results. These segments can capture the different acoustic events contributing to the acoustic scenes. This windowing is performed on the training data and on the test data. Thus, models are trained with a larger number of training instances per class (N_w windows per recording, each for N_{tr} training recordings). Classification is performed on the windowed test data. Each of the N_w windows is separately fed to the classifier so that it recognises one part of the scene. In order to get one decision for the whole instance, a majority voting scheme is employed. Weighting the single classification results by their confidence (which, in the case of SVM as classifier, is obtained by fitting the output of the SVM

to a logistic regression model) brought no improvement and thus, the majority vote is not weighted.

For classification, the SVM classifier [33] is used. An SVM is a non-probabilistic discriminative classifier for a two-class problem. In its simplest, linear form, an SVM is a hyperplane in the multi-dimensional feature space that separates two sets of example feature vectors with the objective of a maximum margin. The working principle of an SVM is illustrated in Figure 2.2. The output of a linear SVM is given by the formula

$$f(x) = w \cdot x - b, \tag{2.1}$$

where w is the normal vector to the hyperplane, x is the input feature vector, and b is an offset. The separating hyperplane is found through the optimisation problem

$$\arg\min_{w,b} \frac{1}{2} \|w\|^2 \tag{2.2}$$

while trying to maximise the so-called margin, which is the distance of the hyperplane to the nearest training data samples (which are called the support vectors). Additionally, so-called *slack variables* are introduced to allow (but penalise) the failure of an example to reach the correct margin. If the two classes are not linearly separable in the original feature space, the feature space can be transformed to a higher-dimensional space. Applying a hyperplane as a decision boundary in the higher-dimensional space leads to a nonlinear classifier in the original feature space. This is realised using a nonlinear kernel function instead of multiplication. The method is extended to multi-class classification by training pair-wise SVMs. In the present experiments, the system employs SVMs with a linear kernel and a complexity of 1.0 (this parameter trades off the width of the margin with the number of margin failures). SVMs are very well suited for the problem of this study because of the small number of classes and the small amount of training data per class. They are trained with the sequential minimal optimisation algorithm [170] using the windowed training data. The implementation of the Weka toolkit [94] is used.

For comparison, a nearest neighbour classifier is tested. Preliminary experiments showed that nearest neighbour performed better than k-nearest neighbour. As a distance function, the Euclidean distance is used for the nearest neighbour classifier. Since the features are normalised, this distance function gives better results compared to, for example, the cosine distance.

2.1.5 Latent Perceptual Indexing

In addition to SVM and nearest neighbour, an approach for acoustic scene classification based on latent perceptual indexing, as presented in [120], was implemented and evaluated.

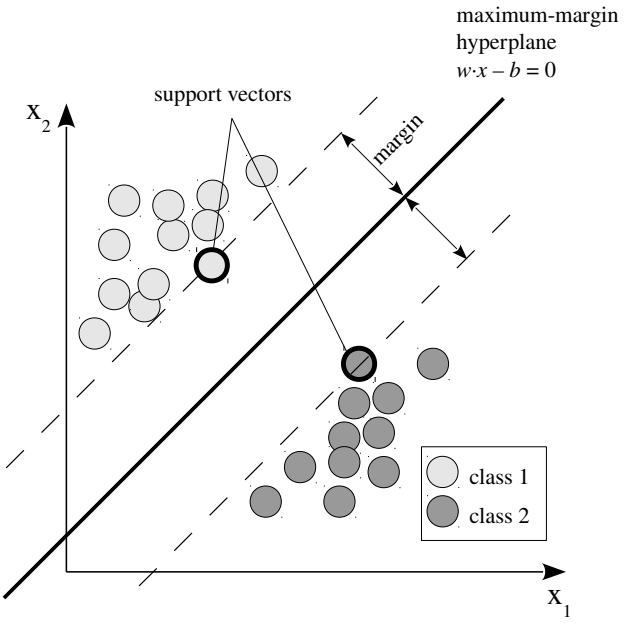


Figure 2.2: Working principle of an SVM for a two-dimensional feature space. Adapted from [14].

With the employed SVM or nearest neighbour approach, the contributing parts of each acoustic scene are recognised separately due to the windowing of training and test data. Thereby, all acoustic events are processed and classified on their own. With a majority voting, a decision for the whole recording is made. This approach ignores the overall composition of a sound scene to a certain extent, recognising only the distinct sound sources, without using a higher-level decision logic. Such a decision logic could take account of all sounds and decide on the acoustic scene based on the mixture of single acoustic events.

Latent perceptual indexing is an approach in which the classification of an acoustic scene is based on the composition of the contributing sounds. Each recording is represented as a vector in a latent perceptual space. First, a clustering (using k-means) of all (windowed) training data is performed to obtain a number of reference clusters that is higher than the number of classes. Then, for each recording, a *bag of feature vectors* is computed using the windows. The recording is transferred into the latent space by counting the occurrence of each of the reference clusters for this recording. The dimension of this latent space is the number of reference clusters. Thus, each training and test recording is described by a vector in the latent space. Transformed test recordings are classified using a nearest neighbour classifier and the cosine vector similarity. Preliminary experiments showed that in the latent space, a nearest neighbour classifier is as good as using SVMs. To obtain a more

fine-grained representation in the latent space, smaller window sizes are used in the latent perceptual indexing method, compared to the SVM or nearest neighbour classifier. A more detailed description of this method is given in [120].

One disadvantage of latent perceptual indexing compared to SVMs is that it requires more training data. In the SVM approach, classification is performed on the window level, whereby the training data are windowed as well. Thus, in the SVM approach, from N_{tr} training recordings, $N_{tr} \times N_w$ training instances are available. On the contrary, the latent perceptual indexing approach represents each training recording as a single vector in the latent space, and therefore, the number of training instances is equal to the number of training recordings N_{tr}. In the experimental validation in this study, there are only small amounts of training data, which is probably not enough for latent perceptual indexing to work properly.

2.1.6 Experimental Evaluation

This section describes the experimental setup and results. Several experiments were performed in order to investigate the behaviour of the system and to analyse the contribution of different parameters and system components to the classification performance. All implemented systems were first evaluated with the development set of the database of acoustic scenes from the challenge on detection and classification of acoustic scenes and events. Finally, the best system configuration was evaluated using the non-public test set of the database. For evaluation, 5-fold cross validation was performed, which is the official protocol of the challenge. The evaluation measure is the average accuracy (in %) over all folds, whereby additionally, the 95 % confidence interval is reported. It has to be noted that, given the small size of the dataset, the minimum significant improvement is relatively large. When results in the order of 60 % are achieved, the improvement in accuracy has to be roughly 12 % to be significant. Significance was evaluated using a one-sided z-test and a p-value of 5 %.

2.1.6.1 Database

For evaluation of the system, the official dataset of the challenge on detection and classification of acoustic scenes and events [77] was employed. Thereby, only the data of the scene classification track were used. This dataset contains 30 second long recordings of various acoustic scenes, categorised into ten different classes: *bus, busystreet, office, openairmarket, park, quietstreet, restaurant, supermarket, tube, tubestation*. For each of the ten classes, the database contains ten recordings, summing up to 100 recordings and 50 minutes total length. In addition, for the challenge, the systems were evaluated with a non-public test set of the same size, containing similar data. Sounds were recorded with a high-quality binaural recording system, whereby the portability and subtlety of the system allowed to obtain unobstructed

Table 2.3: Scene classification accuracies (development set) for two different feature sets (MFCC vs. all features), using the simple feature extraction method or the window approach.

Features	**Simple**	**Window**
MFCC	50 %	68 %
All	60 %	73 %

Table 2.4: Scene classification accuracies (development set) for different window lengths, keeping the window shift constant at 2 s (except for window length 1 s, where the shift is 1 s). SVMs are used as classifier, with the full feature set.

	Window length [s]					
	1	2	3	4	5	6
Accuracy [%]	62	71	67	73	66	64

everyday recordings with relative ease. Since the recordings were performed with binaural microphones on the ears of a person, the head-related transfer function of that person is intrinsically incorporated.

2.1.6.2 Windowing Approach

As a first experiment, the influence of applying the windowing approach was analysed. In this experiment, all 6 669 features or only the MFCC features (MFCCs 0–12, delta and delta-delta coefficients, 39 functionals, summing up to 1 521 features) were used in combination with an SVM classifier. Table 2.3 lists the experimental results for these two feature sets, either using the window approach (4 s long windows with 50 % overlap) or performing classification on the whole recordings. It can be seen that with MFCC features and without the windowing method, an accuracy of 50 % was obtained, which can be considered a rough baseline. Segmenting training and test data into 4 s windows and making a majority vote over single window classification results increased the accuracy to 68 %. On top of that, adding the other energy, spectral, and voicing-related features improved the result (not significantly) to 73 %.

Next, the influence of varying the window length was investigated. Results (using the whole feature set and SVM) are shown in Table 2.4. With a window shift of 2 s, smaller window sizes led to better accuracy, whereby the best result was achieved with a window length of 4 s. This is in contrast to findings in [166], where a window size of 1 s was found to be optimal. One possible reason for this difference is that in that study, more training data were used, which made it possible to apply a finer resolution of the data. Furthermore, the database was divided into more acoustic

Table 2.5: Results for different features, functionals, and classifiers with accuracy [%] supplemented by the 95 % confidence interval, using the development set.

Features	**Functionals**	**Classifier**	**Accuracy [%]**
All LLDs	All	SVM	73 ± 5
MFCC 0-12	All	SVM	68 ± 5
All LLDs	mean, var	SVM	64 ± 9
MFCC 0-12	mean, var	SVM	64 ± 9
All LLDs	mean, var	nearest neighbour	50 ± 12
MFCC 0-12	mean, var	nearest neighbour	39 ± 6

classes, which made such a finer resolution necessary in order to distinguish between these classes.

2.1.6.3 SVM and Nearest Neighbour Results

Different feature configurations were tested when comparing SVM and nearest neighbour classifiers. In addition to using the full set of all LLDs and all functionals (6 669 features), smaller feature sets were obtained by taking only the MFCCs (1 521 features) and/or using only mean and variance instead of all functionals (342 features for all LLDs and 78 for MFCCs). Table 2.5 shows the results for the experiments with different feature configurations. The best performance (73 %) was obtained with the full feature set and SVM as classifier, while using only MFCC features (68 %) was slightly worse. Reducing the set of functionals to only the mean and variance of each of the LLDs led to a degradation in performance to 64 % for both feature sets. These interesting results show that the additional features (compared to only MFCC) are only relevant when being combined with the large set of functionals. For the nearest neighbour classifier, acceptable results were only obtained with the reduced set of functionals, resulting in an accuracy of 50 % for all LLDs and 39 % for MFCC. Interestingly, for the nearest neighbour classifier, there was a larger difference in the performance of MFCC vs. all LLDs. Here, the LLD seem to introduce important information that improves the classification results.

Table 2.6 shows the confusion matrix for the best-performing system, using all proposed features with SVM and the employed window approach. Some classes (*bus*, *busystreet*) were recognised with 100 % accuracy, while for others (*park, restaurant, tube*), scores as low as 40 % were obtained. Most confusions were made between the classes *park* and *quietstreet* or between *restaurant* and *supermarket*. The recordings of the classes *park* and *quietstreet* are partly very similar. Recordings of the classes *tube* and *tubestation* contain a high variability, depending on the actual acoustic events. Therefore, these classes were confused with several other classes.

Table 2.6: Confusion matrix of the development data for the proposed system, achieving an accuracy of 73 %.

	bus	busyst.	office	open.	park	quietst.	rest.	superm.	tube	tubest.
bus	**10**	0	0	0	0	0	0	0	0	0
busystreet	0	**10**	0	0	0	0	0	0	0	0
office	0	0	**9**	0	0	0	0	1	0	0
openairmarket	0	0	0	**9**	0	0	0	1	0	0
park	0	0	0	0	**5**	5	0	0	0	0
quietstreet	0	0	0	1	2	**7**	0	0	0	0
restaurant	0	0	0	2	0	0	**4**	4	0	0
supermarket	0	0	0	0	0	0	1	**8**	0	1
tube	0	1	0	0	0	0	1	2	**5**	1
tubestation	1	0	0	0	1	1	1	0	0	**6**

Table 2.7: Results for the latent perceptual indexing approach with MFCC features for different numbers of clusters.

	# clusters						
	10	20	50	100	200	500	1000
Accuracy [%]	32	36	44	42	43	46	44

2.1.6.4 Results for Latent Perceptual Indexing

For latent perceptual indexing, the best results were obtained with MFCC features and all functionals. Furthermore, a smaller window size led to better results. Therefore, 1 s windows without overlap were employed. The results for different numbers of clusters are shown in Table 2.7.

The best result (46 ± 10 %) was obtained with 500 clusters. Generally, the performance was similar to the nearest neighbour approach. Considering that the classification was performed on the whole recordings instead of windowed data, the accuracy was also comparable to the SVM system with MFCC and without the window approach. Generally speaking, such an approach based on latent perceptual indexing requires more training data to deliver better results.

2.1.6.5 Feature Analysis

In order to better understand the contribution of different features to the classification result, a method for feature analysis was developed and applied to the system. For each of the employed 6 669 features, a score was computed using a t-test. The

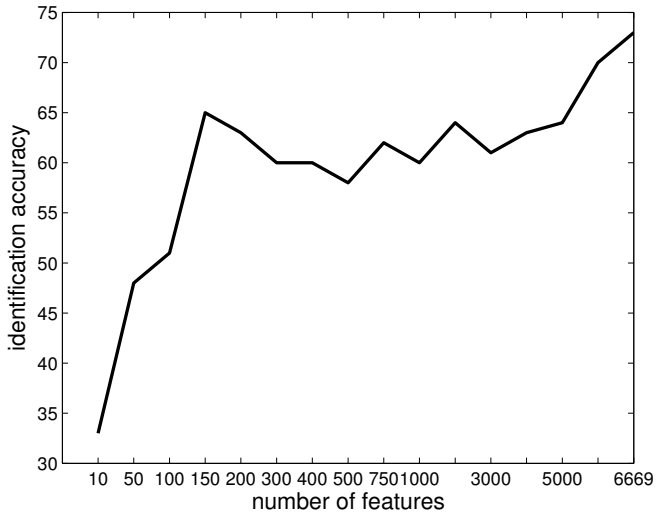


Figure 2.3: Influence of number of features on the accuracy [57] (note the nonlinear scale of the horizontal axis).

t-statistic was computed for each pair of acoustic classes and was summed up over all pairs to obtain a single score for each feature. According to Welch's t-test [231], the t-statistic t_{ij} for two different classes i and j is computed as

$$t_{ij} = \frac{\mu_i - \mu_j}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}} \tag{2.3}$$

using the expected value μ_i of the considered feature's distribution for the class i, the corresponding standard deviation σ_i and the number of samples belonging to this class n_i. This score was computed for each fold, using the training data of this fold. Summing up the scores over all folds resulted in a feature ranking. Using this feature ranking, another set of experiments was conducted, starting with the ten best features and gradually adding more features until the whole feature set was used. For this analysis, SVM was chosen as a classifier. Figure 2.3 shows the results of this experiment. Generally, with increasing number of features, the accuracy increased constantly, with an outlier representing a local maximum at 150 features and 65 % accuracy. This could either be a (statistically not significant) outlier, or it could be the case that the subsequent features are worse and therefore deteriorate the performance. The top 150 features contained mostly Mel spectra (116, whereby lower-order components were represented more often), but also energy (14), MFCC (14, only from the 12 th component), spectral flux (2), and position of spectral minimum (4). Comparing these results to Table 2.5, in which MFCCs achieved a similar performance, it can be concluded that MFCC and Mel spectra perform

equally well in this task. Most of the functionals were equally represented in the top 150 features. However, it stands out that 88 of them are variants of a mean value (e. g. mean of absolute values, mean of non-zero values, quadratic mean). The results of the feature analysis are in line with the results presented in Table 2.5, indicating the performance gain of the employed large feature set. Already with a comparably small feature set (MFCC or Mel spectra, with mean and variance as functionals), a relatively good performance can be achieved. Adding more LLDs and functionals leads to a small but substantial improvement.

2.1.6.6 Test Set

The best system configuration (all LLDs, all functionals, SVM classifier on 4 s windows with 50 % overlap) achieved an accuracy of 69 ± 12 % on the non-public test set of the employed corpus. Generally, the same tendencies can be observed as for the development set and the system made similar confusions. This result shows that the proposed system generalises well to a previously unseen test set. Although the accuracy was slightly worse than on the development set (73 %), it can be concluded that, even with such a large feature set, there is no overfitting of the system.

Detailed test set results for all systems that participated in the scene classification track of the challenge on detection and classification of acoustic scenes and events are publicly available [76] [1]. Here, a comparison of the experimental results of the participants in the challenge is included. Accuracies obtained with the official challenge test set are listed in Table 2.8. The baseline system (MFCC + GMM) implemented by the organisers of the challenge achieved an accuracy of 55 %. Almost all of the challenge participants were able to beat this result. The best performing system used features derived by recurrence quantification analysis of MFCCs which describe the temporal dynamics of the acoustic scene. A large portion of the submitted systems used various simple audio features together with SVMs for classification. As in the system proposed in this thesis, the classification was performed on larger windows, followed by majority voting, in many of the participating systems. Due to the larger feature set, the proposed system achieved better results than many of these similar systems.

2.1.7 Conclusions

In this section, a system for acoustic scene classification was presented and evaluated. Using large-scale audio feature extraction and SVM, an accuracy of 73 % was obtained on the development partition (69 % on the test set) of the dataset of acoustic scenes from the challenge on detection and classification of acoustic scenes

[1] http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/resultsSC.html, last accessed in April 2014

Table 2.8: Test set results in terms of accuracy (Acc.) for all participants in the scene classification track of the challenge on detection and classification of acoustic scenes and events.

Features	**Classifier**	**Acc. [%]**
recurrence quantification analysis of MFCCs [188]	SVM	76 ± 7.2
Wavelets+MFCCs [137]	Treebagger	72 ± 8.4
histogram of oriented gradients of spectrum [177]	SVM	69 ± 7.8
proposed: spectral, energy, voicing features	SVM	69 ± 12.0
i-vector of MFCC [42]	LDA	65 ± 3.1
spectral features, loudness [32]	HMM	65 ± 6.9
learnt by restricted Boltzmann machine [152]	SVM	60 ± 8.2
MFCC + others [157]	SVM	60 ± 8.7
spectrotemporal modulations [164]	SVM	58 ± 9.0
Cochleogram features [129]	SVM	55 ± 4.4
baseline: MFCC [77]	GMM	55 ± 10.0
normalised compression distance [159]	Random Forest	17 ± 2.4

and events [77]. This is an improvement compared to using only MFCC features (68 %).

In a detailed experimental section, the influence of different features was investigated. A substantial performance gain was achieved by adding energy, spectral, and voicing-related features to a feature set consisting of MFCCs. This shows that MFCCs can not capture all the relevant information that is needed to characterise complex acoustic scenes. Furthermore, it was found that these additional audio features are only relevant when coupled with the employed set of statistical functionals. The approach of window-based classification followed by majority voting proved to be highly efficient. A feature analysis showed that Mel spectra were an important factor for the high performance. In addition, the other employed energy and spectral features help to better capture the information in the acoustic scenes. Some acoustic scenes (*park, restaurant, tube, tubestation*) are difficult to recognise due to the high variability in the class and the similarity between the different classes. Future work could apply source separation techniques such as non-negative matrix factorization (NMF) to better handle overlapping acoustic sources.

In a system that tries to analyse an arbitrary audio recording, categorising the acoustic scene is the first step of the processing chain. Such knowledge is helpful in the next processing steps, since the system can now adapt to the expected acoustic events. One example is the detection and classification of individual acoustic events. If the acoustic scene can be classified, prior probabilities of acoustic events that might happen can be preadjusted accordingly. In the next section, a system for acoustic event classification is presented.

2.2 Supervised Learning of New Sound Events

Once the type of an acoustic scene is determined, one of the next goals is to detect and classify individual acoustic events. In this section, a system for acoustic event classification in an office environment is presented and evaluated. Hidden Markov models (HMMs) are used to model different acoustic event classes. A special focus is on the system's ability to learn new acoustic events. This problem is known as the open-set case, in which a class of acoustic events appears that was not included in the training phase. It is evaluated how new classes can be learnt using maximum a posteriori (MAP) adaptation [14] or conventional training methods. Experiments are performed with a database of acoustic events in an office environment.

In a typical office environment, many different sounds are produced either by a human or by objects handled by humans. The detection and classification of such acoustic events is a less explored research area, compared to the fields of speech or speaker recognition. Examples of these acoustic events are human non-vocal sounds such as coughing or other sounds such as keyboard typing or closing a door. Being able to detect and identify these acoustic events helps to analyse the human activity that takes place. Acoustic event detection and classification can also be used to enhance automatic speech recognition. Both can be regarded as disciplines in the area of computational auditory scene analysis [228].

A good overview of recent advances in acoustic event detection and classification technology in an office environment is given in [216]. Several international evaluation campaigns are described, in which different approaches have been deployed, mainly using HMMs or SVMs. The classification of events, activities and relationships (CLEAR) evaluation provided a testbed for different systems for the detection and classification of acoustic events [217]. Another domain for acoustic event recognition was considered in [128], where the goal was to classify acoustic events in a kitchen environment. Small improvements over an MFCC baseline were obtained with features derived from a temporal extension of independent component analysis, together with GMMs or HMMs as a classifier. Other possible domains for the application of acoustic event detection and classification are the detection of key audio events in sports games [106] or affective video content analysis [248]. Acoustic event detection and classification can also be part of a robot audition system as described in [158].

Whereas most of the listed studies describe only closed-set recognition systems in which the same classes appear during training and testing, open-set recognition is the challenge in which previously unknown classes may appear in the test phase. In this case, the fact that an acoustic event belongs to a previously unknown class needs to be detected. This is known as novelty detection [8]. After detecting an acoustic event as being novel, the problem is to add a new class to the classification system. This problem is analogous to the enrolment of a new speaker in a speaker recognition system. The standard approach utilised in speaker verification is to use a universal

background model which contains training data of many different speakers and to use MAP adaptation to derive a model for a new speaker based on limited amounts of enrolment data [182]. However, constructing a universal background model from many different acoustic events, which can be of very diverse nature, might not be efficient.

This thesis presents a system that classifies acoustic events. As a classifier, this system uses HMMs with MFCC features, whereby each class of acoustic events is modelled by one HMM. The present study wants to concentrate on the case in which an acoustic event appears that is not known to the system and is already detected as novel. Two different approaches to add new classes (with limited data) to the system are evaluated. The first approach is to simply perform a complete expectation maximisation training cycle with the instances of the new class. The second approach uses MAP adaptation to derive the model of a new class from the model of one of the known classes. To evaluate the system, a small database of distinctive acoustic events from an office environment was recorded. A possible application scenario is a robotic platform which can learn typical sounds of its daily environment.

In Section 2.2.1, a general system overview and a description of the methods to learn new classes is given. The recorded database, experiments, and results are described in Section 2.2.2, followed by some concluding remarks.

2.2.1 Acoustic Event Classification

This study tackles only the problem of classification of acoustic events and ignores the processing step of event detection. In a silent environment, this step could be performed using an energy-based voice activity detection (VAD) system. As a baseline system for event classification, MFCCs are extracted as features and continuous HMMs are employed for the classification task. The system is subsequently improved by optimising the number of HMM states using an approach based on the Bakis length modelling method. The main part of this study addresses the problem of adding a new class to the classifier. When an acoustic event is detected as previously unknown, it can be added to the system such that when it appears the next time, it can be regarded as known to the system. In order to do this with an HMM-based system, a new model needs to be created for the new class. Two possibilities to achieve this are compared here. The first is to repeat the training phase as it was done with the other classes (using expectation maximisation training). In this approach, the parameters of all models are initialised with global statistics. As a second possibility, MAP adaptation can be used to create a model for the new class, leaving the other models unchanged.

2.2.1.1 Audio Features

As acoustic features, standard MFCCs (+ energy) with delta and delta-delta coefficients were used. Whereas the baseline system used 12 MFCC coefficients, preliminary experiments showed that the best results could be achieved using 8 MFCC coefficients, which, together with energy, delta, and delta-delta coefficients, leads to a total number of 27 extracted features. The features were calculated for overlapping windows of 25 ms size using 60 % overlap (which corresponds to a frame shift of 10 ms). During feature extraction, the stereo signal was converted to mono by averaging over both channels.

2.2.1.2 Pattern Recognition with Hidden Markov Models

An HMM is a statistical generative classifier [175]. In an HMM system, one model is constructed for each class. This model can estimate the probability for a specific observation of belonging to the class represented by this model. HMMs became popular in speech recognition because of the ability to handle dynamic sequences.

An HMM can be regarded as a combination of two stochastic processes, where the first is a state machine and the second describes the emission of observations. Formally, an HMM can be defined as the tuple

$$\lambda = (S, V, \Pi, A, B), \tag{2.4}$$

where $S = (S_1, \ldots, S_N)$ is the set of possible states in the state machine, $V = (v_1, \ldots, v_K)$ are the possible observations, $\Pi = (\pi_1, \ldots, \pi_N)$ are the starting probabilities for the states in the state machine, $A = (a_{ij}), 1 \leq i, j \leq N$ are the state transition probabilities, and $B = (b_{ik}), 1 \leq i \leq N, 1 \leq k \leq K$ are the observation probabilities $b_{ik} = p(O_t = v_k | q_t = S_i)$ for observation O_t given HMM state q_t at time steps $1 \leq t \leq T$. The states of an HMM are hidden and only the observations can be observed. Figure 2.4 shows an exemplary three-state HMM, where the black arrows represent the state transitions and the grey arrows depict the observation probabilities. With a given set of parameters λ, the probability $p(O|\lambda)$ corresponds to the probability that the HMM λ produces the observation O. A classification problem can be solved by providing an HMM for each class k and performing maximum-likelihood classification

$$\hat{k} = \arg\max_k p(O|\lambda_k), \tag{2.5}$$

choosing the model with the highest observation probability. Algorithms are available to estimate optimal model parameters given a set of training observations or to determine the observation probability of a test observation in an efficient way [175].

Often, the discrete observation probabilities b are replaced by continuous distributions (e. g. using GMMs) in audio processing. Furthermore, it is common to employ left-right HMMs, where no back-stepping is allowed in the state machine.

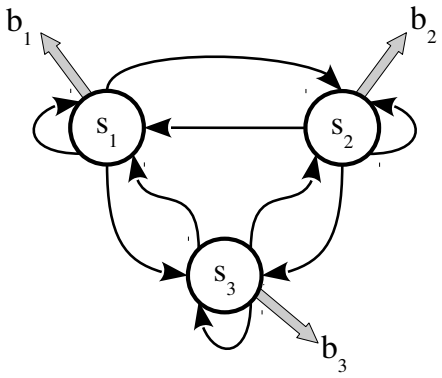


Figure 2.4: Diagram of a three-state HMM.

2.2.1.3 Classifier

HMMs, implemented in the hidden Markov model toolkit [252], were used as a classifier to recognise the acoustic events. Each model corresponds to one class, and the observations were represented by audio feature vectors.

The models were designed as continuous HMMs with a left-to-right topology; the observations were modelled by a mixture of Gaussians, defined by a mean vector, covariance matrix, and mixture weight for each Gaussian. For each class of acoustic events, one model was created in the training phase of the classifier.

In order to optimise the number of HMM states, an approach based on Bakis length modelling [9] was adopted, as proposed in [55]. This methodology follows the finding that the length of the acoustic events is not constant for all different types of events. Each HMM should be able to adequately model the corresponding class and therefore, the number of HMM states should be adapted to the expected length of the acoustic events. The baseline system used the same fixed number of states for every model (also referred to as fixed length modelling). The Bakis length modelling method proposes to set the number of states of each model to a fraction of the average length of the instances of the class. Here the length corresponds to the length of the recording of the acoustic event in seconds, which can be derived from the recorded database. In the employed variant of the Bakis length modelling method, the number of states $N(k)$ for a model corresponding to class k was set to

$$N(k) = c + f \cdot \bar{t}(k), \tag{2.6}$$

where $\bar{t}(k)$ is the average length of the recordings corresponding to class k while c and f are an additive constant and a factor, respectively, which were optimised heuristically. Using this method, the number of HMM states can be modelled to better fit the class statistics.

2.2.1.4 Learning New Acoustic Events

Once an acoustic event has been detected as being novel (novelty detection is not covered in this study), it can be added to the database of known acoustic events. Two ways to learn a new class of acoustic events are compared. The first, conventional way (which is referred to as *train* method in the following,) is to perform a normal training cycle using expectation maximisation. The training is performed with all classes to ensure proper model initialisation.

Another possibility to add a new class to the classifier is to use MAP adaptation. This approach will be called *adapt* method in the following. The model for the new class is not built up from scratch, but it is derived from another model using MAP adaptation. First of all, the new, unknown acoustic event is classified as one of the known classes, then this class is used as a starting point for MAP adaptation. The new model is created by copying and adapting the model of the most similar class. As adaptation data, the recorded (one or more) instances of the new class are used.

MAP adaptation (also known as Bayesian adaptation [14]) is used to adapt the means, mixture weights, and variances of the output probabilities of the HMMs. The mean of mixture component m is adapted as

$$\hat{\mu}_m = \frac{N_m}{N_m + \tau}\bar{\mu}_m + \frac{\tau}{N_m + \tau}\mu_m, \tag{2.7}$$

where $\hat{\mu}_m$ is the adapted mean, $\bar{\mu}_m$ is the mean of the observed adaptation data, μ_m is the old mean of the Gaussian, N_m is the occupation likelihood of the adaptation data for mixture component m, and τ is a weighting factor. Equation (2.8) is used to adapt the variances of the output probability distributions of the HMMs:

$$\hat{\sigma}_m^2 = \frac{N_m}{N_m + \tau}\mathrm{E}_m(x^2) + \frac{\tau}{N_m + \tau}(\sigma_m^2 + \mu_m^2) - \hat{\mu}_m^2. \tag{2.8}$$

Here, $\hat{\sigma}_m^2$ is the adapted variance, σ_m^2 is the old variance, and $\mathrm{E}_m(x^2)$ is the expected value of the squared observation vector x^2 for mixture component m. The adaptation of mixture weight w_m for mixture m follows

$$\hat{w}_m = (\frac{N_m}{N_m + \tau}\frac{N_m}{T} + \frac{\tau}{N_m + \tau}w_m)\gamma, \tag{2.9}$$

where $\hat{w}_m$ is the adapted mixture weight, w_m is the original mixture weight, T is the length of the adaptation data, and γ is a normalising factor which is needed to ensure that all mixture weights sum up to 1.

The weighting factor τ was optimised heuristically. Smaller values of τ lead to higher adaptation, which means (in case for the mean) that the new mean is nearer at the mean of the adaptation data than at the old mean. If τ is set to zero, the new model corresponds to a model trained only with the adaptation data. Conversely,

higher values of τ lead to less adaptation. In the case of $\tau \to \infty$, the old model is copied in order to get the new model, while neglecting the adaptation data.

2.2.2 Experimental Evaluation

This section presents the experimental validation of the proposed system.

2.2.2.1 Database

For testing purposes, a database of non-overlapping acoustic events was recorded with a pair of microphones of a robotic platform in a silent office environment. During the recording process, an automatic energy-based VAD algorithm was used to detect and segment the acoustic events. Thus, the database contains segmented recordings of acoustic events. 15 different classes of acoustic events from an office environment were chosen to be included in the database. These classes include ambient sound events and command-oriented social signals as well as gestures that are intended to provide a scene analysis system with a better understanding of its environment. Two special acoustic events that were included in the database are *speech* for normal speech and *garbage* for all sounds that appeared during the recording that are not included in any of the other classes. Table 2.9 shows the 15 different classes, their frequency in the database, and the average length of the recordings (which is needed for the optimisation of HMM state numbers). The average recording length includes a short phase of silence at the beginning and at the end of each recording. The recorded database is not intended to be used for a very generalised recognition system. For example, it will not be able to recognise all kinds of closing doors. The database is rather intended to be a small sample of very specific sounds from a small environment. In total, the database is made up of 506 single acoustic events.

2.2.2.2 Baseline Results

For the baseline system, 12 MFCC features were calculated during feature extraction. Together with the energy and their delta and delta-delta coefficients, this sums up to a total of 39 features. The baseline system used fixed length modelling (with 6 HMM states for each class) to determine the number of HMM states.

In order to get reliable results, a 5-fold stratified cross validation was applied. Therefore, the instances of each class in the database were randomly divided into five subsets and the experiments were conducted five times, whereby each time another one of the subsets was used for testing. Each time, the remaining four subsets were used to train the models. The final classification result was obtained by averaging over the results of the five single experiments. Table 2.10 shows the confusion matrix for the baseline system. Eight out of the 15 total classes were recognised with 100 %

Table 2.9: Overview of the 15 classes of different acoustic events in the database, their frequency, and the average length of the recordings in seconds. The total number of acoustic events in the database is 506, with an average length of 1.69 s.

Class	**# files**	**avg. length [s]**
chair rolling	22	1.68
chair squeak	24	1.16
clap	36	1.27
cough	51	1.56
door closing	21	1.42
finger snap	30	1.13
garbage	25	1.96
glass placement	51	1.26
key laydown	16	1.25
key rattle	41	2.43
keyboard	39	1.43
paper rustle	42	2.30
paper tear	44	1.38
speech	36	2.69
steps	28	2.03

accuracy. The total accuracy was 95.9 %, which corresponds to an error rate of 4.1 %.

2.2.2.3 Results for Optimisation of HMM State Numbers

Starting with the baseline system, the number of HMM states was optimised using Equation (2.6). The optimal parameters turned out to be $f = 2.5$ and $c = 4$. The application of these parameters (which corresponds to state numbers between 7 and 11) led to an error rate of 3.2 %, which is a relative improvement of 23.7 % compared to the baseline system. Finally, the number of MFCC coefficients was surveyed and the best result was achieved with eight coefficients (27 features in total) with an error rate of 2.8 %. The results of the baseline system and the optimisations are summarised in Table 2.11.

2.2.2.4 Results for Learning New Classes

To evaluate how good a new class can be learnt by the system, the experimental setup as described in the following was used. The described experiment was repeated 15 times, whereby each time, one of the acoustic event classes was used as the 'new' class.

Using the structure of the improved baseline system, the models were trained using all but one of the classes, whereby the same amount of training data was used

Table 2.10: Confusion matrix for the baseline recognition system. Many of the classes are recognised without errors.

	a chair rolling	b chair squeak	c clap	d cough	e door closing	f finger snap	g garbage	h glass	i key laydown	j key rattle	k keyboard	l paper rustle	m paper tear	n speech	o steps	**total error in %**
a	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	34	0	0	2	0	0	0	0	0	0	0	0	0	5.6
d	0	0	0	51	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	1	0	28	0	0	0	0	0	0	1	0	0	6.7
g	0	1	0	0	0	0	20	0	0	0	1	0	0	1	2	20
h	0	1	1	0	0	0	0	49	0	0	0	0	0	0	0	4
i	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0
j	0	0	0	0	0	0	0	0	0	41	0	0	0	0	0	0
k	0	0	2	0	0	1	0	0	0	0	35	1	0	0	0	10.3
l	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0	0
m	0	0	0	0	0	1	0	0	0	0	1	2	40	0	0	9.1
n	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0
o	0	0	0	0	2	0	0	0	0	0	0	0	0	0	26	7.2

Table 2.11: Summary of the recognition results of the baseline system and its improvements. Using a separate number of states for each HMM and adjusting the number of MFCC components reduced the error rate by roughly one third compared to the baseline.

System setup	**Error rate** [%]
baseline	4.1
improved state numbers	3.2
improved state numbers + MFCC	2.8

as for the baseline system. Then, the new class was added to the system using one of the two methods described in Section 2.2.1.4 (complete retraining or adaptation). One, two or three instances of the unknown class were used to create the new model. The remaining data of the new class were used for testing. A 5-fold cross validation was used to have each instance (of the 'old' classes) in the test set once. In addition, this experiment was repeated several times, whereby each time different instances of the new class were used to create the new model. The final average results are shown in Table 2.12. These results show that when using the *train* method, the average error rate for the newly added class was very high when the model was created from

Table 2.12: Error rates [%] for learning a new class using either full training (*train*) or MAP adaptation (*adapt*) to create the model of the new class. Either one, two, or three instances of the new class are used, whereby each time, error rates are reported for the new class (*new*) as well as for the other classes (*rest*).

	1 instance		2 instances		3 instances	
	new	**rest**	**new**	**rest**	**new**	**rest**
train	85.4	3.0	57.7	3.0	35.6	3.1
adapt	33.9	4.6	25.6	4.4	21.4	4.3

only one instance of the class, but, not surprising, it decreases very strongly when more instances are used. For the *adapt* method, using only one instance of the new class already delivered results that were much better than for the *train* method. Here, the performance increased as well when more instances of the new class were used, but not so strongly as with the first method. The classification results for the other classes should not be neglected. With the *train* method, the effect of adding a new class on the classification performance of the previously known classes was almost neglectable. This was not the case when MAP adaptation was used. In this case, the error rate for the previously known classes increased slightly. Looking at the detailed results shows that this was due to confusion of the class that was taken as a base for MAP adaptation with the new class.

2.2.3 Conclusions

This section presented a system for acoustic event classification based on HMMs. The system was tested with a database of acoustic events recorded in an office environment. Then, the case of open-set recognition was considered. When an acoustic event is detected as being previously unknown, the question was how it can be added to the system as a new class. Two approaches were compared, using full expectation maximisation training or MAP adaptation. The results showed that when using full training, the accuracy for the new class was not acceptable when only one instance of this class was used to train the model of this new class. However, using two or three instances improved the results. The classification performance of the other classes was not affected by adding a new class. As another approach to add a new class, MAP adaptation was evaluated. With only very few instances of a new class, MAP adaptation outperformed full training regarding the error rate for the new class. The error rate for the other classes was affected slightly negatively by adding a new class. Thus, it can be concluded that the proposed method is very efficient for learning new acoustic events in the case that only small amounts of training data are available.

2.3 Chapter Summary

This chapter presented the first processing steps in a system that analyses the contents of an acoustic scene. The first step is to categorise the acoustic scene (as a whole) into one class. A method for classifying the acoustic scene of longer audio recordings (30 seconds in the case of the present study) was introduced. Using a large set of cepstral, spectral, energy, and voicing-related audio features together with a set of statistical functionals, a description of the acoustic scene is created on the window-level. The windows (with a length of several seconds) are classified separately and a majority voting is employed to find a decision for the whole scene. In the experimental evaluation with the scene classification database of the challenge on detection and classification of acoustic scenes and events, the system achieved excellent results that beat many other participating systems. The second part of the chapter focussed on acoustic event classification, dealing with the special case of learning new acoustic events. An HMM-based system for acoustic event classification was implemented and evaluated with a database of acoustic events in an office environment. Two different methods for learning a new acoustic event were compared, whereby the proposed method achieved substantially better results.

With the developed methods, an acoustic scene can be categorised into one among a set of classes, and the appearing acoustic events can be classified. The result is the basis for further analysis of the audio sources. In the present study, a special focus is on human activity. The most prevalent sound produced by humans is the human voice, whereby the next goals of the system are to identify the speaker and to recognise the spoken contents, which will be the focus of later chapters in this thesis. Another example of sounds produced by humans are footsteps. Footstep sounds can also carry important information. The following chapter focuses on analysing footstep sounds in order to identify the walking person.

3

Acoustic Gait-based Person Identification

This chapter presents contributions and solutions in the field of person identification from step sounds. Two different approaches for acoustic gait-based person identification are presented.

3.1 Introduction

Recognising people by the way they walk (also known as gait recognition or gait-based person identification) is a relatively new field of research. Person identification can be found in many practical applications and is implemented in public places (e. g. shopping malls or train stations) or smart homes [114]. Approaches for biometric person identification can be categorised into physiological and behavioural methods. The most common methods for physiological biometrics use iris [259] or fingerprint patterns [148]. Behavioural approaches, in contrast, use voice (as illustrated in Chapter 4 of this thesis), gait [34], or other characteristics. Early research has already demonstrated that gait can be seen as a biometric feature [151] and that it can be used to identify humans [35]. Most of the previously proposed methods for gait-based person identification work in the visual domain, where this topic has been an active field of research for the last decade [134]. Acoustic information, however, can also be used for gait recognition. Here one has to differentiate between active and passive systems. Active systems emit ultra-sonic sounds and estimate gait signatures from the reflections [160, 257]. On the other hand, passive systems (which are the focus of this thesis) directly measure the step sounds emitted by a person walking. Even though the focus on this modality (using passive systems) has so far been significantly less, results are promising.

While in the visual domain gait identification systems can rely on analysing the silhouette [229], the task is much more difficult for systems working only with audio information. The relevant information that can be exploited by such systems

consists not only of the sounds of the steps, but also of the sounds of clothing on moving arms and legs. These sounds are influenced by the gait pattern of the person, making them suitable to be used for person identification. Furthermore, the sounds produced during walking are highly dependent on factors such as the type of floor, shoes, and clothing. A user study from Mäkela *et al.* [143] evaluated the potential of humans to recognise others by their walking sounds. After a training phase, twelve subjects were able to identify their co-workers by their walking sounds with an accuracy of 66 %. This result indicates that walking sounds convey characteristic information about the subject and can thus be used for person identification.

Potential applications of gait-based person identification using audio information are indoor surveillance scenarios as well as access control systems. Such an audio-based system can be used to enhance visual surveillance and facilitate multimodal approaches. Unlike video-based person identification, acoustic systems also work in the darkness, require less expensive hardware and often lower sensor density, and are less obtrusive. Acoustic gait-based person identification is also known as *acoustic gait recognition*.

3.1.1 Contributions

In this chapter, two different methods for acoustic gait-based person identification are proposed. First, a system is presented employing a static classifier. A large candidate feature set adopted from speech processing is employed, and features that are relevant for the addressed problem are selected systematically. Support vector machines (SVMs) are used for classification, and features are ranked and selected using a wrapper approach. In addition, it is shown how audio features can improve a conventional video-based system through multimodal fusion. Second, a person identification system using a dynamic classifier is presented. This system extracts cepstral features from the recorded audio signals and uses hidden Markov models (HMMs) for dynamic classification. A cyclic model topology is employed to represent individual gait cycles. This topology allows the system to model and detect individual steps, leading to very promising identification rates.

All experiments were performed with audio data from the TUM gait from audio, image and depth (GAID) database, which contains recordings of a large number (> 300) of subjects. Furthermore, the database features three different variations (normal walking, walking with a backpack, and walking with shoe covers), which allows for realistic experiments.

Next follows a discussion of related work. After introducing the employed database in Section 3.2, two identification systems are presented. First, a system based on a static classifier is proposed in Section 3.3, followed by a system that uses a dynamic classifier in Section 3.4. The chapter concludes with a short summary in Section 3.5. Sections 3.3 and 3.4 are based on the results published in [66] and [67], respectively.

3.1.2 Related Work

The most widespread approach for video-based gait recognition is the gait energy image (GEI) [95], which is a simple silhouette-based method. It can be combined with face recognition [103] or with depth information [102] to improve the recognition performance. Furthermore, model-based approaches have been proposed for visual gait recognition [249]. Besides using video or audio information, other methods to identify walking persons include using acoustic Doppler sonar [119] or pressure sensors in the floor [253]. For example, the authors of the study presented in [212] combined different classifiers to improve identification performance using footstep data obtained with a pressure-sensitive floor. The potential application of a biometric verification system exploiting footstep signals for a smart home environment is described in [187]. In [15], similar methods are used for detecting footsteps, employing information from acoustic and seismic sensors for this problem.

Using audio information is a relatively new approach for the task of gait-based person identification. In [201], footstep sounds were detected in a database of various environmental sounds. A system for person identification using footstep detection was introduced by Shoji *et al.* in [202]. This system used Mel-cepstrum analysis, walking intervals, and the degree of similarity of spectral envelope for classification, together with a method based on k-means clustering. The system was tested with a database of five persons. This study was extended in [112] by adding psychoacoustic features such as loudness, sharpness, fluctuation strength, and roughness. Finally, in [113], dynamic time warping was used for classification and the database was extended to contain ten persons. The system achieves almost 100 % perfect classification rates (using ten persons). This can, however, be attributed to the simplified task of classifying pre-segmented footsteps and to the small number of subjects in the database. A similar task is addressed in the recently published study by Altaf *et al.* [3], in which a database of segmented footstep sounds from ten persons is used. Instead of extracting spectral features, the shape and properties of a footstep sound are examined in a temporal energy domain. Furthermore, the person-dependent asymetry between left and right footsteps is used as a feature for identification; for each footstep, the system is given the information whether it was produced by the left or the right foot. As a result, an identification accuracy above 90 % is achieved when a large number of footsteps are used during testing. In the case that only data from three consecutive footsteps are available, which is more comparable to the study presented in this thesis, an accuracy of 45 % is obtained. In [38], a system for person identification based on walking sounds is presented. Employing mainly spectral features and static classifiers, classification rates range from 33.5 % to 97.5 % using a database with 15 subjects. The system presented in [2] recognises people walking on a staircase by exploiting the fact that the number of steps is known. Inter-peak distances and peak height are used as features.

The weakness of all previous studies about acoustic gait-based person identification is the fact that only small databases have been employed, which are overly prototypical and mostly contain no more than ten subjects. In addition, very often, classification is performed using pre-segmented footsteps, which requires an additional preprocessing step that cannot trivially be solved.

Compared to the discussed previous studies on acoustic gait recognition, this thesis introduces a larger number of features and a much larger database. To recap, in the mentioned studies, the employed audio features include gait frequency, spectral envelope, linear predictive coding coefficients, MFCCs, and loudness. In [2], inter-peak distances and peak height were used as audio features. In [215], acoustic features from the speech domain are used for the classification of acoustic events, which is similar to the problem in the present study, since the authors of the mentioned study also try to adopt features which were originally developed for speech processing to another audio recognition task. Another study about feature selection for acoustic event detection is [260], where the discriminant capability of each feature (candidate features are MFCCs and log frequency filter bank parameters) is quantified according to the approximated Bayesian accuracy. In the present study, similar features and methods for feature selection are considered for the problem of acoustic gait-based person identification.

3.2 The TUM GAID Database

The freely available[1] TUM gait from audio, image and depth (GAID) database [104] is used for the experiments in this study. This database was recorded with the goal of providing a possibility to evaluate and compare multimodal gait recognition systems. Therefore, data were recorded with a colour and depth sensor, as well as with a four-channel microphone array. Thus, a typical colour video stream, a depth stream, and an audio stream are simultaneously available. The database contains recordings of 305 subjects walking perpendicular to the recording device in a 3.5 m wide hallway corridor with a solid floor. In each recorded sequence, the subject walks for roughly 4 m, typically performing between 1.5 and 2.5 gait cycles (each of them consisting of two steps). Most of the sequences have a length of approximately two to three seconds. Three variations were recorded for each subject: normal walking ($\mathcal{N}$), walking with a backpack ($\mathcal{B}$), and walking with shoe covers ($\mathcal{S}$). The backpack constitutes a significant variation in gait pattern and sound, and the shoe covers pose a considerable change in acoustic condition. Figure 3.1 shows screenshots of the three different walking conditions for one subject. There are six recordings of the $\mathcal{N}$ setup and two each of the $\mathcal{B}$ and $\mathcal{S}$ setups for each subject. This sums up to a total number of 3 050 recordings. All instances of the same condition were recorded directly after each other for each subject, meaning that the same shoes and

[1] www.mmk.ei.tum.de/tumgaid, last accessed in April 2014

(a) Normal recording (b) Backpack recording (c) Shoe cover recording

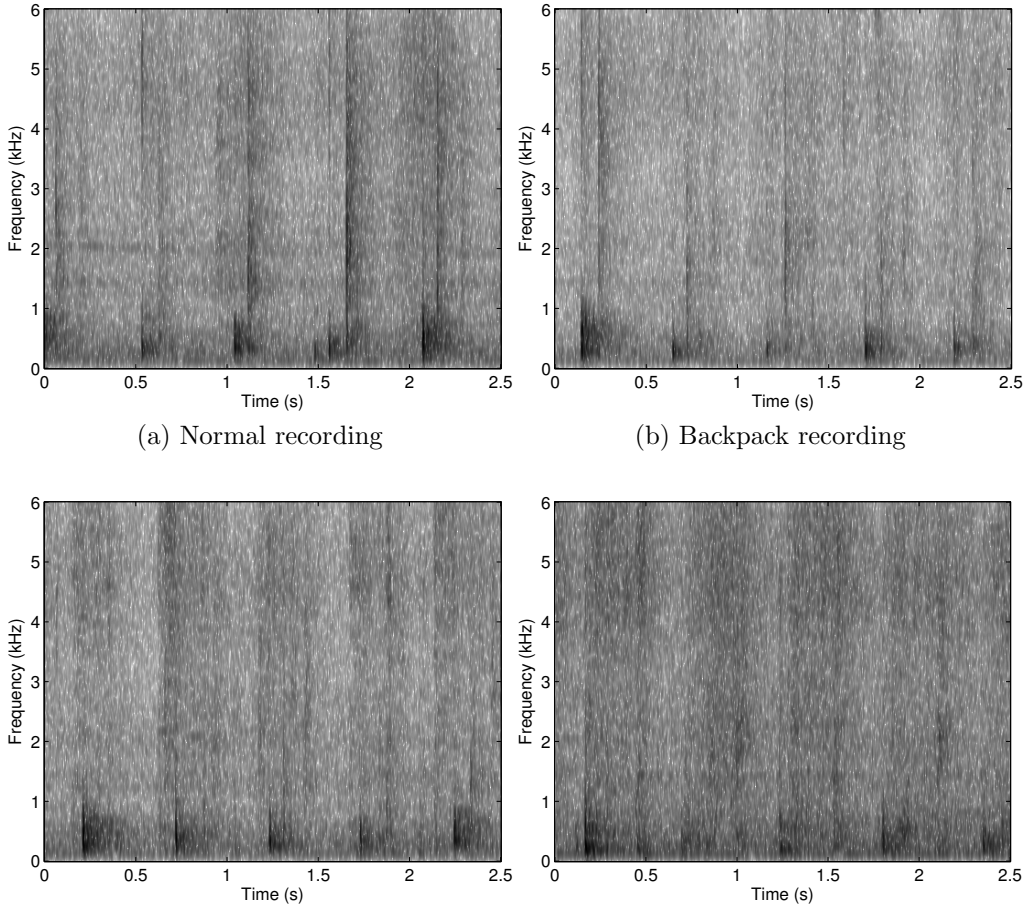
Figure 3.1: Screenshots of three recordings in the TUM GAID database.

clothes were used, which corresponds to a re-identification scenario. The metadata distribution of the database is well-balanced with a female proportion of 39 % and ages ranging from 18 to 55 years (average 24.8 years and standard deviation 6.3 years). More than half of the subjects were wearing sneakers while other commonly-used types of shoes are boots and loafers.

Furthermore, the TUM GAID database contains recordings produced at different dates in time (with three months in between) for a small number of subjects. These are, however, not used within the present study, but would allow for performing identification experiments independent of the type of shoes and clothes.

Figure 3.2 shows the spectrograms of four exemplary recordings in the database. The spectrograms reveal a lot of static background noise, which is due to the recording environment. Figure 3.2 (a) shows a normal recording of one subject, where each walking step is characterised by two successive sounds, whereby the first sound has stronger low-frequency components and the second sound has stronger high-frequency components. With a backpack (Figure 3.2 (b)), the steps get softer and other audible sounds are added by the backpack. When the subject wears shoe covers (Figure 3.2 (c)), more and longer high-frequency components are introduced, which are the rustle-like sounds of the shoe covers. For reference, Figure 3.2 (d) shows the spectrogram of another subject with sounds between the steps, which are the result of the legs of the trousers rubbing against each other.

To allow for a proper scientific evaluation and to prevent overfitting on the test data, the database is divided into a development set and a test set. The two sets are person-disjunct and contain 150 and 155 subjects, respectively. The partition of the database is shown in Table 3.1. Both for the development and for the test set, the first four $\mathcal{N}$ recordings of each subject are envisaged for the enrolment process. The other two $\mathcal{N}$ recordings as well as the $\mathcal{B}$ and $\mathcal{S}$ recordings are used to perform



(a) Normal recording

(b) Backpack recording

(c) Shoe cover recording

(d) Normal recording of another subject

Figure 3.2: Spectrograms of four recordings in the TUM GAID database [66].

the identification experiments. This means that models are learnt only using the $\mathcal{N}$ recordings, while the $\mathcal{B}$ and $\mathcal{S}$ conditions constitute previously unseen variations during the identification experiments and are therefore expected to deteriorate the identification performance.

3.3 Acoustic Gait-based Person Identification using SVM

In this study, several acoustic features are examined for their suitability for acoustic gait recognition. The candidate features are established in audio processing tasks like speech recognition, emotion recognition or acoustic event classification. Furthermore, the feature set also includes features which have been used in previous

Table 3.1: Partition of the TUM GAID database.

	Development (150 subj.)	**Test** (155 subj.)
$\mathcal{N}1$ – $\mathcal{N}4$	enrolment	enrolment
$\mathcal{N}5$ – $\mathcal{N}6$	identification	identification
$\mathcal{B}1$ – $\mathcal{B}2$	identification	identification
$\mathcal{S}1$ – $\mathcal{S}2$	identification	identification

studies of acoustic gait-based person identification. The size of the feature set is reduced using a wrapper-based feature selection technique. This, at the same time, increases the identification accuracy. For classification, SVMs are employed.

3.3.1 Candidate Features

The TUM GAID database provides audio signals with four audio channels recorded with a sampling rate of 16 kHz. Before the feature extraction step, the recordings were converted to mono by averaging over the four individual channels. In order to provide a first well reproducible and transparent baseline system, a brute-force large-scale feature extraction approach was used by employing the open-source toolkit openSMILE [45].

The employed audio feature set is based on the baseline audio features provided for the 2011 Audio/Visual Emotion Challenge [196] and it contains a number of energy, spectral, and cepstral features. Compared to that feature set, the voicing related features were omitted, as it was found that these are not relevant for the problem of gait recognition. The employed features are of supra-segmental nature. This means that the acoustic descriptors such as energy and spectral entropy (which are sampled at a fixed rate) are summarised over a recording (of variable length) into a single feature vector of constant length. This is achieved by applying functionals to the acoustic low-level descriptors (LLDs). Thereby, each functional maps each LLD signal into a single value for the given segment. Examples for functionals are mean, standard deviation, higher order statistical moments, and quartiles.

The set of LLDs and the functionals are listed in Table 3.2 and Table 3.3, respectively. All 25 LLDs were computed every 10 ms, where a window size of 60 ms was applied for the MFCCs and loudness features while all other features were computed based on windows with a length of 25 ms. Features that were analysed in previous studies about acoustic gait-based person identification [112, 113] such as the loudness, psychoacoustic sharpness or MFCCs are included in the tested feature set. In addition, the feature set provides a substantial amount of further acoustic information. Furthermore, for each LLD, first order delta coefficients (equivalent to the first derivative) were computed. The final feature set was made up of 25 LLDs

Table 3.2: 25 energy and spectral-related acoustic low-level descriptors (LLDs).

Energy-related features (3)
loudness (auditory model based), energy in bands from 250 Hz – 650 Hz, 1 kHz – 4 kHz,
Spectral features (12)
zero crossing rate, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacoustic sharpness, harmonicity,
Cepstral features (10)
MFCCs 1 – 10

× 42 functionals and 25 delta coefficients × 23 functionals, summing up to 1 625 features in total per recording.

3.3.2 Classification

SVMs with a linear Kernel function (as implemented in the Weka toolkit [94]) were applied as a classifier. A short introduction to SVMs is given in Section 2.1.4 of this thesis. Sequential minimal optimisation [170] (complexity 1.0) was used for training. The multi-class classification problem was handled by constructing pairwise SVMs. SVMs are discriminative classifiers which do not require large amounts of training data. This makes them especially suited for the problem considered in this study.

3.3.3 Baseline Results

Results for the candidate feature groups and their combinations (using the development set) are shown in Table 3.4. Altogether, the best results were obtained in the normal ($\mathcal{N}$) setup. Carrying a backpack ($\mathcal{B}$ setup) leads to a different walking pattern as well as to additional sounds and therefore to a decreased identification performance. Using shoe covers ($\mathcal{S}$ setup) completely changes the characteristics of the footstep sounds. All results in the $\mathcal{S}$ setup were however still better than the chance level (0.7 % accuracy).

The best single feature group, consisting of the MFCCs, led to an average accuracy of 23.1 %. Energy features (which constitute the smallest feature group with only three features) achieved the worst performance. Looking at the different combinations of the feature groups, it can be concluded that MFCCs and spectral features do not complement each other very well since there is no significant improvement when combining those two feature groups (significance was evaluated using a

Table 3.3: Set of all 42 functionals used for audio feature extraction. Only the statistical functionals are applied to delta coefficients, whereby the mean is only computed for positive values.

Statistical functionals (23)
(positive) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1 %, 99 % percentile, percentile range 1 % – 99 %, percentage of frames contour is above: minimum + 25 %, 50 %, and 90 % of the range, percentage of frames contour is rising, maximum, mean, minimum segment length, standard deviation of segment length
Regression functionals (4)
linear regression slope and corresponding approximation error (linear), quadratic regression coefficient and approximation error (linear)
Local minima/maxima related functionals (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude mean of minima, amplitude range of maxima
Other (6)
Linear Predictive Coding gain/coefficients 1 – 5

one-sided z-test). The best result was obtained by combining MFCCs with energy features, with an average accuracy of 28.1 %. This result is significantly better than with only MFCCs (significant with $p < 0.01$) and even better than the combination of all three feature groups.

3.3.4 Feature Analysis

In order to reduce the size of the feature set and to identify the relevant descriptors, an automatic wrapper-based [127] selection technique is applied. In general, wrapper-based means that the classifier is evaluated during the process of feature selection (e. g. in order to optimise the feature set).

A simplified version of sequential forward selection [174] was used for feature analysis. The classifier was trained and evaluated on the $\mathcal{N}$ setup of the development set for each of the $N = 1\,625$ candidate features (including all the delta coefficients

Table 3.4: Results on the development set (150 subjects), using different combinations of feature groups, the best 400 features as determined by the described feature selection technique, and the best acoustic low-level descriptors (the three energy features, spectral kurtosis, flux, and skewness as well as MFCC 1). Shown is the identification accuracy for the three setups $\mathcal{N}$ (normal walking), $\mathcal{B}$ (backpack), and $\mathcal{S}$ (shoe covers) together with the average (Avg.).

Features	**Condition** $\mathcal{N}$	$\mathcal{B}$	$\mathcal{S}$	**Avg.**
MFCC	41.7	22.3	5.3	23.1
energy	29.3	17.0	5.0	17.1
spectral	41.0	20.7	4.0	21.9
MFCC + energy	49.7	28.7	**6.0**	28.1
MFCC + spectral	44.0	26.3	4.7	25.0
energy + spectral	43.0	24.7	4.3	24.0
MFCC + energy + spectral	48.3	29.0	4.3	27.2
best 400 features	**57.7**	**30.7**	3.3	**30.6**
best LLDs	43.7	29.7	4.3	25.9

and functionals). For each of the features in the feature set $F = f_1, f_2, ..., f_N$, the accuracy $a(f_n)$ was computed as

$$a(f_n) = \frac{1}{T} \sum_{t=1}^{T} \delta\Big(\arg\max_c P(x_t | m(c, f_n)) - l(t) \Big) \tag{3.1}$$

where $x_t, t = 1, \ldots, T$ are the instances of the development set, $l(t)$ denotes the true label for instance t, $\delta(\cdot)$ is the Kronecker delta function, and $m(c, f_n)$ is the model of the classifier for class c, built using only feature f_n. In this experiment, accuracies $a(f_n)$ between 0 % and 7 % were achieved, with a mean of 1.8 %. The features were sorted according to their classification accuracy $a(f_n)$. Then, starting with the single best feature, more and more features were added to the feature set according to their ranking until the whole feature set was used. Figure 3.3 shows the results of this experiment on the development set. The best result was obtained using 400 out of the total 1 625 features. This result is also included in Table 3.4. Out of these 400 features, 89 are derived from MFCCs, 90 from energy features and 221 from spectral features. This composition suggests that in all of the three feature groups, there are relevant features. While an accuracy of 48.3 % was achieved with all 1 625 features in the $\mathcal{N}$ setup, this accuracy was raised by 19 % relative to 57.7 % with the 400 best features. This improvement is significant ($p < 0.05$). For the $\mathcal{B}$ setup, there was a non-significant improvement to 30.7 %, while the performance

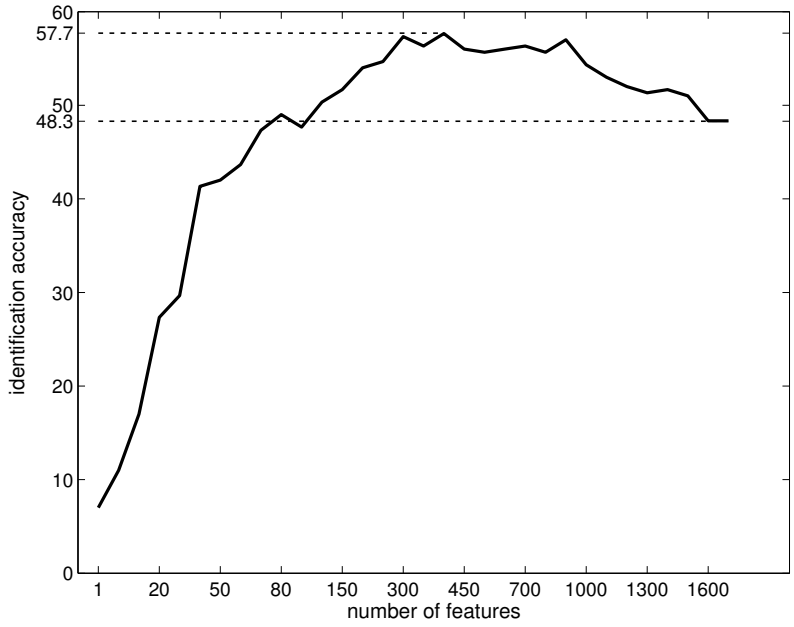


Figure 3.3: Identification accuracy on the $\mathcal{N}$ (normal) setup of the development set for different numbers of features. More features were added according to their identification accuracy as a single feature [66] (note the nonlinear scale of the horizontal axis).

in the $\mathcal{S}$ setup underwent a slight (non-significant) decrease to 3.3 %, compared to using all features.

The employed low-level features and functionals were further analysed in the following. In order to understand the relevance of each feature, a single score was obtained for each low-level descriptor and for each of the functionals. For each low-level descriptor, the average of $a(f_n)$ over the top 50 % performing functionals was computed. Analogously, for each of the functionals, the average accuracy was computed for the top low-level descriptors for this functional. The features which got the highest scores are the three energy features, spectral kurtosis, flux, and skewness as well as the first MFCC coefficient. These are also the features which are among the most common in the feature set of the best 400 features as determined by the simplified sequential forward selection technique. When only these LLDs were used (together with all delta coefficients and functionals, summing up to 455 features in total), an average accuracy of 25.9 % was obtained with the development set (cf. Table 3.4). Since these features have been ranked and selected independently of each other, it is understandable that the result is slightly worse than the baseline result of 27.2 % when using all features. Among the functionals, the best scores were achieved by the means (arithmetic and root quadratic), the standard deviation, the quartiles and quartile ranges, and the percentiles and percentile ranges. Similarly as

Table 3.5: Results on the test set (155 subjects), using different combinations of feature groups, the best 400 features as determined by the described feature selection technique, and the best acoustic low-level descriptors (the three energy features, spectral kurtosis, flux, and skewness as well as MFCC 1). Shown is the identification accuracy for the three setups $\mathcal{N}$ (normal walking), $\mathcal{B}$ (backpack), and $\mathcal{S}$ (shoe covers) together with the average (Avg.).

	Condition			
Features	$\mathcal{N}$	$\mathcal{B}$	$\mathcal{S}$	**Avg.**
MFCC	42.3	21.9	**7.4**	23.9
energy	24.2	17.7	3.6	15.2
spectral	33.2	10.0	1.9	15.1
MFCC + energy	46.5	25.5	**7.4**	26.5
MFCC + spectral	43.6	24.8	3.6	24.0
energy + spectral	37.1	22.9	3.2	21.1
MFCC + energy + spectral	44.5	27.4	4.8	25.6
best 400 features	**51.9**	**28.4**	4.2	**28.2**
best LLDs	38.1	21.6	4.5	21.4

with the best scored features, these functionals were also among the most common functionals in the best 400 features.

Table 3.5 shows the results for the experiments using the test set containing 155 subjects. In general, the same trends can be observed as on the development set. The combination of MFCCs and energy features was better than the combination of all three feature groups. Using the best 400 features which were selected as described in Section 3.3.4, the best result was achieved with an average identification accuracy of 28.2 %. This is a 10 % relative improvement compared to using the whole feature set.

3.3.5 Multimodal Fusion

Although gait-based person identification using audio features cannot yet reach the performance of video systems, it has the potential of improving such a system through multimodal fusion. Since the TUM GAID database provides video, depth, and audio data, it is perfectly suited to perform experiments for multimodal fusion. Therefore, in the following, results are shown for the separate modalities (video, depth, audio[2]) as well as for all possible combinations.

A detailed description of the employed methods for video and depth feature extraction as well as for classification is given in [104]. From the video data, the

[2]Thanks to Martin Hofmann for providing the features and recognition results for the video and depth modalities.

Table 3.6: Multimodal fusion results on the test set (155 subjects), for all combinations of audio, video, and depth features. Shown is the identification accuracy for the three setups $\mathcal{N}$ (normal walking), $\mathcal{B}$ (backpack), and $\mathcal{S}$ (shoe covers) together with the average (Avg.).

	Condition			
Modality	$\mathcal{N}$	$\mathcal{B}$	$\mathcal{S}$	**Avg.**
video	99.1	27.1	52.6	59.6
depth	99.0	40.3	**96.1**	78.5
audio	44.5	27.4	4.8	25.6
video + depth	**99.4**	51.3	94.8	81.8
video + audio	**99.4**	45.2	44.2	62.9
depth + audio	99.0	52.3	95.5	82.3
video + depth + audio	**99.4**	**59.4**	94.5	**84.4**

gait energy image (GEI) [95] is extracted. The GEI averages the information (in the form of silhouettes) of each frame within a complete gait cycle. The depth gradient histogram energy image, which is motivated by the concept of histograms of oriented gradients [37], is used for feature extraction from depth data. It is similar to the GEI, but, instead of only processing simple silhouettes, it makes use of the edges and depth gradients within the silhouettes. In this way, it exploits the available depth information. The video and depth features are processed by principal component analysis and linear discriminant analysis (LDA) before being fed to a 1-nearest neighbour classifier with a cosine distance measure. For simplicity, the baseline audio system (using the whole audio feature set) was selected for the fusion experiments.

For combining multiple modalities, score-level fusion was applied. The video and depth scores were represented by the normalised cosine distances, and the audio scores were obtained from the pairwise voting of the multiclass SVM (majority scores are normalised and transformed to distances). Scores were combined with the sum rule (i. e. added together), and the scores of video, depth, and audio features were weighted in the ratio 2:4:1. Finally, classification was performed with a nearest neighbour classifier using these fusion scores.

The experimental results of all possible combinations of multimodal fusion are listed in Table 3.6. First, it can be seen that, unsurprisingly, the video and depth modalities achieved much better results than the system with only audio features. However, in situations where the video system struggled, such as the backpack condition (27.1 %), the audio system was able to keep up (27.4 %). Furthermore, the performance in the backpack condition was significantly improved by combining the three modalities, compared to the single-modality results. Moreover, the best overall result (averaged over all recording conditions) was also obtained by combining all three modalities. This shows that gait-based person identification using only audio

features is a much more challenging task than with video or depth information; however, it has the potential to improve the overall system performance when all modalities are combined.

3.3.6 Conclusions

In this section, an extensive feature analysis for acoustic gait-based person identification was presented. Using the development set of the TUM GAID database, suitable features were analysed and selected from a large candidate feature set. Out of all 1 625 features, a subset of 400 features was selected with a wrapper-based feature selection technique, which led to the best recognition results while at the same time reducing the system complexity. This feature set contains features from all three tested feature groups (energy, spectral, and cepstral). When features were examined independently of each other, the three energy features, spectral kurtosis, flux, and skewness as well as the first MFCC coefficient were found to be relevant for acoustic gait-based person identification. The best results were achieved on the normal recordings ($\mathcal{N}$ condition), while wearing a backpack ($\mathcal{B}$) or shoe covers ($\mathcal{S}$) influenced the identification accuracy in a negative way. Furthermore, it was shown how audio features are capable of improving a video-based gait recognition system through multimodal fusion.

3.4 Acoustic Gait-based Person Identification using HMMs

A system for static classification of walking persons was presented in the last section. While a certain amount of dynamics was modelled in the features, this system is not capable of a dynamic classification of gait sequences. The contribution of this section is a system for acoustic gait-based person identification using a dynamic classifier. HMMs are a dynamic classifier (an introduction is given in Section 2.2.1.2 of this thesis) that is capable of modelling the dynamic behaviour of step sounds. In [236], cyclic HMMs were applied for animal sound classification. The cyclic model topology proved to be efficient for modelling the repetitive structure of these sounds. The system proposed in this thesis uses Mel-frequency cepstral coefficients (MFCCs) as audio features and HMMs with a cyclic topology for classification, in order to model the dynamics of gait patterns. With the cyclic topology, one pass through the model corresponds to half a gait cycle (containing one step). Thus, it is expected that the system is capable of detecting the individual steps in a recording and using them for person identification. Experiments were conducted using the TUM GAID database. The experimental results show that the developed system achieves large improvements in the recognition rates compared to the static classification system presented in the last section.

3.4.1 System Description

The proposed system uses HMMs for classification. Each individual subject is modelled by one HMM. While the starting point of the system was to employ settings from simple word-based speech recognition methods, the system properties were modified and improved to fit to the problem of acoustic gait recognition. In the experimental section, it is shown how these modifications improved the system performance.

3.4.1.1 Audio Features

Section 3.3 explored the suitability of different audio features for the addressed problem. Adding and selecting relevant features improved the average identification accuracy from 23.9 % (only MFCCs) to 28.2 %. In the present section, the focus is not on the front-end processing but rather on the back-end recognition system. Therefore the front-end is kept fixed to using only MFCCs. MFCC features are extracted in the standard configuration: MFCCs $0 - 12$ including their delta and delta-delta coefficients, computed every 10 ms from a 25 ms Hamming window, resulting in 39 features in total. While the database provides four-channel audio recordings, features are extracted from monaural recordings, which are obtained by averaging over the four channels. In addition, slight improvements were obtained by processing the audio features with principal component analysis, without reducing the number of components. Here, the transformations are computed only on the enrolment data and applied on both the enrolment and identification data.

Figure 3.4 shows the spectrograms and corresponding first MFCC coefficients for two exemplary recordings ($\mathcal{N}$ setup) of two different subjects. The spectrograms reveal a considerable static background noise, which is due to the recording environment. Several spectral peaks can be identified which correspond to the footsteps and the sounds between the steps, which are mostly made by the legs of the trousers rubbing against each other. In the plot of the MFCCs, the temporal position of the steps are marked. The behaviour of the MFCCs indicates that these features are useful to detect the position of the steps and to distinguish between different persons.

3.4.1.2 HMM Identification System

The starting point in this work is a simple HMM system that can be compared to a whole-word recognition system (each subject representing one *word*) in speech recognition. Each subject in the dataset is represented by one HMM. The models are designed with a linear left-right topology. With such a model topology, the HMM has to pass through all of its states sequentially without skipping a state. Before introducing an appropriate step modelling method, which is described later, each recording containing several steps is modelled by one pass through an HMM.

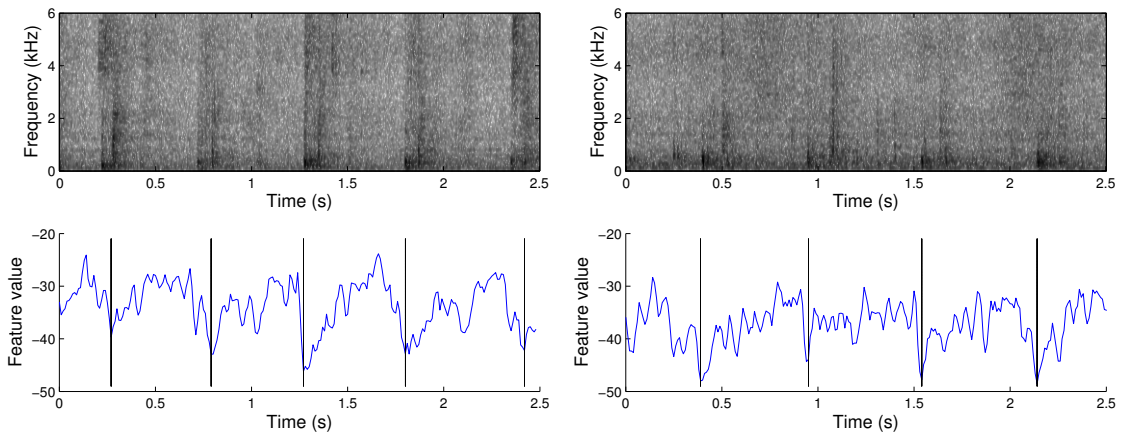


Figure 3.4: Spectrograms (top) and corresponding first MFCC coefficients (bottom), each for a normal-type recording of two different subjects. The temporal position of footsteps is marked with a vertical line.

As a result, rather large numbers of states (generally more than ten) are required to be able to model the dynamic sequence of sounds during walking.

In a standard HMM system, the observations are modelled with a mixture of Gaussians. However, the first experiments showed that better results are obtained with HMMs having a single Gaussian state model, as the amount of training data is very small and hence probably not sufficient to train a more fine-grained distribution of the features. Another reason could be that a higher number of components leads to overfitting, modelling also the noise in the recordings.

During decoding, a grammar controls the possible recognition output. The simplest employed grammar follows the basic HMM system setup where exactly one pass through a model is allowed for each recording. A multi-step grammar was introduced to let the system automatically segment the recording; any number of repetitions of the same model (subject) are allowed.

In order to train the HMMs to model the individual steps, an approach using a cyclic HMM topology was employed as described in the following. In the basic HMM system, each recording (containing several gait cycles) is modelled by one pass through the HMM. The strategy of representing each gait cycle separately by one pass through all states of the HMM is better suited to model the observations. The two halves of each gait cycle can be considered to be equivalent (although in fact, there is a person-dependent asymetry [156]), and therefore, the system was designed to model half a gait cycle (containing one step) by each HMM. In this way, one pass through the HMM models the sounds of one step and adjacent sounds (produced by the arms and legs). This method of step modelling was implemented in the system configuration and training in the following way: the state transition matrix of each HMM has a left-right topology, and jumps from the last state to the

Table 3.7: HMM results on the development set (150 subjects) for different systems. Shown is the identification accuracy for the three setups $\mathcal{N}$ (normal walking), $\mathcal{B}$ (backpack), and $\mathcal{S}$ (shoe covers) together with the average (Avg.).

	Condition			
System	$\mathcal{N}$	$\mathcal{B}$	$\mathcal{S}$	**Avg.**
basic HMM	53.3	30.7	7.0	28.2
+ multi-step decoding	56.3	31.3	7.3	31.6
+ principal component analysis	57.7	34.3	**9.7**	33.9
+ step modelling	**69.7**	**44.7**	9.3	**41.2**

first state are allowed. Models were trained with embedded re-estimation, where the number of steps was known from the transcription of the training data. As a result, the position of the steps in the training data are automatically estimated during model training. Together with the introduced multi-step decoding grammar, the developed system is capable of detecting, segmenting, and recognising the steps occurring in the recordings.

3.4.2 Experimental Evaluation

Person identification experiments were performed wi th the TUM GAID database that was described in Section 3.2. The system was designed and tuned using the development set of the database. Finally, the test set was employed to evaluate the best system configuration. For all systems evaluated using the development set, 15 HMM states appeared to be the optimal configuration. In addition, the best results were obtained with six training iterations. For each system setup, experimental results (identification accuracy) are reported separately for the three different recording conditions (normal, backpack, and shoe covers). Furthermore, the average accuracy over these three conditions is given.

3.4.2.1 Development Set

Table 3.7 shows the results for different system configurations on the development set. The basic HMM system without explicit modelling of separate steps is the first evaluated system. In the normal recording condition, slightly more than half of the test samples were classified correctly. An accuracy of 28.2 % was obtained when averaging over the three different conditions, which serves as a baseline for further experiments.

The first step towards the improved recognition system was the introduction of a decoding grammar which allows the recognition of multiple sequential instances of the same subject in the recordings. This modification improved the average

Table 3.8: HMM results on the test set (155 subjects) for different systems, compared to previous studies. Shown is the identification accuracy for the three setups $\mathcal{N}$ (normal walking), $\mathcal{B}$ (backpack), and $\mathcal{S}$ (shoe covers) together with the average (Avg.).

	Condition			
Accuracy [%]	$\mathcal{N}$	$\mathcal{B}$	$\mathcal{S}$	**Avg.**
video (GEI) [104]	**99.4**	27.1	**52.6**	**59.7**
baseline SVM	44.5	27.4	4.8	25.6
SVM + feat. sel.	51.9	28.4	4.2	28.2
basic HMM	41.0	24.2	7.1	24.1
improved HMM	65.5	**36.5**	9.0	37.0

accuracy to 31.6 % (mostly due to improvements in the $\mathcal{N}$ setup). Applying principal component analysis to the features improved the accuracy for all three recording conditions. Finally, training the system to model each step by one pass through an HMM led to the largest improvement in accuracy. In the normal walking condition, more than two thirds of the samples were then identified correctly. The accuracy in the backpack walking condition was also greatly improved, whereas the performance in the shoe cover condition remained largely unaffected. While the improvements obtained with the multi-step grammar and principal component analysis are not significant, improved step modelling led to a significant improvement (compared to the basic HMM) in the $\mathcal{N}$, and $\mathcal{B}$ conditions and for the average accuracy (evaluated with a one-tailed z-test with a significance level of $p = 0.05$).

The system's ability to correctly detect the individual steps was examined with a simple analysis. To this end, the best-performing developed system (last row in Table 3.7) was used. For the test samples of the normal walking conditions, the number of steps detected by the system was observed. The average number of steps in these test recordings is 5.3, while the system predicted 4.3 steps on average. For correctly identified *subjects*, the average number of predicted *steps* was 5.0, while for incorrectly identified subjects it was 3.5. This shows that when the subjects are identified correctly, the step segmentation works very well.

3.4.2.2 Test Set

Table 3.8 shows the results on the test set for the baseline system and the best system configuration. For comparison, results from previous studies on the same dataset are listed, including a video processing method as well as the results obtained with SVMs as presented in Section 3.3. The first row shows results of a state-of-the-art gait recognition method working with video data, namely the GEI [104]. This method achieves almost perfect results in the normal walking condition, while espe-

cially the backpack and the shoe variation constitute a real difficulty for the system (59.7 % on average). However, these results have to be interpreted carefully, since the GEI utilises mainly the appearance (the silhouette of a person) and not the behaviour (the gait pattern). A first audio-domain baseline system used a large set of different audio features (1 625 static features per recording) and SVMs for classification (second row), as presented in Section 3.3. Naturally, the task is much more difficult when dealing only with audio data (average accuracy 25.6 %). However, this system can compete with the GEI in the backpack recording variation. The SVM system was improved by employing a feature selection technique to choose relevant features for the task, obtaining an average identification accuracy of 28.2 %. Now, with the basic HMM setup, the resulting accuracy of 24.1 % is comparable to the baseline SVM system. The methods introduced in this study (primarily modelling each step separately during model training and decoding) were able to bring a large improvement, reaching 37.0 %. In the $\mathcal{N}$ and $\mathcal{B}$ recording conditions, the accuracy was improved significantly (compared to the baseline SVM) by more than one third. The accuracy of the video processing method (GEI) in the backpack recording condition was surpassed by 26 % relatively. Compared to the previous best-performing audio system (the SVM system including feature selection) the average accuracy was improved by 24 % relatively (significant in all recording conditions).

3.4.3 Conclusions

This section introduced a model-based system for recognising people from walking sounds. The system used HMMs in a cyclic topology to automatically segment the recordings according to separate steps. Experiments were conducted using the TUM GAID database. The results showed that a basic HMM system (without explicit modelling of separate steps) achieved a similar performance as the SVM system presented in the first part of this chapter. Improving the system with the methods introduced in this thesis resulted in large performance gains in identification accuracy. With this system, the gait cycles are properly modelled using cyclic HMMs. One pass through the model covers the sound of one step and adjacent sounds, which are mainly produced by the person's arms and legs. Thus, it is clear that the backpack or shoe cover variation influence the identification performance in a negative way. However, when identification experiments were carried out with the same walking style and shoe type as the model was trained with (the normal walking condition), almost two thirds of the subjects were identified correctly from the test set containing 155 individuals. While this demonstrates the capability and potential of the developed system, it has to be noted that this setup corresponds to a re-identification scenario, in which, for example, a person can be recognised just a few moments after he/she was enrolled in the system.

3.5 Chapter Summary

The goal of this chapter was to present a system that is capable of identifying humans from step sounds. Two different methods were proposed and evaluated with a database of more than 300 subjects.

Moderate identification accuracies were obtained with a static classifier, whereby the relevance of different audio features for this task was investigated. Large improvements were obtained by employing a dynamic classifier in the form of an HMM system that models the distinct gait cycles. The 155 subjects from the test set of the TUM GAID database were identified correctly in almost two thirds of the trials in the normal walking condition, which corresponds to a re-identification scenario where the external conditions (including shoes and clothes) remain unchanged. Further improvements are expected when the number of subjects is decreased or when the top five ranked subjects are looked at. In addition, it has to be noted that the test trials contain only a small number of steps (typically three to five). The identification performance can also be improved when more steps are considered [3]. Thus, it can be concluded that the proposed methods take big steps in the direction of a system that is good enough to be used in a practical application, such as a smart home.

Given the challenging but application-friendly enrolment of only four examples per walking subject and in order to improve the robustness of the system, adopting approaches from speaker recognition such as creation of models through adaption from a background model [182] could be a promising strategy in the future. Furthermore, the system's robustness to variations should be addressed. This includes better coping with the backpack and shoe cover recording conditions. In addition, the TUM GAID database contains a set of subjects with recordings made on two different dates in time (with three months in between). Therewith, the influence of changing types of shoes and clothes as well as possibly higher variation of the walking style on the system performance can be evaluated. In order to improve the system in this direction, approaches to address session variability known from speaker recognition (such as joint factor analysis [121]) could be tested, as well as methods for model adaptation or feature transformation adopted from speech recognition systems.

4

Speaker Diarization

The topic of this chapter is speaker diarization, a variant of speaker recognition in which no prior knowledge of the speakers is available. This study particularly addresses the problem of overlapping speech. Different methods for detecting and handling overlapping speech in a speaker diarization system are proposed. Overlap means that two or more speakers are speaking at the same time, which happens very often in natural conversational speech. Furthermore, an algorithm for speaker diarization capable of online processing is presented.

4.1 Introduction

Speaker diarization is the task of answering the question 'Who speaks when?' [220, 5, 149]. It can be regarded as a subdomain of the problem of audio diarization [183]. General audio diarization systems aim segmenting an audio recording into homogenous regions and to attribute them to the contributing sources. These sources can include speakers, but also music, background noise from the acoustic scene, or other sound sources. Speaker diarization systems, however, mostly deal with human speech with the goal of detecting all different speakers in an audio stream. The general problem is that, in contrast to traditional speaker recognition systems [124], the contributing speakers are initially unknown. Furthermore, an additional challenge is that the number of speakers is also regarded as unknown. Therefore, the system needs to identify the speakers and assign audio segments to them in an unsupervised manner. Figure 4.1 shows the workflow of a diarization system. An audio stream with an unknown number of previously unknown speakers is segmented into speaker-homogenous regions, and these regions are clustered together to represent individual speakers.

Applications of speaker diarization systems can be found in many fields of research of audio processing. Audio indexing and retrieval [144] aims at a thorough analysis of audio documents (e. g. broadcast recordings), resulting in a transcription

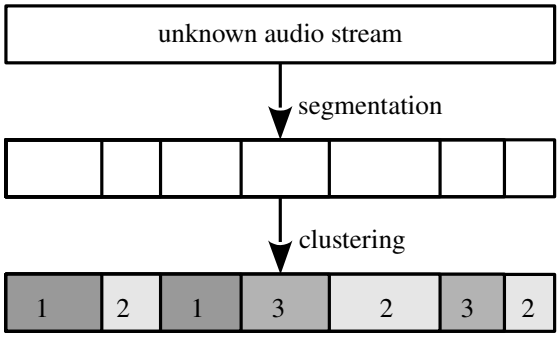


Figure 4.1: Workflow of a speaker diarization process.

of the spoken content, speaker segmentation and clustering, identification of known speakers, topic classification, and story segmentation. Speaker diarization is an important component here, helping to structure the audio document with regard to the individual speakers. Such an analysis of an audio document provides possibilities to summarise and browse the contents. For example, an automatic summary and analysis of a meeting recorded can be produced by such a system.

The task of rich transcription has been the subject of comparative evaluation campaigns of the National Institute of Standards and Technology (NIST)[1], in which speaker diarization has been evaluated several times. Such studies help to compare different approaches and methods by providing standardised evaluation databases and protocols. As a result, the state of the art is advanced by the efforts of different research groups. While initially the NIST rich transcription evaluations were carried out using broadcast news and telephone recordings, later, the domain of meeting speech was addressed. In all of those domains, different persons contribute to the discussion. Telephone conversations demonstrate the most simple scenario, in which normally only two speakers are involved. Broadcast news recordings are characterised by a well-structured setup. However, background music or other sound sources complicate the diarization process. Meeting recordings contain highly natural and spontaneous speech and additionally, various interfering sounds produced by the meeting participants. Audio-visual analysis of meeting recordings has also been the subject of several research projects, such as the Augmented Multi-party Interaction (AMI) project [27].

Another application of speaker diarization is to use it as a preprocessing step to speaker adaptation for speech recognition. Compared to speaker-independent speech recognition systems, training speaker-dependent models or adapting the speaker-independent models to the specific speaker's characteristics can greatly improve the recognition performance [135]. With speaker-adaptive training [4], speaker-dependent models can be applied in a speaker-independent recognition system. A feature

[1] http://www.itl.nist.gov/iad/mig/tests/rt/, last accessed in April 2014

transformation is estimated for each speaker in batch processing. For the speakers in the test set, this transformation has to be estimated from adaptation or test data. This is where speaker diarization plays an important role. A speaker diarization system can segment and cluster unstructured test data to group adaptation data by individual speakers [205].

The field of speaker diarization has increasingly gained popularity in recent years. Initiated by the NIST rich transcription evaluations mentioned above and by efforts in the field of meeting analysis, the state of the art has progressed to a point where small problems can severely degrade the system performance. Open issues include linguistic influences on speaker diarization [22], the problem of overlapping speech [109], or methods capable of online processing [145].

In this chapter, first an introduction to the field of speaker diarization is given by outlining established approaches and methods, presenting commonly used databases, and explaining evaluation measures in Section 4.2. Following this, Sections 4.3 and 4.4 address the issues of detecting and handling overlapping speech, respectively. Finally, a method for online speaker diarization is presented in Section 4.5, before concluding this chapter.

The different approaches for overlap detection presented in this work are based on previously published results. Section 4.3.4 is based on [68], Sections 4.3.5 and 4.4 are based on [71], Section 4.3.6 is based on [63], and Section 4.3.7 is based on [64]. Finally, Section 4.5 is based on [62].

4.2 Fundamentals and Methods

This section gives an introduction to the field of speaker diarization by describing state-of-the-art methods, evaluation techniques, and available databases, as well as some of the remaining open questions in speaker diarization technology.

4.2.1 Speaker Diarization Methods

An early study in the direction of speaker diarization was carried out by Chen and Gopalakrishnan [30]. The objective of their system was to detect change points in broadcast news recordings and to cluster the resulting segments. Both tasks can be accomplished using the Bayesian information criterion [199], which finds suitable change points and appropriate speaker clusters for merging. Most diarization systems are based on the same principle. An overview of different systems is given in [5]. Figure 4.2 shows the structure of such a typical diarization system. First, preprocessing takes place. Most often, this step consists of voice activity detection (VAD) [247], which aims at detecting all regions in the audio recording containing human speech. Additionally, methods for filtering noise are commonly applied in the preprocessing step. In the following feature extraction step, audio features that are suitable for

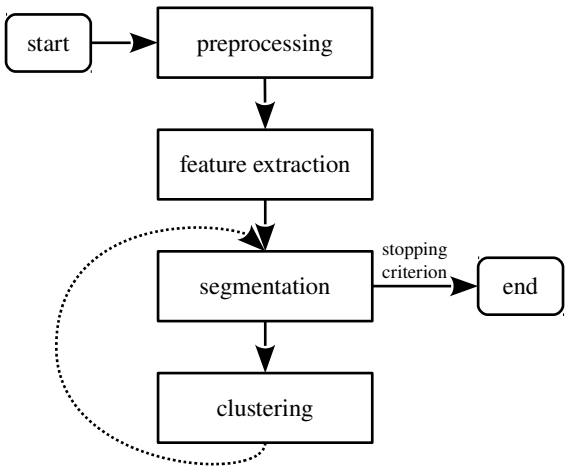


Figure 4.2: Structure of a speaker diarization system.

processing in the later steps are computed. Most widespread features are MFCCs, while some studies propose to include other features, such as prosodic or other long-term features [50]. Prosodic features can capture speaker-specific information about intonation, timing or loudness. Furthermore, when multi-channel audio recordings are available, spatial features exploiting the inter-channel time differences can be taken into account [162]. On the other hand, most diarization systems work with features extracted from single-channel audio. For that reason, typically, delay-and-sum beamforming is applied to convert the multi-channel recordings to a single channel [247]. In the case that the preprocessing steps (e. g. VAD) use the same features as the actual diarization process, the order of preprocessing and feature extraction is changed. The segmentation step that follows the feature extraction aims at dividing the audio recording into speaker-homogenous regions (cf. Figure 4.1). In its simplest form, the initial segmentation directly uses the speech segments resulting from the VAD module. Other variants include uniform initialisation, whereby the recording is segmented into small segments of equal length [5]. The most important part of a diarization system is the next step, clustering. As a result of the clustering step, segments that belong to the same speaker are grouped together, as can also be seen in Figure 4.1. In bottom-up diarization approaches (e. g. [247, 154]), agglomerative hierarchical clustering [41] is used to successively merge the closest clusters until an optimal number of clusters is reached. Here, the choice of initial clusters has a strong influence on the system performance. Each cluster is usually represented by a Gaussian mixture model (GMM) and closest clusters are found by appropriate distance metrics. After each clustering step, a new model is trained for the merged cluster and the audio data are re-segmented with the new models. A stopping criterion is then evaluated in order to decide whether the optimal number of clusters has already been reached. Alternatively, top-down approaches (e. g. [48, 147])

start by grouping all segments into one cluster. Gradually, more and more speaker clusters are created, again, until an optimal number of clusters is reached. The study presented in [43] investigates the differences between top-down and bottom-up diarization systems, leading to the conclusion that bottom-up systems produce more discriminative speaker models while top-down systems are less vulnerable to nuisance variation.

There are several toolkits available with different implementations of methods for speaker diarization. For example, the LIUM_SpkDiarization toolkit [190] is optimised for radio or television shows and uses a bottom-up clustering component employing agglomerative hierarchical clustering with the Bayesian information criterion. This toolkit was developed for the French ESTER2 evaluation campaign in 2008 [52], where it obtained the best results in the task of speaker diarization. The SHoUT toolkit [110] developed by Marijn Huijbregts also contains methods for speaker diarization. Another widely used toolkit that contains implementations of speaker diarization algorithms is the ALIZE/LIA_RAL toolkit [131]. The speaker diarization experiments in the present thesis have been performed with a system based on the latter.

4.2.2 The Diarization Error Rate

The performance of a speaker diarization system can be measured by the diarization error rate (DER),

$$\mathrm{DER} = \frac{\sum_{all\ \ segs} \mathrm{dur}(seg) \cdot \Big(\max\big(N_{ref}(seg), N_{sys}(seg)\big) - N_{corr}(seg) \Big)}{\sum_{all\ \ segs} \mathrm{dur}(seg) \cdot N_{ref}(seg)}, \tag{4.1}$$

which counts the fraction of speaker time that is not attributed correctly to a speaker [155]. The terms N_{ref}, N_{sys}, and N_{corr} represent the numbers of speakers in a segment, as transcribed in the reference, hypothesised by the system, and correctly detected by the system, respectively, while $\mathrm{dur}(seg)$ is the duration of a segment. In the denominator, the length of all speech segments is added up, while overlapping speech is not ignored: segments with more than one active speaker are weighted by the respective number of speakers. The numerator counts all errors of the system, which can be divided into three groups: missed speaker times, false alarms, and speaker errors. Missed speaker times and false alarms can often be traced back to the VAD component, which either detected too few or too many speech segments. Additionally, in the case that the system detects only one speaker during a segment of overlapping speech (and most state-of-the-art diarization systems can actually detect only one speaker at a time), a missed speaker error is

observed, since any overlapping speakers are not detected. Speaker errors are made when the system misclassifies a speaker. To account for imprecise human annotation, normally, a 'forgiveness' collar of 0.25 seconds (in both directions) is used around segment boundaries. Since speaker diarization is an unsupervised recognition process, the detected speaker identities cannot directly be compared to the reference speaker identities. An optimum one-to-one mapping from detected speakers to reference speakers is computed, minimising the DER. Thus, detecting the correct number of distinct speakers is a very important challenge for the system.

4.2.3 Databases

Databases which are commonly used for the evaluation of speaker diarization systems can primarily be divided into three categories: telephone conversations, broadcast recordings, and meeting recordings. This distribution results directly from the different applications. Over the years, the subject of research became more difficult by transitioning from telephone conversations over broadcast recordings to meeting speech. The NIST rich transcription evaluation project was already mentioned. From 2002 to 2009, nine rich transcription evaluation campaigns were conducted by the NIST. Most of them also included a speaker diarization task, whereby appropriate evaluation data were always provided, each time in English language. The first speaker diarization evaluations were carried out with broadcast news and conversational telephone speech, and later on, the focus was set on meeting speech. Furthermore, the NIST conducted speaker recognition evaluations. The employed databases have also found use for speaker diarization systems. For example in [6], two-speaker telephone dialogues from the NIST 2005 speaker recognition evaluation were used to test a diarization system. Other corpora that are used for speaker diarization include the 1996 HUB-4 broadcast news corpus [53], which was employed for example in [163] for online speaker diarization. In the Augmented Multi-party Interaction (AMI) project, the AMI meeting corpus [27] was created. This corpus contains audio-visual recordings of meetings and was used for multimodal meeting analysis, including speaker diarization. Recently, the ETAPE corpus was provided in the context of the ETAPE evaluation campaign, featuring speaker diarization as well as a multiple speaker (overlap) detection task [90]. This database consists of French radio and television broadcast data. The recently released Sheffield Wargames Corpus [46] contains recordings of native English speakers speaking naturally while playing a table-top game. This database contains particularly large amounts of overlapping speech.

4.2.4 Open Issues

While the state of the art in speaker diarization has made substantial progress in recent years, a number of open issues and possible directions for research persist. One

of these open problems is the impact of linguistic content on speaker diarization, which influences the clustering step. The goal of the clustering process in a diarization system is to converge to different clusters for individual speakers, whereas one typical malfunction is to create clusters for other acoustic classes. In this way, the resulting clusters represent the spoken content instead of the speakers. In [22], the influence of linguistic content on speaker diarization performance is examined. It is shown that bottom-up diarization systems are more error-prone to linguistic content compared to top-down systems. The clusters produced by the top-down system are better normalised against phone variations. Following this study, a method for addressing the problem of linguistic influence is presented in [21]. In analogy to speaker-adaptive training in speech recognition, phone-adaptive training is used to reduce the influence of phonetic variation within speaker clusters. Other studies addressing the linguistic information in diarization systems are presented in [29, 261]. Additional open problems for the clustering component of a diarization system are proper methods for cluster initialisation or cluster purification.

Another critical issue, which will be addressed in detail in this thesis, is the presence of overlapping speech. Several studies show that overlapping speech constitutes a major source of error for most speaker diarization systems [47, 108]. In the clustering process, impure speaker models are created which include speech from multiple speakers, since all speech material is aggregated in the clusters. These corrupt models will eventually lead to higher error rates in the segmentation process. In addition, typical diarization systems can recognise only one speaker at a time. This means that all segments of overlapping speech will count towards the missed speaker error, because at most one speaker can be detected correctly. Properly detecting and handling overlapping speech can bring large improvements to diarization systems. In the next section, several solutions to this problem are presented.

Section 4.5 of this chapter introduces an approach for online speaker diarization. Most current systems work in offline mode, which means that the full audio recording has to be present before the diarization process can be started. For several applications (e. g. a dialogue system that needs to react to the speakers), an online diarization system is required. The goal of such a system is to analyse an ongoing audio stream, keep track of the number of speakers, and to create new models as new speakers appear.

4.3 Detection of Overlapping Speech

Overlapping speech is a severe problem for many speech processing applications [204, 203]. For example, it is a problem for speech recognition systems for conversational speech [25]. In general, overlap can degrade the performance of any speech processing system that assumes only one speaker to be active. The presence of overlapping speech is especially a difficult problem for speaker diarization systems. Overlap

appears particularly in spontaneous conversational speech, which is the regular domain for speaker diarization systems. State-of-the-art diarization systems are mostly capable of detecting only a single active speaker, which leads to missed speaker errors. In addition, overlapping speech results in corrupt speaker models during the clustering process.

The earliest prior studies on this topic analysed the general influence of overlapping speech on diarization performance [109, 47, 161]. A more detailed analysis of the influence of different system components on the diarization performance is given in [108]. Using a diarization system where several parts are replaced by 'oracle' components resulted in a quantitative analysis of the contribution to the DER of errors made by different parts of the system. The contribution of overlap to the DER due to missed speaker errors is relatively easy to determine. In a single-speaker diarization system, the amount of overlapping speech in the evaluation data will directly translate into missed speaker errors: in every segment of overlapping speech, only one speaker is detected. Therefore, any overlapping speakers are missed. The study presented in [108] used two different datasets for evaluation, both containing recordings from NIST rich transcription evaluations. In those experiments, the proportion of missed speaker errors due to overlapping speech was 21.21 % and 18.08 % of the total DER, for the two evaluation sets, respectively. By appropriately detecting these segments and adding labels for the overlapping speakers, these errors could be mitigated.

The other spot where overlap leads to system errors is speaker modelling. The clustering part of a diarization system assigns each speech segment in a recording to one of the clusters. Using the allocated speech data, the speaker models are trained. All segments of overlapping speech will thus lead to corrupt speaker models, this again leads to a degraded speaker classification performance. The exact influence of this effect on the DER is difficult to quantify. In [206], a diarization experiment was performed where optimal speaker models were trained using data from only one speaker per model. This corresponds to a supervised speaker identification experiment or to the case where the clustering step works perfectly, resulting in 100 % pure clusters. Furthermore, this means that no overlapping speech is included during model training. Compared to the baseline DER of 28.09 % (using data from previous NIST rich transcription evaluations), the result of this experiment was a DER reduction to 17.29 %. A part of this improvement could be obtained in a real system by detecting and excluding overlap segments prior to the clustering step. In addition, labelling all overlap segments using the reference transcription (by adding labels for the overlapping speakers) reduced the DER to 5.87 %. Again, these results illustrate the potential of an overlap detection system to improve diarization systems.

Two strategies for handling overlapping speech can easily be derived from the two mentioned problems that are posed by overlapping speech. First, detected segments of overlapping speech should be removed prior to the clustering step. After iterative model training and refining, the overlap segments should be put back into the system

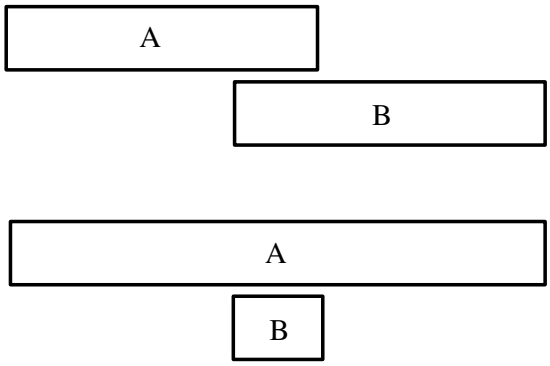


Figure 4.3: Different types of overlapping speech between the speakers A and B. Top: turn-taking, bottom: interruption.

and multiple speaker labels should be assigned to these segments in order to detect all overlapping speakers. The successful detection of overlap thus has the potential to improve the robustness of a speaker diarization system under realistic conditions. Accordingly, there is an increasing effort within the community (a discussion on related work follows in Section 4.3.2) to develop systems that detect and handle overlapping speech appropriately. New algorithms are needed to reliably detect segments of overlap. Overlap detection is an unsolved problem and the focus of this thesis.

Another application of an overlap detection system, besides speaker diarization, is conflict detection. In [123], conversational features for conflict detection are derived from an overlap detection system, assuming that conflicts in conversations can be predicted by conversational overlap patterns between opposing groups. Going in the same direction, it could be possible to use information from an overlap detection system to characterise speakers, for example whether they participate in the conversation more aggressively or rather more passively. Furthermore, there is related, prior work in speaker recognition [210], in which overlap detection is used as a preprocessing step for speaker recognition.

4.3.1 Overlapping Speech in Human Conversations

In spontaneous, conversational speech, it happens very often that two or more speakers are speaking simultaneously [203]. A straightforward method of categorising segments of overlapping speech (from two speakers) takes into account the neighbouring speakers. Figure 4.3 shows two different types of overlapping speech. In cases where the speaker preceding and following an overlap segment is the same (lower part of Figure 4.3), this overlap segment might be a short interruption or backchannel utterance (such as 'uh-huh' or 'mm-hmm') from the second speaker, or a failed attempt to steal the speaking turn. The other case is when the speakers before and after

an overlap segment are the same two speakers that are active during the overlap segment (upper part of Figure 4.3). In this situation, the second speaker took over the speaking turn from the first speaker. It could further be differentiated whether the first speaker was able to finish his speaking turn (making this situation almost like a normal speaker turn change), or whether the first speaker was interrupted in the middle of his speaker turn.

Studies in the literature report different numbers for the proportion of overlapping speech, which depend on the examined databases and especially depend on the style of conversation (e. g. multi-party discussion, interview, or dialogue). The numbers range from 5 % [136] to over 50 % [218]. Another analysis in [96] found that overlapping speech made up up to 40 % of inter-speaker intervals in three databases of conversational speech.

In [1], a detailed analysis of overlapping speech in political interviews is given. The authors analysed a corpus consisting of television shows in which a politician is interviewed by three journalists and a chairman. On average, the database contained three to four overlaps per minute and the global duration of overlap was relatively low (below 5 % of the data). Segments of overlap contained 2.5 words on average. The study introduced four categories of overlapping speech, which were classified as being either intrusive or non-intrusive: complementary (intrusive), backchannel (non-intrusive), turn stealing (intrusive), and anticipated turn taking (non-intrusive). The 'complementary' tag was introduced for overlaps which aim at complementing the main speaker's utterance. Backchannel and turn stealing overlaps each made up around one third of all overlap segments, while the other two types shared the rest of the overlap segments. The authors furthermore proposed measures to characterise speakers, depending on their speaking pattern concerning overlapping speech. These measures categorise speakers as either being more active/aggressive or more passive, with respect to the occurrence of overlap. With a system for detecting overlapping speech and identifying the involved speakers, such an analysis could be performed automatically, without the need of a time-consuming annotation of the data. Alternatively, prior knowledge about the involved speakers could be used to improve an overlap detection system.

The authors of [82] analysed different turn-taking cues, where overlap plays an important role. Backchannels and interruptions, along with turn switches, are distinguished as different categories of turn-taking phenomena. The authors examined the durational distribution of overlapping speech segments. They found that in the employed database, 60 % of the overlap segments had 200 ms or less of simultaneous speech.

Recent studies try to further analyse the nature of overlapping speech, revealing results about properties such as the duration and when it is likely to happen. These studies show that the occurrence of overlapping speech depends on many factors. In [81], backchannel utterances were further classified into eleven categories, where the most important distinctions are (a), whether the words convey acknowledgement

or agreement and (b), whether they introduce the beginning or end of a discourse segment. In the study presented in [118], backchannels were compared with agreements and it was found that backchannels are shorter in duration, have lower pitch and intensity, and are more likely to end with a rising intonation. In [132], the dynamics of overlap were analysed. The authors show that the durations of intervals of overlap are approximately log-normally distributed and that overlapping speakers tend to end speaking simultaneously less frequently than they start. The study presented in [240] investigates how onsets of overlapping speech are timed with respect to syllable boundaries. In [12], it was found that overlapping speakers tend to use different frequency bands. The authors of [83] found that interruptions (which are one type of overlapping speech) are likely to occur during or after speech with certain acoustic/prosodic properties. Such studies reveal important information about the nature of overlap (e. g. acoustic properties) that can be exploited to improve overlap detection systems.

4.3.2 Related Work on Overlap Detection and Handling

Independent of the actual overlap detection system, most studies reported in the literature use very similar approaches for handling detected overlap segments in the diarization system. One of the first proposed systems for overlap detection and handling uses a two-way strategy to process overlapping speech in a diarization system, namely, overlap exclusion and overlap labelling [16]. Most of the later studies also rely on these two steps. Overlap exclusion (the first step) means that detected overlap is excluded from the model creation phase. This ensures that the resulting speaker models are not corrupted by overlapping speech, whereby the quality of the speaker models is improved. Referring back to Figure 4.2, this technique is applied prior to the iterative procedure of segmentation and clustering. The second step (overlap labelling, also known as overlap attribution) is executed during the final segmentation step of the diarization system and has the goal of attributing the overlap segments to a speaker. To begin with, all detected overlap is re-introduced into the system prior to this final segmentation. This means that the decoding and segmentation is also performed for the overlap segments, leading to a detected speaker for these segments. In addition, the system is forced to detect a second speaker in these segments. In one variant, the neighbouring speakers (before and after the segment) are assigned to an overlap segment [107]. This may result in only one speaker being assigned to the segment. This method is based on the assumption that overlap is mostly the result of one speaker interrupting the other, and in this case, every overlap segment has two neighbouring speakers. One the other hand, the most straightforward strategy to detect two speakers is to choose those two with the highest posterior probabilities, as determined in the segmentation process [16]. This approach was predominantly used in the subsequent studies about overlap handling for speaker diarization (e. g. [254]). All approaches for overlap labelling assume

that at most two speakers are active at the same time. Of course, this depends on the actual application of the diarization system, and sometimes, more than two speakers may be active at the same time. However, in meeting recordings, which is the primary domain for speaker diarization systems, this assumption is true for the majority of overlapping speech (around 80 % of the time) [16]. With this strategy of attributing overlap to two speakers, every falsely detected overlap segment will lead to a false alarm contributing to the DER. Therefore, it is especially important to tune the overlap detection system towards a low false detection rate.

While there are not many different approaches for the handling of overlapping speech in diarization systems, more research was done towards improving the actual overlap detection systems. The first system for detection and handling of overlap in speaker diarization was presented in [16]. A hidden Markov model (HMM) system with three classes (nonspeech, speech, and overlapping speech) employing mainly spectral audio features was used to detect overlap. This system can be regarded as an extended VAD system, where the additional class of overlapping speech is added. It is a model-based system, using a statistical model for the three classes, trained on a separate training set. The processed audio recording is segmented using these models, leading to the detection of overlap segments. These overlap segments are subsequently processed by the diarization system using the methods described in the previous paragraph. Huijbregts *et al.* [107] report another model-based overlap detection approach. In their system, however, the overlap model is not trained on held-out data, but using data localised around speaker turns, assuming overlap to be present frequently around speaker turns. The three-class HMM system framework has prevailed and was used by most subsequent studies on overlap detection. What is subject of most research studies is the selection of appropriate audio features that are used to model overlapping speech. As already mentioned, the first system for overlap detection employed spectral audio features within the HMM framework [16]. In addition to MFCCs, root mean square energy, and linear predictive coding residual energy, the output of the diarization system was used in the form of the diarization posterior entropy, computed from the posterior probabilities of the diarization system. This study was extended in [17] and [18] which assessed the use of new features, including spectral flatness, the harmonic energy ratio, modulation spectrogram features, kurtosis, zero-crossing rate, and harmonicity. From a larger set of candidate features, these features had been selected as being most relevant for overlap detection. In these studies, significant improvements in the DER were achieved on a subset of the AMI corpus [27].

More recently, other features have been investigated for overlap detection. Prosodic features [254] were able to improve the performance of an HMM-based overlap detection system. The features were selected from a set containing pitch, intensity, and four formant features, together with long-term statistical characteristics extracted from 500 ms windows. An alternative approach which uses the output of a voice activity detection component and the silence distribution to detect overlap

was reported in [251]. This study was extended by exploiting long-term conversational features for overlap detection [250]. In [226], a system using convolutive non-negative sparse coding (CNSC) for overlap detection was first presented. The present thesis extends the techniques used therein.

The addition of multiple microphones opens more possibilities for overlap detection. Using personal close-talking microphones in a meeting room, overlap detection can be performed as a post-processing step after the segmentation of each individual microphone signal into speech and nonspeech – for example using cross-correlation analysis [168]. Recent work by Zelenák *et al.* [255, 256] reports an HMM system for overlap detection, using spatial/localisation features in addition to conventional acoustic features. The cross-correlation between signals of microphone pairs is exploited to derive spatial features, which are processed in the HMM system. Significant improvements in overlap detection performance are reported. The present study, however, concentrates on a single distant microphone with no localisation features available. This restriction to a single microphone makes the problem more challenging but at the same time, solutions become more versatile. The performance of each approach described above is modest at best and further work is needed to improve overlap detection performance.

4.3.3 Experimental Framework

4.3.3.1 Database

For all experiments on overlap detection reported in this thesis, the same database of audio recordings is employed, namely the AMI corpus of meeting recordings. This database was developed as part of the AMI project[2], aiming at improving meetings in order to be more productive by developing meeting browsers that allow users to find important information quickly. Thus, one goal of the project was to automatically analyse audio-visual meeting recordings. In the course of the project, a comprehensive database of meeting recordings was produced in different recording sites [27]. In total, this publicly available data set[3] consists of 100 hours of meeting recordings. These meetings were either occurring naturally or emanated from a scenario with the participants (mostly four persons) playing different roles according to a script.

A subset of ten recordings from the AMI corpus was used for the evaluation of the systems proposed in this study. The same subset was already used in previous studies by other authors [18]. As a consequence, the present results can directly be compared to the results of that study. For system development (e. g. for parameter tuning), six different recordings were used. Finally, for model training, a selection of 40 meeting recordings were employed. All employed meeting recordings are listed

[2]http://www.amiproject.org, last accessed in April 2014

[3]https://www.idiap.ch/dataset/ami, last accessed in April 2014

Table 4.1: Meeting recordings from the AMI evaluation corpus used for training, development, and testing.

Training set							
EN2001b	EN2002a	EN2002c	ES2005a	ES2005b	ES2006d	ES2007b	ES2008c
ES2010a	ES2010d	ES2013d	ES2014a	ES2015a	IB4003	IB4004	IB4010
IN1013	IS1000c	IS1002b	IS1002d	IS1003c	IS1003d	IS1004a	IS1005b
IS1009a	IS1009b	IS1009c	TS3003c	TS3004b	TS3005c	TS3007a	TS3007b
TS3008c	TS3009a	TS3010d	TS3011a	TS3011c	TS3011d	TS3012a	TS3012d
Development set							
EN2009c	ES2014c	IB4002	IN1016	IS1009d	TS3009d		
Test set							
EN2003a	EN2009b	ES2008a	ES2015d	IN1008	IN1012	IS1002c	IS1003b
IS1008b	TS3009c						

Table 4.2: Confusion matrix for a detection problem with positive (P) and negative (N) examples. The confusions are called true-positive (TP), false-negative (FN), false-positive (FP), and true-negative (TN).

		Prediction	
		P	N
Label	P	TP	FN
	N	FP	TN

in Table 4.1. The composition of recordings was selected to contain meetings from different recording sites. The length of the recordings in the test set varies between 17 and 57 minutes, and in total, the length of the test set is more than six hours. In all cases, only the single-channel far-field microphone recordings were considered. This is the most challenging scenario, but also the most realistic. The test set contains 13 % of nonspeech (with respect to the full duration) and 19 % of overlapping speech, with the rest (68 %) being single-speaker speech.

4.3.3.2 System Evaluation

The primary criteria for evaluating overlap detection systems employed in this study are precision and recall. These measures can be computed from the correct and wrong decisions of the system, as illustrated in Table 4.2. In a detection problem as in overlap detection, the decisions of the system can be categorised as *true-positive* (TP), *false-positive* (FP), *false-negative* (FN) or *true-negative* (TN), depending on the prediction of the system and the actual ground truth label. In the example of overlap detection, the *positive* label corresponds to overlapping speech, while *negative* equals single-speaker speech or nonspeech, and the system decisions are

counted frame-wise. The recall (Rec), also known as *true-positive rate* or *sensitivity*, is defined as

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4.2}$$

which is equal to the fraction of overlap frames that are correctly detected as overlap by the system. The precision (Pre) is defined as

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.3}$$

and corresponds to the fraction of all positively detected frames that are actually overlap. In addition, the overlap detection error (Err) is introduced,

$$\text{Err} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FN}} = 1 - 2 \cdot \text{Rec} + \frac{\text{Rec}}{\text{Pre}}, \tag{4.4}$$

which counts all errors made by the system as a fraction of all overlap frames. It is also possible to compute the error directly from the precision and recall values.

The overlap detection error can be interpreted in the following way. In the case of overlap handling, each TP decision of the system will reduce the DER (through a decreased missed speaker error) while each FP decision increases the DER (through an increased false alarm error). Therefore, in order to improve a diarization system through overlap labelling, the goal of the overlap detection system is to maximise the number of TP decisions while minimising the FP decisions. This can be combined in the objective function

$$\max\ (\text{TP} - \text{FP}), \tag{4.5}$$

which corresponds to

$$\max\ (\text{P} - \text{FN} - \text{FP}). \tag{4.6}$$

Since the amount of overlap P is constant, this can be simplified to

$$\max\ \big(-(\text{FN} + \text{FP})\big) \tag{4.7}$$

or

$$\min\ (\text{FN} + \text{FP}) \tag{4.8}$$

which corresponds to the sum of wrong decisions made by the system. This term is set in relation to the amount of overlap $\text{P} = \text{TP} + \text{FN}$, resulting in the overlap detection error

$$\min\ \frac{\text{FN} + \text{FP}}{\text{P}} = \min\ \text{Err}. \tag{4.9}$$

Thus, tuning the detection system to minimise this error enhances the potential improvement through overlap labelling. By normalising the error with the amount of overlap, it can directly be interpreted as the potential improvement in DER.

Through overlap detection and labelling, the missed speaker part of the DER is reduced, and thus, the error Err can directly be used to estimate the possible DER improvement. For example, an error of Err = 90 % means that the missed speaker part of the DER could be reduced to 90 % of its original value (a 10 % improvement). For the application of overlap exclusion, however, as much overlap as possible should be detected and excluded, which leads to the objective of achieving high recall performance.

Results for precision, recall and error are always reported in % in this study. Where possible, the experimental sections also provide figures plotting precision against recall, similar to a receiver operating characteristic curve. This is the case when the employed classifier uses a tunable parameter to control the trade-off between precision and recall, such as a threshold. In the following sections, different methods for detecting overlapping speech are presented and evaluated.

4.3.4 Overlap Detection using a Source Separation Method

This section presents a method for the application of convolutive non-negative sparse coding (CNSC) to the problem of overlap detection in the context of conference meetings and speaker diarization. CNSC is a signal separation technique based on non-negative matrix factorization (NMF), with the goal of decomposing a mixed signal into its bases and base activations. This technique is used here to project a mixed speaker signal onto separate speaker bases and hence to detect intervals of overlapping speech. While the decomposition of meeting recordings into speaker bases and activations was already introduced in a previous study [226], the novelty in this study is the suggestion of new overlap features which are derived from the speaker activations. Instead of using these features in an HMM system, however, a threshold-based overlap classifier is employed first. In a later section of this thesis, these features will be used in the HMM framework. The system is assessed using a subset of the AMI meeting corpus. Results are reported which are comparable to the state of the art, demonstrating the potential of this new approach to overlap detection. An analysis of system performance highlights the necessity of further work to address weaknesses in detecting particularly short segments of overlapping speech.

4.3.4.1 Convolutive Non-Negative Sparse Coding

The method of non-negative sparse coding (NSC) [133, 105] is an approach to represent a non-negative matrix as a linear combination of lower rank bases. The imposition of non-negative constraints allows only additive combinations in the representation. NSC can be seen as the combination of NMF and sparse coding, which is achieved by adding sparsity conditions to the update rules of NMF.

With NSC, a non-negative matrix $X \in \mathbb{R}^{\geq 0}_{M \times N}$ is represented as

$$X \approx WH \tag{4.10}$$

where $W \in \mathbb{R}^{\geq 0}_{M \times R}$ and $H \in \mathbb{R}^{\geq 0}_{R \times N}$ form the bases and base activations, respectively. These are learnt with the objective function of minimising the regularised least square error between the original matrix and the recomposition $\hat{X} = WH$:

$$(\hat{W}, \hat{H}) = \arg\min_{W,H} \|X - WH\|_F^2 + \lambda \sum_{ij} H_{ij}, \tag{4.11}$$

where λ is a regularisation parameter that controls the sparsity of the resulting representation and $\|\cdot\|_F$ denotes the Frobenius norm. This formulation, however, is not able to capture the correlation between adjacent frames in the data matrix X that is inherent in speech signals. A convolutive variant, known as CNSC [208] addresses this issue. The CNSC decomposition takes the form

$$X \approx \sum_{p=0}^{P-1} W_p \overset{p\rightarrow}{H}, \tag{4.12}$$

where P is the convolution range. The operators $\overset{p\rightarrow}{\cdot}$ and $\overset{p\leftarrow}{\cdot}$ are column shift operators which shift p columns of the matrix to the right and left, respectively.

The learning of bases and activations together according to Equation (4.11) is an optimisation problem that is solved by iteratively updating W and H until convergence using the following update rules [230]:

$$W_p = W_p \odot \frac{X \overset{p\rightarrow T}{H}}{\hat{X} \overset{p\rightarrow T}{H}} \tag{4.13}$$

$$H(p) = H \odot \frac{W_p^T \overset{p\leftarrow}{X}}{W_p^T \overset{p\leftarrow}{\hat{X}} + \lambda I} \tag{4.14}$$

$$H = \frac{1}{P} \sum_{p=0}^{P-1} H(p), \tag{4.15}$$

where I is an $R \times N$ identity matrix, $\odot$ is the Hadamard product, and the division of matrices is performed element-wise. After each update of W, its columns are normalised to unit vectors. This is an essential step in sparse coding since it en-

sures that W does not grow in an uncontrolled manner, and it encourages a sparse representation.

4.3.4.2 CNSC-based Overlap Detection

This section describes how the CNSC algorithm is applied to detect overlapping speech. CNSC bases are learnt for individual speakers such that an interval of overlapping speech can be decomposed into its underlying speaker components. This leads to a natural solution for the overlap detection problem. In this section, first, the CNSC-based decomposition of speech signals is described, followed by the introduction of a new frame-level approach to overlap detection.

CNSC bases W are learnt for each speaker in an audio document using spectral magnitude features extracted from segments of pure (non-overlapping) speech. The base patterns of each speaker $s = 1, \ldots, S$ are then concatenated together to create a global basis W^G,

$$W^G = [W_1, \ldots, W_S], \tag{4.16}$$

which includes the spectral patterns of all speakers. Spectral magnitude features across the whole audio document, including overlapping segments, are then decomposed at the frame level according to Equation (4.11) with W^G kept fixed and only H being updated to minimise the optimisation criterion.

The activations for any given frame and any given speaker therefore serve as an indication of that speaker's activity. While the activations H and corresponding bases W could be used to reconstruct or separate each speaker's contribution to the audio recording, the proposed system uses the activations H directly to detect each speaker's activity and hence segments of overlapping speech.

The sum of a speaker's activations is strongly correlated to the signal energy from that speaker since the bases W are normalised. Therefore, this sum is a good indicator of that speaker's activity. The energy for speaker s during frame t is estimated according to:

$$E_t(s) = \sum_{i \in I_s} H_{it} \tag{4.17}$$

where I_s represents the speaker-specific rows in H or the activations for speaker s.

Figure 4.4 (top) illustrates the CNSC activation energy against time for two speakers during a short interval from an exemplary meeting recording, where the speaker energy is calculated according to Equation (4.17). Ground-truth reference speaker activities are indicated beneath using the same line profile for corresponding speakers. It is seen that the CNSC activation energies are a suitable indicator of speaker activity: both speakers have high activation energy in the overlapping segment between 2 and 3 seconds.

In order to obtain overlap features from the activations, the speaker activation energies calculated as per Equation (4.17) are smoothed with a moving average filter

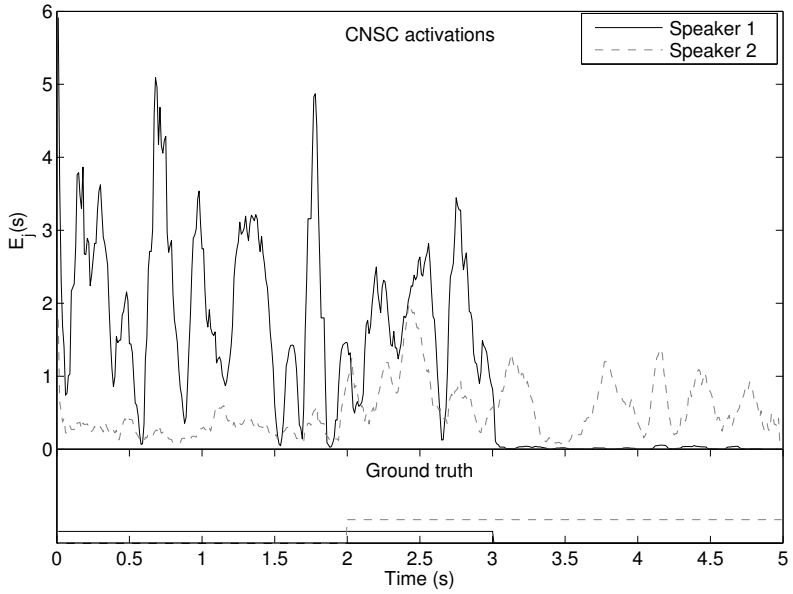


Figure 4.4: An illustration of the correlation between ground-truth speaker activity (bottom) and CNSC activation energies (top) for two speakers in a conversation containing an interval of overlapping speech [68].

and used to implement a frame-based overlap detector. It is based on the energy ratio ER for frame t, estimated as follows:

$$ER_t = \frac{E_t(\hat{s}_2)}{E_t(\hat{s}_1)} \tag{4.18}$$

where $\hat{s}_i$ denotes the speaker with the i^{th} highest energy. The energy ratio reflects the difference in activation energy for the two speakers who are deemed to be most active in the given frame. For overlapping segments the ratio is expected to be closer to unity, while for non-overlapping segments the ratio should be closer to zero. Since overlapping speech segments typically have more energy (they contain speech from multiple speakers), the total energy ET_t is estimated by summing up Equation (4.17) across all speakers and filtering out frames with low total energy. All frames with an energy ratio ER_t and total energy ET_t greater than empirically optimised thresholds δ_{ER} and δ_{ET} are deemed to contain overlapping speech.

In a previous study [226], the variance of speaker activation energy differences in a frame was used as a measure for detecting overlap. However, the energy ratio measure gives much better results when used in conjunction with the total energy threshold introduced in this thesis.

4.3.4.3 Experimental Evaluation

This section reports an assessment of the introduced overlap detection system using a subset of the AMI meeting corpus [27]. For training, tuning, and testing, the collection of meeting recordings as described in Section 4.3.3.1 was used.

In a practical speaker diarization scenario there are no speaker specific training data other than those within the audio recording itself. Consequently, the diarization system hypothesis itself must be used to estimate regions of clean speech for each speaker. Due to diarization errors, this speech material is not entirely pure, but is the only information available with which to learn speaker-specific base matrices for CNSC overlap detection. Any derived results are therefore dependent on the performance of the underlying speaker diarization system and the extraction of generalised results is thus troublesome.

In such scenarios it is typical to use oracle references to marginalise the impact of system elements that are not under direct observation and thus to minimise their influence on observed results. This approach was adopted here; the reference transcription was used to identify intervals of pure speech for each speaker. Accordingly, results presented in this section are independent of errors in an automatically derived speaker segmentation or diarization output and therefore the assessment focuses on CNSC alone. While such an approach does not necessarily give a reliable estimate of performance under practical conditions, it is noted that in a previous study [226], there was little difference in overlap detection performance between using reference segmentations and those obtained with a real speaker diarization system.

The parameters of the CNSC algorithm were tuned on a small, artificial two-speaker test set in which overlapping speech was manually created and controlled in order to better understand system behaviour and the influence of different parametrisations. These parameters were subsequently re-optimised on the AMI development data. The algorithm is applied to magnitude spectra computed on 40 ms windows with a window shift of 20 ms. CNSC speaker activations are calculated with speaker bases of dimension $R = 35$, a convolutional range of $P = 4$, and a sparseness parameter of $\lambda = 0.05$. The relatively large windows capture more discriminative speaker features whereas the use of small numbers of bases leads to more effective modelling and avoids overfitting.

Overlap detection performance was assessed using precision and recall statistics calculated at the frame level, as described in Section 4.3.3.2. Given that overlap detection can be applied in different processing steps of a typical speaker diarization system (overlap exclusion during clustering and overlap labelling during segmentation), different operating points with different precision and recall values are beneficial. Therefore, in addition to precise figures, the dynamic influence of the energy threshold δ_{ET} on the trade-off between precision and recall is also shown. A higher threshold identifies less overlap yielding lower recall but higher precision.

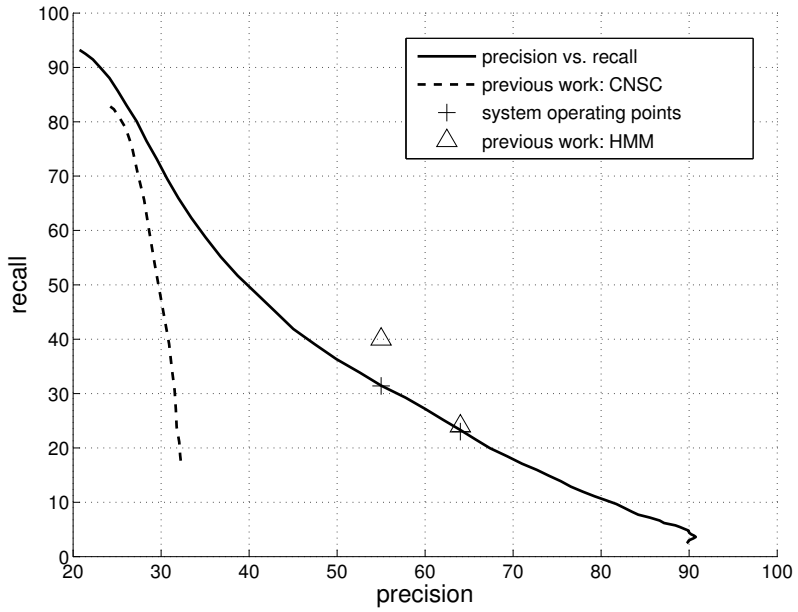


Figure 4.5: Overlap detection performance using CNSC features, in terms of precision and recall on the evaluation dataset [68]. For comparison, results of prior studies using CNSC [226] or HMM [18] are also shown.

The energy ratio threshold was tuned on the development set and fixed to $\delta_{ER} = 0.5$ across all audio recordings. The threshold for the total energy δ_{ET} was set dynamically for each audio recording and according to a fraction t_{ET} of the mean energy over the entire recording. Figure 4.5 shows the overlap detection performance in terms of precision and recall as a function of t_{ET} (solid profile), revealing considerably better performance than a previous system using CNSC (dashed profile) [226]. The new energy ratio and total energy features thus yield a notable improvement in system performance. Boakye *et al.* [18] report experiments using the same evaluation set, with precision/recall values of 55 %/40 % and 64 %/24 %. The system proposed in the present study achieved similar values of 55 %/31 % and 64 %/23 %. The two sets of results are also depicted in Figure 4.5 as triangles and crosses, respectively. The CNSC system achieved comparable performance without a duration model that is implicitly inherent in the HMM-based approaches.

In order to better understand the performance and weaknesses of the presented overlap detection system, the results were analysed as a function of overlap segment duration. For this study, the first operating point with precision/recall of 55/31 % was chosen. Figure 4.6 shows four histogram plots for the test set which illustrate overlap detection performance in terms of detected and missed overlap (top right and bottom left) as well as recall (bottom right). For comparison, a reference overlap

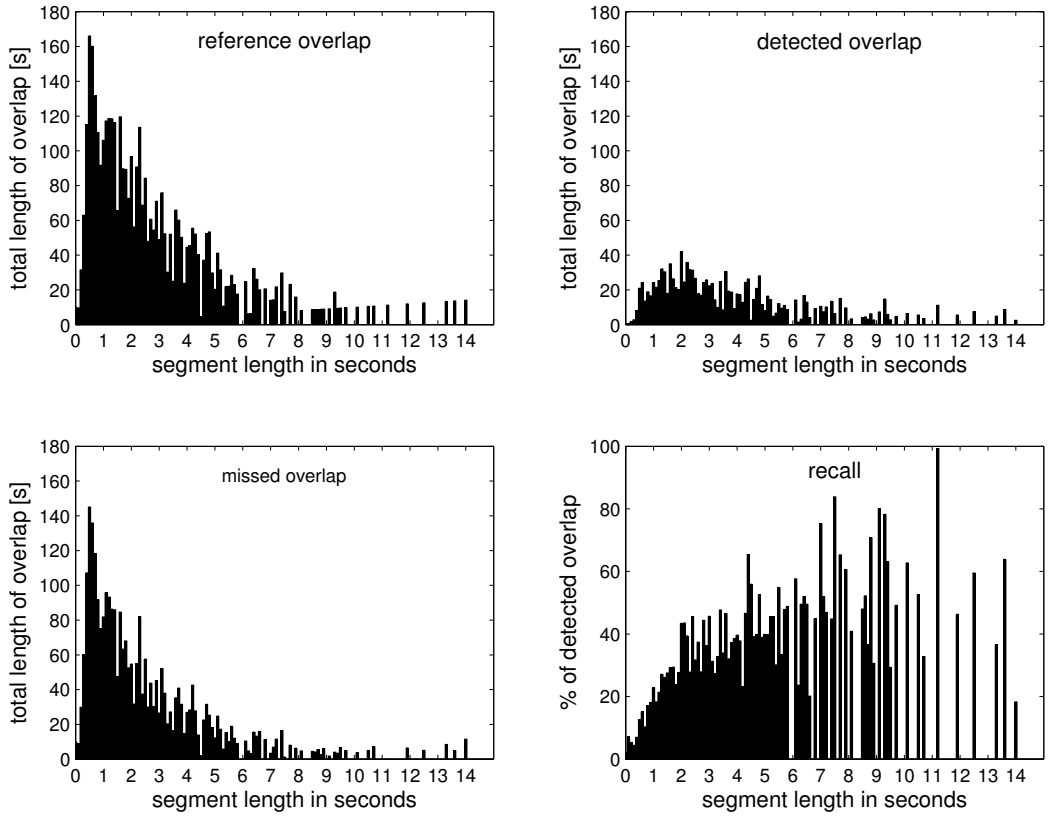


Figure 4.6: Weighted length histograms for reference overlap segments, detected overlap, missed overlap, and recall [68].

histogram is also presented (top left). Note that the distribution of overlap segment durations show the total contribution (in seconds) to the corpus for each bin (not the number of segments with the respective length, as would be the case in a conventional histogram). The plots show that the largest contribution to overlap comes from shorter segments with durations between 0.5 and 1.5 seconds. However, there is a surprising number of longer overlap segments with durations in excess of 4 seconds. The missed overlap and recall histograms show that short segments, which are the most frequent, were the least well detected. The results show that a significant penalty will be incurred if those short segments are not detected reliably. Future work should therefore focus on improving detection of shorter segments.

4.3.4.4 Conclusions

This section reported advances in applying CNSC to the detection of overlapping speech in the context of conference meetings and speaker diarization. CNSC is used

to separate a potentially overlapping speech signal into single-speaker signals. It was shown how the resulting CNSC base activations can be applied to detect overlapping speech segments.

The new CNSC approach yields overlap detection results which are comparable to a state-of-the-art HMM overlap detection system, when evaluated on the AMI meeting corpus. Compared to the HMM, a rather simple classifier is employed, which does not depend on any training data. Optimised parametrisations as well as new energy ratio and total energy features resulted in significantly better performance than previous work. This supports the potential for CNSC-based overlap detection. A new analysis of overlap detection performance highlights the need for continued work to improve overlap detection particularly for shorter segments of between 0.25 and 2 seconds in duration. A large part of these short overlap segments are backchannel utterances, in which one speaker speaks in the middle of a longer utterance of another speaker. However, very often it is not the case that there is a real acoustic overlap between these two speakers. Therefore, these segments can not be detected by overlap detection systems which rely only on acoustic features.

The study presented in the next section aims to integrate the CNSC-based overlap features into an HMM overlap detection framework to exploit the benefits of duration modelling. This approach is expected to improve overlap detection performance for overlapping segments of particularly short and long duration.

4.3.5 Audio Features for Overlap Detection

This section reports the combination of features derived through CNSC and new energy, spectral, and voicing-related features within a conventional HMM system. While in the previous section, new CNSC-based features for overlap detection were presented and tested with a threshold classifier, these are now combined with the duration modelling ability of an HMM system, using a tandem HMM setup.

The contributions of this section are two-fold: first, it reports the use of CNSC base activations within an HMM framework, which inherently includes duration modelling. Second, it introduces new energy, spectral, and voicing-related features which are well-suited for overlap detection.

4.3.5.1 HMM Overlap Detection System

As a classification framework, a standard HMM-based overlap detection system was applied, as first presented in [16]. Speech, nonspeech, and overlapping speech were each modelled by a three-state HMM. A short introduction to HMMs is given in 2.2.1.2. Observations were modelled by a multivariate GMM with diagonal covariance matrices. Due to unbalanced training data, the mixtures of the speech model had 256 components, while those of the nonspeech and overlap models had 64 components each. The models were trained with an iterative mixture splitting tech-

nique with successive re-estimation. The decoding grammar forbid self-transitions and transitions from nonspeech to overlapping speech. In order to trade off false-positive detections versus false-negatives, different operating points were tested. The log-likelihood transition penalty from speech to overlapping speech is the parameter that was tuned to obtain these operating points. This parameter is also referred to as overlap insertion penalty (OIP). A higher OIP leads to fewer false-positive detections which translates to higher precision, but also lower recall.

4.3.5.2 CNSC-based Features

This section investigates the application of the CNSC-based features (as introduced in Section 4.3.4) within the HMM framework. To recapitulate, CNSC is used to decompose the mixed signal into speaker bases and activations. From the activations, an energy term $E_t(s)$ is derived for each speaker s in all audio frames t, using Equation (4.17). Overlap detection features are then computed from the speaker energies. The first feature is the energy ratio ER_t, which is estimated according to Equation (4.18). In addition, the total energy ET_t is used, which is estimated by summing up the energies $E_t(s)$ of all speakers $s \in S$:

$$ET_t = \sum_{s \in S} E_t(s). \tag{4.19}$$

To mitigate variations in energy across different recordings, the average over all speech frames in the respective recording is subtracted from ET_t, resulting in the normalised total energy ET_{nt}:

$$ET_{nt} = ET_t - \frac{r}{|J_{sp}|} \sum_{t \in T_{sp}} ET_t, \tag{4.20}$$

where r is a regularisation factor tuned on held-out development data, and J_{sp} denotes the set of all speech frames in the recording. The latter is determined by the VAD component of the employed diarization system. Whereas the system presented in Section 4.3.4 investigated the thresholding of the normalised energy ratio ER_t and total energy ET_t to detect overlap, this section reports their use as features in an HMM-based overlap detection system.

4.3.5.3 Additional Features and Feature Selection

As already mentioned, in addition to the two CNSC-based features described above, the system proposed in this section also considers new energy, spectral, and voicing-related features which are well-suited to overlap detection. These audio features are a part of the set that was provided for the 2011 Audio/Visual Emotion Challenge [196] and can be extracted using the open source toolkit openSMILE [45]. The 35 candi-

date features, including CNSC and baseline MFCC features, are listed in Table 4.3. Originally, this feature set was designed for speech emotion recognition. It contains MFCCs, and additionally, several energy-related features are included, such as the loudness or the energy in different frequency bands. Other spectral features (e. g. roll-off points, flux, skewness, kurtosis, or harmonicity) provide a comprehensive description of spectral properties. To model the human voice, several voicing-related features (e. g. fundamental frequency f_0, probability of voicing, jitter, or shimmer) are included. Together, these features cover a broad range of properties of speech signals. It is therefore expected that among them, there are several audio features which are especially suited for overlap detection. In addition, this feature set includes many features which have previously been used for overlap detection. First of all, MFCCs can be considered a baseline in speech and audio processing. Furthermore, the application for overlap detection of features such as spectral kurtosis and harmonicity or zero-crossing rate was already investigated in [18]. All features are computed every 20 ms with window sizes between 25 ms and 60 ms, as indicated in Table 4.3.

A feature selection approach based on the Kullback-Leibler divergence, similar to that reported by Zhou *et al.* [258], was used to identify the features that are most relevant for overlap detection. The discriminant value d_f of each feature f was computed according to

$$d_f = D(p_f \mid\mid q_f), \tag{4.21}$$

where $D(\cdot \mid\mid \cdot)$ is the Kullback-Leibler divergence, p_f is the distribution of feature f for overlap frames, and q_f is the distribution over all frames. The Kullback-Leibler divergence of two probability distributions p and q is computed as

$$D(p \mid\mid q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x. \tag{4.22}$$

Under the assumption of Gaussian distributed features with mean μ and variance σ^2, Equation (4.22) can be computed as [167]:

$$D(p \mid\mid q) = \log \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}. \tag{4.23}$$

The resulting Kullback-Leibler divergence scores for all features are displayed in Table 4.3, which shows that a small selection is particularly well-suited to overlap detection. Scores for loudness, the two energy features, spectral flux, kurtosis, harmonicity, probability of voicing, jitter, shimmer, and the two CNSC features (illustrated in boldface in Table 4.3) are all higher than those for MFCC features. The energy-related features achieved the highest scores, which is expected to some extent, since the signal energy should be a good indicator of overlap. Jitter is a measure of fluctuations in fundamental frequency, while shimmer is a measure of

Table 4.3: Candidate features with window sizes [ms] and score of the Kullback-Leibler divergence based feature selection on the training set. Selected features are indicated in boldface.

Feature	**Win. size**	**Score**
Energy & spectral (27)		
MFCC 1 – 12	**60**	**0.01 – 0.06**
loudness (auditory model based)	**60**	**0.29**
zero crossing rate	25	0.04
energy in band 250 – 650 Hz	**25**	**0.98**
energy in band 1 kHz – 4 kHz	**25**	**1.15**
25 % spectral roll-off point	25	0.03
50 % spectral roll-off point	25	0.02
75 % spectral roll-off point	25	0.02
90 % spectral roll-off point	25	0.01
spectral flux	**25**	**0.43**
spectral entropy	25	0.02
spectral variance	25	0.00
spectral skewness	25	0.02
spectral kurtosis	**25**	**0.06**
psychoacoustic sharpness	25	0.00
spectral harmonicity	**25**	**0.09**
Voicing-related (6)		
f_0 (subharmonic summation followed by Viterbi smoothing)	60	0.03
probability of voicing	**60**	**0.18**
jitter	**60**	**0.08**
shimmer (local)	**60**	**0.11**
jitter (delta: 'jitter of jitter')	60	0.02
logarithmic harmonic-to-noise ratio	60	0.01
CNSC-based (2)		
energy ratio	**40**	**0.05**
CNSC energy	**40**	**0.28**

amplitude variability. It is thus of no surprise that they are also good indicators of overlap.

The set of selected energy, spectral, and voicing-related features is called ESV features in the following. All of these features were used together with MFCCs as additional inputs to an HMM overlap classifier. The feature set was augmented with first order regression coefficients. All features were normalised to have zero mean and unity variance, using the statistics of the training data only.

4.3.5.4 Experimental Evaluation

CNSC bases for each speaker were learnt from the standard diarization system output (which is regarded as pure speech). For this purpose, the top-down LIA-EURECOM speaker diarization system described in [20] was used. This means that the speaker bases were not always extracted from pure, non-overlapping speech of only a single speaker, but may also have contained overlapping speech and material from other speakers. However, it was found that this is no relevant degradation compared to using only clean speech in an oracle-style experiment, as was done for the experiments described in Section 4.3.4. The parametrisation of CNSC decomposition was the same as in Section 4.3.4. The factor r in Equation (4.20) was tuned on the development data and set to $r = 1.2$.

Overlap detection performance was assessed using averaged frame-level precision (Pre) and recall (Rec) statistics. In addition, the overlap detection error (Err) is reported, which is defined as the sum of false alarm and missed overlap times divided by the reference overlap time, as expressed in (4.4). This measure is a good indicator for the possible improvement in DER through overlap handling.

For the first experiment, only MFCCs were used as features. Then, the selected ESV features were added. Finally, this feature set was augmented by the CNSC overlap features. Each of these systems was evaluated for two different operating points, using different values of OIP. Table 4.4 shows overlap detection results for each of the different system setups. In addition, results from previous studies are shown for comparison. This includes another study in which the same test set was used for the experiments [18], as well as results from Section 4.3.4 in the present thesis. Results for an HMM-based system with MFCCs and other features, reported in [18], are illustrated in the first row for high recall (left) and high precision (right) setups. These results are slightly better than those for the CNSC-only system. However, it can be seen that it is difficult to obtain high precision results, even at the cost of lower recall. The last three rows of Table 4.4 show the results for the system described in this section. First, using only baseline MFCC features, then the same system with additional ESV features, and finally with additional CNSC features. In all cases, for the high recall setup, the parametrisation OIP = 100 was employed, whereas for the high precision setup OIP = 0. The results for MFCCs were slightly worse than the ones presented in [18]. This is not surprising since in

Table 4.4: Overlap detection results on the test set, comparing previously published results as well as results from Section 4.3.4 with the proposed HMM system with various features. For each of the proposed systems, two different precision (Pre) vs. recall (Rec) operating points with their respective overlap detection error (Err) are shown, depending on the OIP.

System			Pre	Rec	Err	Pre	Rec	Err
MFCC [18]			55.0	40.0	92.7	64.0	24.0	89.5
CNSC			55.0	31.0	94.4	64.0	23.0	89.9
MFCC	ESV	CNSC	OIP = 0			OIP = 100		
✓	-	-	45.2	50.1	110.8	54.3	25.8	96.0
✓	✓	-	60.9	24.0	91.4	85.7	13.2	89.0
✓	✓	✓	65.8	31.5	84.9	81.6	22.9	82.3

that study, there are also other features employed in addition to MFCCs. In the case of OIP = 0, it can be seen that the overlap detection error Err was above 100 %, which was caused by the low precision, since there are many false-positive detections. Adding ESV features led to a substantial improvement in precision and error over the MFCC baseline but a drop in recall. The inclusion of CNSC features brought further substantial improvements to recall performance which was then comparable to previous work [18] but with better precision and also the lowest error.

The results presented in this section show that the inclusion of more audio features in addition to MFCCs is capable of improving the overlap detection performance of an HMM system. Furthermore, features derived from CNSC decomposition lead to an additional improvement.

4.3.6 Overlap Detection using Lexical Information

Almost all of the prior studies in overlap detection focus on the use of acoustic cues. This includes audio features such as MFCCs, prosodic information, or, as introduced in the previous sections, features based on CNSC or other energy, spectral, or voicing-based features. While backchannel utterances such as 'yeah' or 'mm-hmm' are very frequent in spontaneous, overlapping speech [81], they do not necessarily overlap acoustically with competing speech. In Section 4.3.4, the system analysis showed that especially short overlap segments, such as those from backchannel utterances, are particularly difficult to detect using acoustic features on their own. New approaches, exploiting different cues are thus required.

This section introduces the application of lexical information, in which an analysis of the spoken content is used to improve overlap detection. Using language models for single-speaker speech and overlap, an overlap score is created for every

spoken word and used as an additional feature within the HMM framework. The motivation to use lexical features for overlap detection stems from the hypothesis that some words are more likely to occur during overlap than others, and that thus, spoken words can be used to detect overlap. This is intuitively the case for floor grabbers, backchannels, and interruptions, for example.

The lexical or linguistic content of a speech signal has previously been used for example for speech emotion recognition [194] or for speaker diarization [21]. This section considers the use of such higher-level information for overlap detection. The spoken content of the audio signal is one such source of information. Central to the idea is the use of language models to detect backchannel words and other language characteristics which typify instances of overlap. In particular, this section presents a new approach to overlap detection using lexical features, which uses language models to characterise the spoken content in overlapping and single-speaker speech. The language models are used to assign scores to each word in a dictionary and thus to estimate the probability that the lexical content reflects overlapping speech. Using the output of an automatic speech recognition (ASR) system, these scores are estimated for an unknown audio segment. The resulting scores are used within a conventional HMM framework to detect overlap. Experiments conducted on the AMI corpus illustrate that the proposed lexical features lead to improved performance.

4.3.6.1 System Overview

An overview of the proposed system is illustrated in Figure 4.7. It shows the integration of the new lexical features into an HMM-based overlap detection system with baseline features. Unigram language models are learnt for single-speaker and overlapping speech using independent training data and ground-truth, word-level transcriptions. Test data are processed with an ASR system to produce a comparable word-level transcription. This transcription is used together with the two language models to estimate a score which reflects the relative likelihood that the signal contains speech from a single, or more than one speaker. The score is combined with the baseline features and used in an HMM detection system (tandem system setup) which classifies the signal as either nonspeech, speech (from a single speaker), or overlap. The HMM system is the same as described in Section 4.3.5.1.

4.3.6.2 Lexical Cues for Overlap Detection

In this study, language models are used to characterise the distribution of words during non-overlapping and overlapping speech. Generally, a language model is used to assign a probability to a sequence of words $p(w_1, ..., w_m)$. Of practical use are N-gram language models, in which the probability for a word depends on the last $N-1$ words. In speech recognition, it is common to use bigram or trigram language

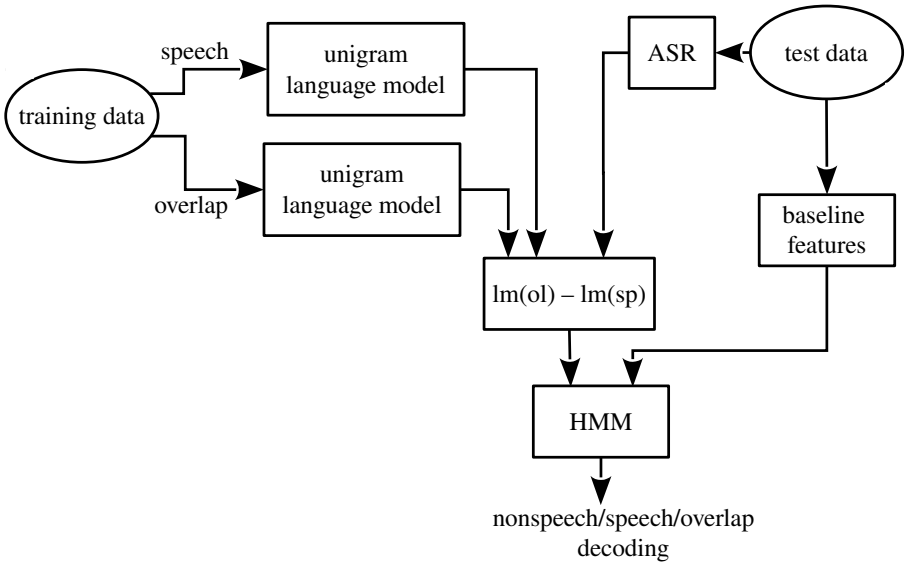


Figure 4.7: System overview for the overlap detection system using lexical information [63].

models. In contrast, a unigram language model describes only the probability of a single word $p(w)$. Such unigram language models $p(w|c_w = sp)$ and $p(w|c_w = ol)$ for speech and overlap (as the word class c_w) are computed using training data for single-speaker and overlapping speech, respectively. In practice, however, to permit the automatic recognition of only a single word at a time, only the longest spoken word per overlap interval is taken into account.

The detection of overlap using lexical content is therefore equivalent to determining the probability of overlap $p(c_w = ol|w)$ for any given word. With Bayes' theorem, this can be expressed as:

$$p(c_w = ol|w) = \frac{p(w|c_w = ol) \cdot p(c_w = ol)}{p(w)}. \tag{4.24}$$

Since the prior probability $p(c_w = ol)$ is independent of the word and $p(w)$ is approximated by the language model probability for single-speaker speech $p(w|c_w = sp)$, Equation (4.24) can be reduced to

$$p(c_w = ol|w) \sim \frac{p(w|c_w = ol)}{p(w|c_w = sp)}. \tag{4.25}$$

Using log-likelihoods, the probability of overlap $s_{lex}(w)$ is finally expressed as:

$$s_{lex}(w) = \log\big(p(c_w = ol|w)\big) \sim \log\big(p(w|c_w = ol)\big) - \log\big(p(w|c_w = sp)\big). \tag{4.26}$$

Equation (4.26) reflects the relative likelihood of speech of a single speaker compared to multiple speakers. The value $s_{lex}(w)$ can be computed for every word in the

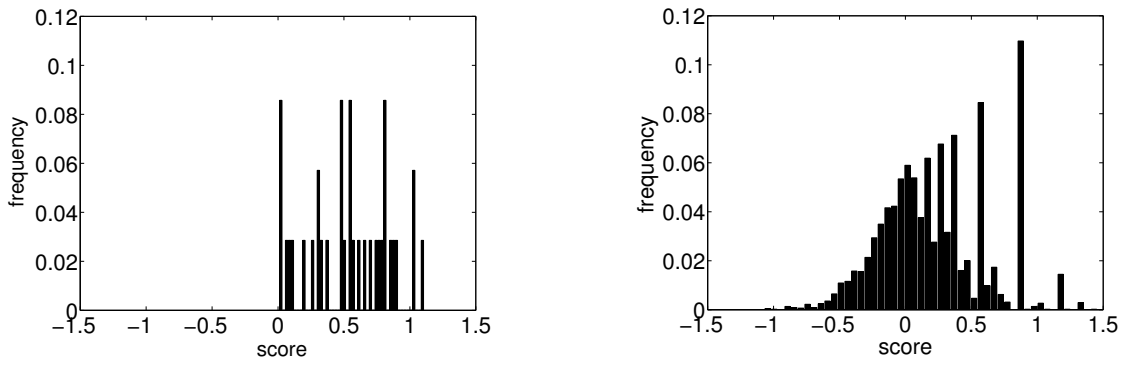


Figure 4.8: Distribution of overlap scores for backchannel words (a) and other words (b).

training set. A ground-truth annotation is subsequently used to construct a word-level look-up table and to identify those words which are most or least indicative of overlap. Words which appear more often in overlapping speech and less often in non-overlapping speech are assigned higher scores, and vice versa. The score forms the new feature to be used for overlap detection.

Words typically used in overlap segments (e. g. in backchannel utterances) like 'mh-hmm', 'uh-huh', 'ah', 'yeah', 'yep', 'okay', 'right' were all shown to be among those with the highest scores. Figure 4.8 shows the distribution of overlap scores computed with Equation (4.26) for a choice of 35 typical backchannel words (left) and for all other words (right) in the training set. The distribution of the non-backchannel words is centred around $s_{lex} = 0$, except for some outliers, while all of the backchannel words obtained positive scores. Most of the outliers for the non-backchannel words are represented by words that occur very rarely, which leads to bad estimates of the overlap scores. These observations support the hypothesis that lexical cues have potential for overlap detection.

For the training data, the reference transcriptions were used to determine $s_{lex}(w)$ on a frame-by-frame basis according to Equation (4.26). The value of $s_{lex}(w)$ was then added to the baseline feature set. As with all other features, first order regression coefficients were added and the features were normalised. The corresponding value of $s_{lex}(w)$ was assigned to recognised words in test data from the look-up table. For both training and test data, a value of $s_{lex}(w) = 0$ was assigned in the absence of any recognised words.

To assess the potential of lexical cues for overlap detection independently from the performance of an ASR system, the performance of the proposed system was evaluated using a so-called *oracle-style* ASR system, that is, ground-truth transcripts. Accordingly, to simulate the output of a more realistic ASR system, the transcripts were purged of overlapping words so as to retain only those with the

Table 4.5: Precision (Pre), recall (Rec), and overlap detection error (Err), each given in %, on the test set for the four feature combinations. Operating points were tuned (by varying OIP) to achieve minimum overlap detection error on the tuning set.

Features	**OIP**	**Pre**	**Rec**	**Err**
MFCC	50	55.3	34.2	93.4
MFCC + ESVC	50	77.8	25.8	81.6
MFCC + lexical	65	72.6	23.2	85.5
MFCC + ESVC + lexical	85	81.7	28.0	78.3

largest duration. The output of such an oracle-style ASR system thus corresponds to that of a perfect ASR system, but capable only of single-word recognition. This is the same strategy as applied for language model training.

4.3.6.3 Experimental Evaluation

The proposed system using lexical information for overlap detection was evaluated using the same partition of the AMI database as described in 4.3.3.1. The required language models were estimated using a larger training set of 161 meeting recordings. This step helped to provide a better estimate of the language models. The following system parameters were applied for feature extraction: energy, spectral, and voicing-related features were computed every 20 ms. A window size of 60 ms was applied for MFCC and voicing-related features, whereas other energy and spectral features were determined using a window size of 25 ms. In addition, CNSC-based features as developed in previous sections of this thesis were used. CNSC was applied using magnitude spectra computed from 40 ms windows with a window shift of 20 ms. For the CNSC features, the number of bases per speaker was set to $R = 35$. Furthermore, the algorithm used a convolutional range $P = 4$ and a sparseness parameter $\lambda = 0.05$. The regularisation factor in Equation (4.20) was set to $r = 1.2$. Speaker bases were learnt using speaker-specific training data obtained with the LIA-EURECOM speaker diarization system [20]. The system performance was measured in terms of frame-wise precision, recall, and overlap detection error.

4.3.6.4 Results and Conclusions

Results are reported for four different combinations of MFCC features, ESV, and CNSC-based features (denoted as ESVC in combination) as well as for the newly proposed lexical features (using an oracle-style ASR system). Results are illustrated in Figure 4.9 for each tested feature combination as a function of OIP. In addition, Table 4.5 lists the test set results for all four systems at one operating point. This operating point was determined by varying OIP and evaluating on the tuning set to achieve a minimum overlap detection error.

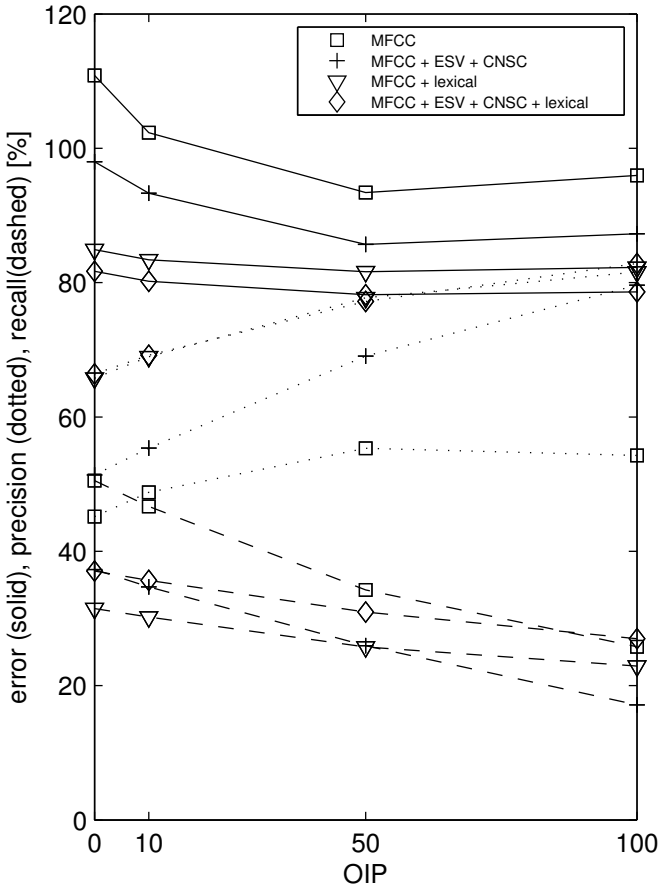


Figure 4.9: Overlap detection performance as a function of the OIP for different feature combinations [63]. Performance illustrated in terms of detection error (solid line), precision (dotted line), and recall (dashed line).

The lowest detection error achieved with MFCC features alone was 93.4 % at an OIP of 50. In combination with the new lexical feature, the error dropped to 85.5 % as a result of improvements in precision. For other system operating points (other values for OIP), the addition of lexical information to the MFCC feature set also helped to decrease the overlap detection error. The best feature set without lexical features combines MFCCs, energy, spectral, and voicing-related features with CNSC features. With this feature set, the minimum detection error was 81.6 %, again with an OIP of 50. When lexical features were added to this feature set, the error dropped to 78.3 %. This time, however, the drop in error can be attributed to an increase in both, precision and recall. Here again, the addition of the new lexical feature led to a consistent performance gain for all tested values of OIP. The increased recall in the case of using lexical features can be attributed to a better detection of small overlap segments, for example those containing backchannel utterances.

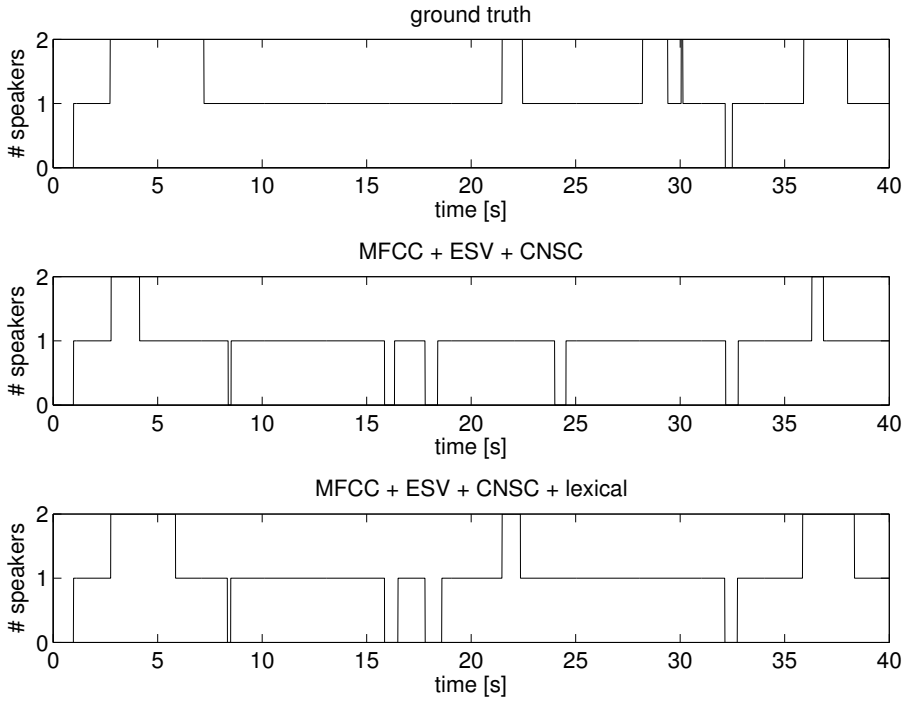


Figure 4.10: Illustration of overlap detection, showing annotations for the ground truth, the baseline system, and the baseline combined with the lexical feature (from top to bottom) for a 40-second excerpt of a recording from the test set [63].

Figure 4.10 illustrates overlap detection performance for a 40 second excerpt of a recording from the test set. The three plots show the ground-truth annotation (0, 1 or 2 active speakers), the output of the baseline system and the output when the new lexical feature is added. Together, these plots illustrate how the new feature has the potential not only to improve overlap detection accuracy of those segments already detected with the baseline approach, but also smaller segments which are not at all detected by the baseline system.

The experimental results confirm the hypothesis that lexical features can help to improve overlap detection; the results show improved precision and recall performance and reduced overlap detection error. Lexical features were derived from an oracle-style ASR system. It has to be tested how these results transfer to a real-case scenario with a more error-prone ASR system.

To extend the idea further, future work should not only consider the use of language model scores, but also those of acoustic models. In addition, continuing the idea of using more information beyond pure acoustic features, further work should study the nature of overlapping speech in greater detail so that new insights can stimulate the development of future overlap detection systems.

4.3.7 Overlap Detection with Memory-Enhanced Recurrent Neural Networks

In this section, neural networks exploiting long-range temporal context are applied to the problem of overlapping speech detection. Together with the HMM overlap detection system, neural networks are used in a tandem architecture.

In Section 4.3.4, an analysis of detected overlap segments showed that especially short segments of overlapping speech are hard to detect. Such segments often include backchannel utterances or interruptions, which are characterised by a low degree of actual acoustic overlap. Therefore, systems that go beyond pure acoustic features should be considered in order to improve overlap detection. For example, an approach presented in [251] uses the output of a voice activity detection system and the silence distribution to detect overlap. This study was extended by exploiting long-term conversational features for overlap detection [250]. In the previous section, it was shown how lexical information can be used to detect overlapping speech.

Neural networks could potentially improve overlap detection by putting a larger emphasis on analysing temporal context. This idea is motivated by the findings of [83], which show that interruptions (which are a common type of overlapping speech) are likely to occur after speech with certain acoustic properties. Thus, incorporating more temporal context should help to improve an overlap detection system. In the domain of speech recognition, tandem architectures which combine neural networks with HMMs were applied successfully [28]. Neural networks can be equipped with capabilities for context modelling with the introduction of recurrent neural networks (RNNs). However, the amount of context a conventional RNN can exploit is limited. Long short-term memory (LSTM) RNNs have been proposed to overcome this so-called vanishing gradient problem [100]. Recently, LSTM RNNs have been shown to deliver good results for VAD [44], which is a related problem to overlap detection.

In this section, LSTM RNNs (the acronym LSTM will be used to denote an LSTM RNN in the following) are applied to the task of speech overlap detection. Using conventional MFCC features as well as energy, spectral, vocing-related, and CNSC-based features that were introduced for overlap detection in the previous sections, LSTMs are used for regression to predict frame-wise overlap scores. These scores are employed to detect segments of overlapping speech. In addition, the predicted overlap scores are applied as features within the HMM framework. Experiments were conducted with the AMI corpus of meeting recordings containing spontaneous speech.

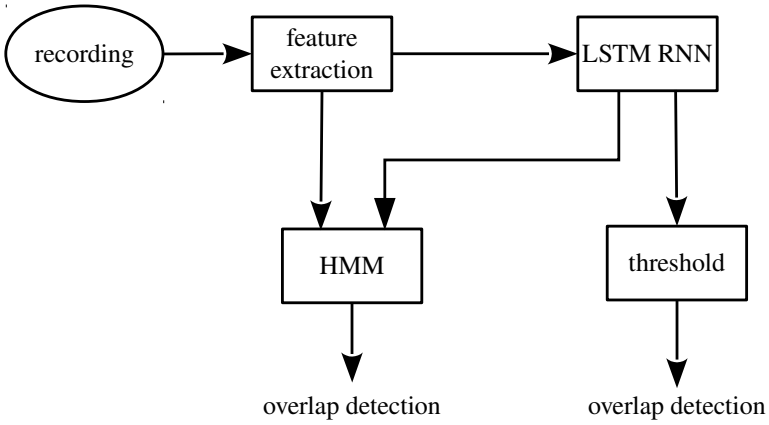


Figure 4.11: System overview for the LSTM overlap detection system [64].

4.3.7.1 System Overview

The proposed overlap detection system is depicted in Figure 4.11. It consists of a conventional HMM system for overlap detection. In addition, extracted audio features can also be fed to the LSTM to generate overlap predictions. These overlap predictions are either used directly (by applying a threshold) to detect overlap or they are added to the other features and decoded with the HMM, resulting in a tandem LSTM-HMM system.

4.3.7.2 Long Short-Term Memory Recurrent Neural Networks for Overlap Detection

RNNs are a widely used technique for context-sensitive sequence labelling. They exploit context in the form of inputs from past time steps by using cyclic connections. Due to the so-called vanishing gradient problem (the influence of a certain input on the hidden and output layers of the network decays exponentially over time), the context that is used by an RNN is limited. In order to overcome this problem, LSTM networks have been introduced in [100]. LSTMs use memory cells to store information over a longer period of time. An LSTM hidden layer is composed of so-called memory blocks. Each memory block consists of a memory cell and three multiplicative gate units (input, output, and forget gate). These gates allow for write, read, and reset operations of the memory cell. The amount of context information that the network uses is learnt during training. Due to their ability to model long-range dependencies between the inputs, LSTMs seem to be a promising approach for overlap detection. A more detailed description of LSTM networks is given in Section 5.2 of this thesis.

In this study, LSTMs are applied for linear regression to predict framewise overlap scores. For this purpose, the output layer of the network consists of a single

linear unit with output $o(t)$ at time t. The input feature vectors of the network are defined as

$$X_f = [x_{f1}, ..., x_{fT}] \tag{4.27}$$

where T is the number of frames in the target sequence. The output $o(t)$ depends on the past input vectors $X_{ft} = [x_{f1}, \ldots, x_{ft}]$,

$$o(t) = \mathcal{F}(X_{ft}) \tag{4.28}$$

due to the LSTM principle and the recurrent nature of the network. During network training, output targets $\hat{o}(t)$ are defined as

$$\hat{o}(t) = \begin{cases} 1 & \text{if } x_{ft} \in \text{overlap} \\ 0 & \text{if } x_{ft} \in \text{speech} \\ -1 & \text{if } x_{ft} \in \text{nonspeech} \end{cases} \tag{4.29}$$

which results in a larger distance between nonspeech and overlap. This punishes the confusion of nonspeech and overlap more than that of speech and overlap. The idea underlying this principle is that nonspeech and overlap should also have a larger distance in the feature space (e. g. when using energy features) than speech and overlap. The predictions $o(t)$ of the trained network are used for classification by applying a threshold θ,

$$c_w(t) = \begin{cases} 1 & \text{if } o(t) \geq \theta \\ -1 & \text{if } o(t) < \theta \end{cases} \tag{4.30}$$

whereby the predicted class $c_w(t)$ differentiates only between overlap ($c_w = 1$) and non-overlap ($c_w = -1$). The threshold θ is varied to obtain different system operating points as a trade-off between precision and recall.

The size of the input layer of the network was equivalent to the number of employed audio features. One recurrent hidden layer with 50 LSTM blocks was used. This topology proved to be efficient for voice activity detection [44] and therefore, it was also considered for this study. Networks were trained and evaluated with the RNNLIB software by Graves *et al.* [85]. LSTM training was performed with the backpropagation through time algorithm; the weights were updated using gradient descent with a learning rate of 10^{-5} and a momentum of 0.9. Moreover, the weights were required to be initialised with non-zero values, leading to the choice of uniform random values sampled from $]0; 0.1]$. To enhance generalisation, Gaussian noise with zero mean and standard deviation of 0.3 was added to all inputs. A maximum of 40 training epochs was run to avoid over-adaptation. Training was stopped if there was no error improvement on the development set for the ten most recent epochs. The frame-wise mean quadratic error between the targets $\hat{o}(t)$ and the network predictions $o(t)$ served as an error measure during network training.

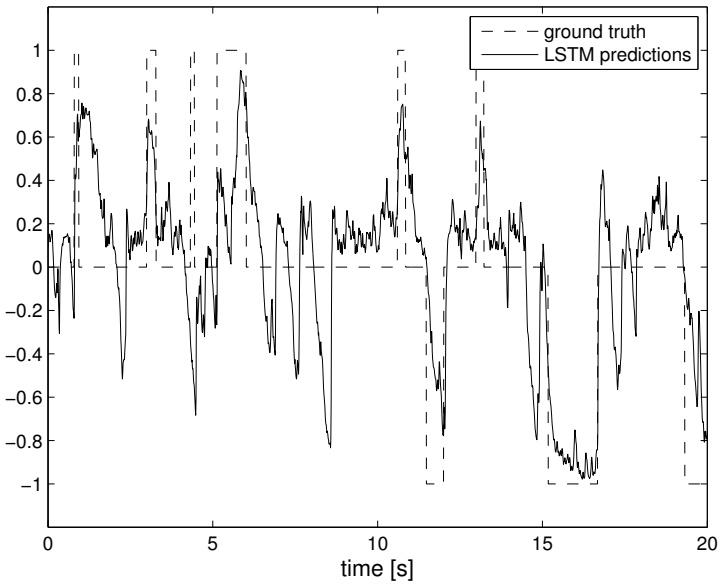


Figure 4.12: LSTM predictions for a 20-second excerpt from the test set [64]. The dashed line marks the ground truth (-1: nonspeech, 0: speech, 1: overlap).

Figure 4.12 shows LSTM predictions $o(t)$ for a 20-second excerpt from the test set. It can be seen that the LSTM predictions are well correlated with the ground truth, yielding low values for nonspeech regions and high values for overlap segments. By applying a threshold, overlap segments can be detected from these LSTM predictions.

To combine LSTM with HMM overlap detection, the LSTM predictions $o(t)$ were used as an additional feature for the HMM system. As for all other features, delta coefficients were computed for LSTM predictions. However, LSTM predictions were not normalised, based on the findings of preliminary experiments.

4.3.7.3 Experimental Setup

Experiments were conducted using a subset of the AMI corpus, using the partition into training, tuning/development, and test set as specified in Section 4.3.3.1. Feature extraction for MFCC and ESVC features was performed using the methodology described in Section 4.3.5.

Both the HMM and LSTM-based overlap detection systems were used to perform experiments. The systems were tested both with 24 MFCCs and with the 46 ESVC features. In addition, the combination of HMM and LSTM was evaluated, where LSTM predictions are added to the HMM feature set. This tandem system was also tested with MFCC features and with ESVC features. Thus, in total, six different system configurations were tested.

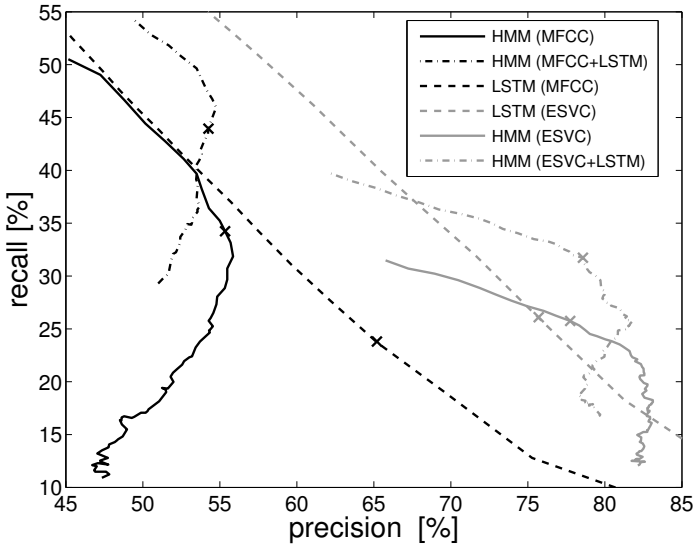


Figure 4.13: Precision and recall for HMM (solid lines), LSTM (dashed lines), and their combination (dashed-dotted lines), each time using either MFCC (black) or ESVC (grey) audio features [64]. An 'X' marks the operating point with minimal overlap detection error as determined with the development set.

4.3.7.4 Results and Conclusions

Figure 4.13 plots recall against precision for all six tested systems. In addition, each system's operating point is marked as 'X' and displayed in Table 4.6. These operating points are the points with minimum overlap detection error on the development set. For the LSTM system, different operating points were obtained by varying the threshold for overlap detection, leading to rather straight curve in the plot. Different operating points for the HMM systems were obtained by increasing OIP. At first, increasing OIP resulted in higher precision and lower recall, while for higher OIP, both measures decreased. This can be seen from the results for all HMM systems.

With MFCC features, LSTM-based overlap detection performance was comparable to the HMM in the high-recall region. Beyond this, LSTMs with MFCC features were capable of achieving high-precision operating points. The two different classifiers achieve an error of 93.4 % and 88.9 %. This difference was the result of higher precision with the LSTM system. Using ESVC features instead of MFCCs in the HMM system resulted in higher precision and lower recall. Yet again, the performance of the LSTM was comparable to that of the HMM, with error rates of 81.6 % and 82.3 %, respectively. For both feature sets, adding LSTM predictions (resulting in the tandem system) greatly improved the recall performance while achieving similar precision values. In the case of using MFCC features, for the minimum-error operating point, recall increased from 34.2 % to 44.0 % while precision stayed roughly

Table 4.6: Precision (Pre), recall (Rec), and overlap detection error (Err) on the test set for the six tested system and feature combinations. Operating points were tuned to achieve minimum overlap detection error on the development set.

Features	**System**	**Pre**	**Rec**	**Err**
MFCC	HMM	55.3	34.2	93.4
ESVC	HMM	77.8	25.8	81.6
MFCC	LSTM	65.2	23.8	88.9
ESVC	LSTM	75.7	26.1	82.3
MFCC + LSTM predictions	HMM	54.3	44.0	93.1
ESVC + LSTM predictions	HMM	78.6	31.7	76.9

at the same level. Due to the relatively low precision, the overlap detection error did not improve. With ESVC features, the system combination increased recall from 25.8 % to 31.7 %, while the precision remained constant. The error decreased from 81.6 % to 76.9 %, which is the consequence of the increased recall performance.

It is worth noting that in the case of ESVC features, the combination of LSTM and HMM was Pareto-dominant to the single systems. This means that the system combination was in (almost) all cases (operating points) better than the single systems. This effect was not observed for MFCC features.

In summary, the experimental results show that LSTMs alone perform comparably to HMMs in terms of overlap detection error. The combination of HMM and LSTM can substantially improve the overlap detection performance due to higher recall. This combination is obtained by adding LSTM predictions to the HMM feature set. Overlap detection recall is improved (23 % relative improvement in the case of using ESVC features) while keeping precision constant. Thereby, the overlap detection error is substantially reduced. One reason for higher recall with the tandem system could be that the ability of LSTMs to exploit long-range context helps to identify overlap segments which are hard to detect by the HMM with acoustic features alone. Examples of such segments are short backchannel utterances.

4.3.8 Summary of Overlap Detection Results

In Section 4.3, different methods and approaches for speech overlap detection were presented. Because all systems were evaluated with the same database of meeting speech recordings, the results can directly be compared. The results (in terms of precision, recall, and overlap detection error) of selected experiments described in the previous sections are summarised in Table 4.7. For simplicity, for all systems, results are only reported for one operating point. This operating point was tuned to achieve a minimum overlap detection error, which is the most important performance measure for the application of overlap handling for speaker diarization. In addition,

Table 4.7: Comparison of overlap detection results for different systems evaluated in this study. Results are given in terms of precision (Pre), recall (Rec), and overlap detection error (Err) on the employed test set of the AMI corpus. Operating points were selected to achieve minimum overlap detection error.

System	**Pre**	**Rec**	**Err**
HMM (MFCC + other) [18]	64.0	24.0	89.5
CNSC	64.0	23.0	89.9
HMM (MFCC)	55.3	34.2	93.4
HMM (ESVC)	77.8	25.8	81.6
HMM (ESVC + lexical)	81.7	28.0	78.3
HMM (ESVC + LSTM)	78.6	31.7	76.9

results from Boakye *et al.* [18] are included, as they use the same test set of meeting recordings. In their study, only precision and recall are reported as performance measures. The overlap detection error can, however, directly be computed from the precision and recall values, as was shown in Equation (4.4). This system can be considered a baseline system as it is an HMM overlap detector employing MFCCs as well as other spectral audio features, which are rather standard techniques.

In Section 4.3.4, an approach for overlap detection using a signal separation technique, namely CNSC, was presented. It is based on the idea of separating possibly overlapping speech signals into the contributing speakers using CNSC. From the resulting speaker activation energies, features were constructed for overlap detection. This method works without the HMM framework, but achieved comparable results (cf. row two in Table 4.7). The study presented in Section 4.3.5 consisted of selecting relevant audio features for overlap detection. These features were selected from a large set of candidate features, including MFCCs, energy, spectral, and voicing-related (ESV) features. Furthermore, the CNSC-based features were added to the feature set. For comparison, a baseline system that uses only MFCCs was tested (row three). Substantial improvements were obtained with the larger feature set and the overlap detection error was further decreased to 81.6 %. An approach exploiting lexical information for overlap detection was introduced in Section 4.3.6. The spoken content was analysed for words that are cues for overlap. Language models were used to create a score that was added to the feature set of the HMM system. With this new feature, improvements in all performance measures were obtained (row five). Finally, in Section 4.3.7, the application of LSTM neural networks for overlap detection was described. This classifier can exploit long-range temporal context, which proved to be successful in improving overlap detection results. The LSTM system was combined with the HMM in a tandem setup, which resulted in the best overlap detection error among all systems presented in this thesis (row six).

Table 4.8: Overlap detection performance depending on the overlap segment length, in terms of statistics derived from the number of positive overlap (P) samples as well as true-positive (TP) and false-negative (FN) detections.

	Length l [s]					
	0 – 0.33	0.33 – 1	1 – 1.5	1.5 – 5	>5	all
P [s]	56	816	568	2116	727	4283
$\frac{\mathrm{P}}{\Sigma\mathrm{P}}$ [%]	1.3	19.0	13.3	49.4	17.0	100
TP [s]	1	57	110	795	396	1359
Rec = $\frac{\mathrm{TP}}{\mathrm{P}}$ [%]	2.3	7.0	19.3	37.6	54.5	31.7
FN [s]	55	759	458	1321	331	2924
$\frac{\mathrm{FN}}{\mathrm{P}}$ [%]	97.7	93	80.7	62.5	45.5	68.3
$\frac{\mathrm{FN}}{\Sigma\mathrm{P}}$ [%]	1.3	17.7	10.7	30.9	7.7	68.3

Similar to the analysis presented in Section 4.3.4.3, the performance of the best overlap detection system in terms of detection error (HMM with ESVC + LSTM features, cf. last row in Table 4.7) was analysed for different overlap segment lengths. The results of this analysis are presented in Table 4.8, in which all results are given for different intervals of overlap segment length. The first row shows the sum of the lengths (in seconds) of all overlap segments in these intervals, with the percentages given in the second row. It can be seen that segments with a length between 1.5 and 5 s made up around half the amount of total overlap. The next pair of rows show the amount of correctly detected overlap in seconds and in percentage (which is the recall value). In the same manner, the table lists the missed detections in rows five to seven. These numbers clearly show that short overlap segments ($l < 1.5$ s) are especially difficult to detect, resulting in high missed detection (FN) rates. On the other hand, due to the large amount of overlap segments with a length between 1.5 and 5 s, the overall contribution to the missed detections is comparably small. It can be concluded that the potential for improvement is larger for short overlap segments. Overall, however, overlap detection can profit the most from improved detection of longer overlap segments.

4.4 Overlap Handling

This section describes the full integration of overlap detection results into a top-down diarization system through the application of overlap exclusion and overlap labelling. Experiments on a subset of the AMI corpus show that the system delivers significant reductions in missed speech and speaker error.

4.4.1 Methodology

Overlap handling is achieved with two setups which correspond to different system operating points for overlap detection with a different trade-off in precision versus recall. In the case of an HMM system, these operating points are obtained by varying OIP. First, detected intervals of overlapping speech are excluded from the diarization clustering process to reduce speaker model impurities. For this approach, a high overlap detection recall is desired in order to detect and discard as much overlapping speech as possible. Second, overlap labelling is applied by adding a second speaker in the diarization output for all intervals of detected overlapping speech. While this can reduce missed speaker time, it can also introduce false alarms, and thus a high precision operating point is desirable in the overlap detection system. In the present study, two approaches are used to determine the second speaker: either the GMM log-likelihoods (LLKs) from the diarization system or the CNSC energies according to Equation (4.17) are summed up over the detected overlap segment. The speaker with the highest summed score (or the second highest, if the speaker with the highest score is already that one detected by the baseline system) is then added as a second speaker.

4.4.2 Results and Conclusions

In the experiments, the overlap detection system presented in Section 4.3.5 was utilised. This is an HMM overlap detection system using the ESVC audio feature set including MFCCs, energy, spectral, voicing-related, and CNSC-based audio features. For overlap exclusion, the OIP was set to zero (resulting in a higher recall, with precision of 66 % and recall of 31 %), while for overlap labelling the OIP was set to 100 (resulting in a higher precision, with precision of 82 % and recall of 23 %). The speaker diarization system used for the experiments was the top-down LIA-EURECOM system reported in [20]. Finally, so that all results are independent of speech activity detection, reference speech/nonspeech segmentations were used in all cases.

Results are presented in Table 4.9. For the baseline system, overlapping speech was shown to contribute 15 % to the missed speaker error[4] whereas there were no false alarms due to the use of reference speech/nonspeech transcriptions. With a speaker error of 18.2 %, a baseline DER of 33.2 % was obtained. The DER fell marginally to 32.7 % (1.5 % relative improvement) when overlap exclusion was used

[4]Note that the missed speaker error is computed as a percentage of the total speaker time, which weights all speech segments by the number of contributing speakers and thus counts overlap segments twice (in the case of two overlapping speakers). Therefore the number of 15 % for the missed speaker time is lower than the number of 20 % reported in Section 4.3.3.1 as the percentage of overlapping speech (which is computed as a fraction of all speech frames without counting overlap twice).

Table 4.9: Influence of overlap handling (applying either overlap exclusion, overlap labelling or both), showing the missed speaker error (Miss), false alarm error (FA), speaker error (SpkE), diarization error rate (DER), and relative improvement in DER over the baseline. Overlap labelling was performed using either LLK scores or CNSC energy scores.

System	Miss	FA	SpkE	DER	Improvement
baseline [20]	15.0	0.0	18.2	33.2	
+ exclusion	15.0	0.0	17.7	32.7	+ 1.5 %
+ labelling LLK	11.6	0.6	20.1	32.3	+ 2.7 %
+ labelling CNSC	11.6	0.6	19.6	31.9	+ 4.0 %
+ exc. + lab. LLK	11.6	0.6	19.4	31.6	+ 4.8 %
+ exc. + lab. CNSC	11.6	0.6	18.9	31.1	+ 6.4 %

to reduce clustering impurities. This improvement stemmed from a reduced speaker error due to improved speaker modelling. On its own (without exclusion), overlap labelling had a slightly larger impact on performance. The DER improved by 2.7 % relative when labelling was performed using LLK scores and by 4.0 % relative for CNSC scores. With a small increase in false alarms, the average missed speech rate fell to 11.6 %, whereas there was a small increase in speaker error due to erroneous labelling. Here, CNSC speaker labelling was more potent in determining the overlapping speaker, as compared to labelling using the GMM LLK scores. When used in addition to overlap exclusion, the LLK and CNSC-based overlap labelling approaches gave relative improvements of 4.8 % and 6.4 %, respectively.

The diarization results show the expected behaviour of decreased missed speaker error and thereby reduced DER. Overlap labelling requires a high-precision operating point of the overlap detection system. Otherwise, false positive overlap detections would result in an increased false alarm error contributing to the DER. However, a high-precision operating point comes at the cost of lower recall. Thereby, only a part of the missed speaker error is eliminated, as can be seen in the experimental results. With improved overlap detection performance in terms of recall, even larger DER improvements could be obtained in the future.

4.5 Online Speaker Diarization

While the previous sections focussed on improving a speaker diarization system by detecting and handling overlapping speech, this section introduces an algorithm for online speaker diarization. Conventional speaker diarization systems mostly work offline, which means that the whole audio stream is processed at the same time. All audio segments must be present before the speakers are compared and clustered. In online speaker diarization, the segments are processed as soon as they are recorded, meaning that as soon as a new segment of the audio stream arrives, it must be

assigned to a speaker. Therefore, offline systems have access to more information, which makes the results more stable. In addition, online systems have to fulfil real-time constraints (except a small latency not longer than a couple of seconds), which is not the case for offline systems.

In this section, an online speaker diarization system is presented. The system is based on GMMs, which are used as speaker models. The system starts with three such models (one each for both genders and one for nonspeech) and creates new models for individual speakers just when the speakers occur. As more and more speakers appear, more models are created. The system implicitly performs audio segmentation, speech/nonspeech classification, gender recognition, and speaker identification. Experiments are performed with the HUB4–1996 radio broadcast news database [80].

Most of the offline speaker diarization systems work in the following way: a complete audio stream is segmented in smaller homogenous parts, each of them containing only one speaker. After the complete audio stream has been segmented, the segments are compared and clustered. In this way, one cluster is created for every speaker. This is done for example using the Bayesian information criterion [222]. Online speaker diarization is different, the system cannot wait for all segments to arrive before the clustering process begins. Therefore, no hierarchical clustering algorithms (as in offline speaker diarization) can be applied. More sophisticated clustering algorithms have to be used. In [141], a leader-follower clustering for k-means clustering and a dispersion-based speaker clustering are proposed for online speaker diarization.

GMMs in addition with universal background models and maximum a posteriori (MAP) adaptation (also known as Bayesian adaptation [41]) were proposed for speaker verification in [182]. From a large training database, a universal background model is created, and to enrol a new speaker in the verification system, very small amounts of data are needed, because the speaker's GMM is created from the universal background model with MAP adaptation. The present study proposes to use these techniques for a real-time speaker diarization system.

A system which is similar to the proposed system was presented in [145]. However, there are some substantial differences. In [145], a different model adaptation technique is used and the main difference is that a simpler database is used for the experiments. Whereas in [145], a database of recordings of European Parliament plenary speeches is used, in the present study, a radio broadcast news database is used. This database contains not only speech but also music, which is responsible for a lot of errors in a diarization systems. Both types of databases are well suited for online diarization systems, because relatively long speaker turns and low amounts of spontaneous speech are beneficial for a stable performance of such a system. A hybrid system for speaker diarization was presented in [223]. This system repeatedly performs offline diarization and uses the resulting speaker models to perform online speaker identification. Another similar system was proposed in [141]. In that sys-

tem, however, only speaker clustering is performed. The system uses the reference segmentation instead of performing segmentation itself. In [207], GMMs with MAP adaptation were used for offline speaker diarization.

One important property of the present system is that it performs online speaker diarization, in contrast to many other speaker diarization systems, like the one proposed in [219]. In addition, it does not only perform online speaker clustering or audio segmentation, but carries out both steps. Furthermore, as the system works with GMMs, the results of one pass of the system are not only speaker clusters, but also completely trained GMMs that can be used for speaker recognition. Due to the properties of the database that is used for validation, the system has to work not only with clean speech, but also with speech overlapped by music.

4.5.1 Methodology

The system introduced in this study can be divided into an offline part and an online part. In the first phase, the offline part of the system is used to train GMMs for each gender (*male* and *female*) and *garbage*. These three models are used as initialisation for the online phase, in which the *male* and *female* models play the role of universal background models.

The online part of the system performs speaker clustering. This is done as explained in the following. Initially, the continuous audio stream is segmented. Each time a new segment (typically with a length of several seconds) is created, model-based classification is performed and a speaker label is assigned to the segment based on the classification result. To be able to perform model-based classification, a model has to be trained for each speaker. The two gender models as well as the *garbage* model are constructed in the offline phase of the system. Models for individual speakers are generated sequentially in the online phase.

The operating principle can best be described by an example as shown in Figure 4.14. When the first segment of the audio stream arrives, recognition is performed with the existing three models. If the segment is classified as *male* or *female*, a new speaker model is created by copying the corresponding gender model and adapting the model with the audio data of the segment, using MAP adaptation. In the example, the first audio segment is classified as *male*, thus a new speaker model is created by copying the *male* model and adapting it with the audio data of the corresponding segment. At this stage, the following models exist: the gender models, the *garbage* model, and the model for the first speaker, named *s001*. If an audio segment is classified as *garbage*, as is segment 2 in this example, no model adaptation is performed. Segment 3 is classified as *female* here, leading to a new speaker model *s002*. The next segment is recognised as *s001*. In this case, the model is once again adapted with the new audio data. The system continues in this way: whenever a new segment is recognised as *male* or *female*, a new speaker model is created by copying and adapting the gender model. If an already seen speaker is recognised, its model is adapted as well.

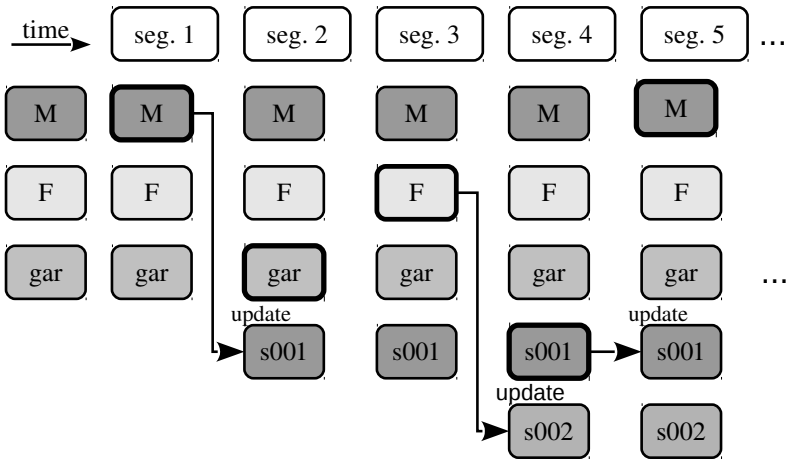


Figure 4.14: System operation of the proposed online speaker diarization system [62]. At the beginning, only three models exist, for male (M), female (F), and garbage (gar). New speaker models (e. g. s001 for speaker 1) are created when the speakers occur. At every time step, the model to which the segment is classified is highlighted by bold border lines.

The system implicitly performs several tasks: audio segmentation, speech/nonspeech classification, gender recognition, speaker novelty detection, and speaker identification. Audio segmentation is performed with an energy-based algorithm, as described below. Speech/nonspeech classification is done model-based: the decision between the speaker models and the *garbage* model constitutes the decision between speech and nonspeech. The decision between the two gender models and any of the speaker models (which have been derived from one of the gender models) corresponds to the gender recognition. For example, the process of creating new models could be turned off to get a gender recognition system. Speaker novelty detection is achieved by the maximum likelihood decision in the classification step. Considering the case in which several speaker models have already been created, then, if one of the gender models has the highest likelihood in the classification step for a new audio segment, the segment is classified as being from a new speaker.

The system works online, there is no clustering process that uses the whole audio recording. The recording is processed online (with a small latency) and labelled with speaker names. After a complete pass with one recording, the performance can be evaluated. The advantages of the system are the following: no trained speaker models are required (beside the gender and garbage models), the models are created on the fly. Additionally, the number of possible speakers need not be known to the system, which is often referred to as open-set speaker recognition. A result of the online pass of the system are trained speaker models (GMMs) that can be used for speaker recognition. The model adaptation process of the system was implemented with the hidden Markov model toolkit [252].

4.5.1.1 Segmentation and Feature Extraction

An energy-based segmentation method is applied to segment the audio stream. This method is a modified version of the segmentation as implemented in [252]: each audio frame is either declared as speech or silence. When the energy of the frame exceeds a certain threshold, it is declared as speech, otherwise as silence. Various rules are used to determine start and end points of speech segments. For example, a maximum segment length is applied, guaranteeing for a low system latency. Different maximum segment lengths are evaluated in the experimental section. As acoustic features, 12 standard MFCCs (+ energy) were used, together with their corresponding delta and delta-delta coefficients, which sums up to a total of 39 features. The features were extracted with a frame rate of 10 ms and window size of 25 ms.

4.5.1.2 Learning New Speakers

The online phase is the centrepiece of the system. In the recognition phase of the system, the speaker models are constantly adapted using the new audio data. In order to create a model for a new speaker, the corresponding gender model is copied and adapted with the new data of this speaker. This process uses MAP adaptation in the same way as it is used in GMM systems with universal background models [182] to adapt the means, mixture weights, and variances of the speaker GMMs. This adaptation process follows the equations presented in Section 2.2.1.4. Here again, the weighting factor τ was optimised on development data. Using small values of τ causes the system to produce a new speaker for almost every audio segment, because in this case, the new model is overfitted to the small amount of adaptation data and not generalised enough to be able to recognise other segments from the same speaker. If τ is set to zero, the new speaker model corresponds to a model trained only with the adaptation data.

4.5.2 Experimental Evaluation

The HUB4-1996 radio broadcast news database [80] was selected for testing purposes. This database consists of radio broadcast news recordings. A total number of eleven recordings was used in this study, six for training the gender and garbage models, three in the development set and two in the test set. Each recording is about 28 minutes long and most of them contain at least 15 different speakers. Furthermore, the recordings are interspersed with segments of music, as it appears often in radio broadcast news recordings. Music occurs not only alone, but also in the background of speakers.

The composition of the three files of the development set is 74 % *male*, 19 % *female*, and 7 % *garbage*. The *garbage* parts consist mostly of music segments and

Table 4.10: Speech/nonspeech classification and gender recognition error rates [%] on the development set for different numbers of mixture components.

# Mixtures	Speech/ Nonspeech	Gender
1	9.1	23.0
2	14.2	20.2
4	5.5	22.7
8	5.1	19.9
16	4.1	12.1
32	3.9	11.7
64	3.7	7.1
128	4.3	5.9
256	4.6	4.8
512	4.5	4.1
1024	4.7	3.6

pauses between speaker turns. Additional difficulty is added because speech is overlapped with music from time to time.

4.5.2.1 Speech/Nonspeech and Gender Recognition Results

In order to get results for the speech/nonspeech classification and gender recognition performance of the system, the online phase of the system was started without using the adaptation technique. In this way, no individual speaker models are created and recognition is performed with just the three models *male*, *female* and *garbage*. To get the speech/nonspeech classification error rate, the amount of time that *male* or *female* were classified as *garbage* and vice versa was summed up and divided by the total length of the audio test material. The gender recognition performance was obtained by determining the amount of time that *male* and *female* are confused with each other. With this experiment, the speech/nonspeech classification and gender recognition could be evaluated independently of the subsequent adaptation process of the system. This makes it easier to track down possible error sources in the system. The speech/nonspeech classification and gender recognition error rates with different numbers of mixture components for the GMMs are shown in Table 4.10. As can be seen in the results, more mixture components generally promised better results. However, the speech/nonspeech classification performance stagnated at a medium number of mixture components and even slightly worsened for higher numbers. Therefore, it was decided to use 128 mixtures: the classification performance was good enough and the computational effort was small enough to let the system work in real-time. In sum, the error rate with 128 mixtures was 10.2 %.

One part of the errors occurred due to segmentation errors: if the segmentation step erroneously under-segments the audio stream, there are relatively long segments that contain more than one speaker, for example a female speaker followed by a male speaker. In the following classification step, it is thus inevitable to make a small error for this segment, as only one label can be assigned to each segment. Another large portion of the errors was made because of music overlapping with speech.

4.5.2.2 Online Recognition Results

The online phase of the system including the adaptation process is the main component of the system that was evaluated. The system makes similar segmentation, speech/nonspeech classification, and gender recognition errors as shown above. Thus, when comparing the clustering results with similar studies, for example [145] or [141], it must be taken into account that the system presented in this study performs the whole processing chain, while for example in [141], the reference segmentation is used. Following the results from the previously presented experiment, it was decided to train the gender and garbage models with 128 mixtures for the online test, as this number represents the best trade-off between error rate and computation time. The best results were achieved with the MAP adaptation parameter τ set to $\tau = 135$.

The full system was evaluated using the DER as introduced in Section 4.2.2. Due to the nature of the evaluation data (relatively long speaker turns), no forgiveness collar was used. In addition, overlapping speech was ignored, since it is not very frequent in the database that was used for evaluation. Another measure for the performance of the clustering process is the cluster purity P_c, which is calculated as

$$P_c = \frac{\sum_{j=1}^{\hat{S}} \max_{1<i<S} f_{ij}}{\sum_{i=1}^{S} \sum_{j=1}^{\hat{S}} f_{ij}}, \tag{4.31}$$

with S being the true number of speakers while $\hat{S}$ is the number of speakers hypothesised by the system, and f_{ij} equals the length of all audio frames that belong to speaker i and were classified as speaker j. The cluster purity shows how well the clustering process generates clusters which contain only one single speaker.

Table 4.11 shows the results of the proposed speaker diarization system for the three recordings of the development set and the two recordings of the test set. The table reports the true number of speakers S as well as the hypothesised number of speakers $\hat{S}$, the DER for three different maximum segment lengths in the segmentation step, and the cluster purity P_c. The results show that the system was capable of creating a reasonable number of speaker clusters. The DER and cluster

Table 4.11: True and hypothesised number of speakers S and $\hat{S}$, respectively, DER [%] for different maximum segment lengths, and cluster purity P_c [%]. Results are reported for the files of the development (dev.) and test set.

			Max. segment length			
Audio file	S	$\hat{S}$	1 s	2 s	3 s	P_c
dev. file 1	15	19	42.3	21.9	29.9	82.3
dev. file 2	27	20	52.7	38.4	45.7	70.9
dev. file 3	17	21	51.4	29.8	32.1	83.1
dev. avg.			48.8	30.0	35.9	78.8
test file 1	18	22	42.1	35.0	40.1	72.7
test file 2	23	20	42.2	55.4	38.1	70.3

purity P_c both contain errors from each of segmentation, gender recognition, and speaker recognition. The best results (with the development set) were achieved with a maximum segment length of 2 s, resulting in a DER as low as 30 %, and the cluster purity reached almost 80 %, on average.

The system had problems with speech overlapped by music: if a person speaks several times in the recording, sometimes with overlapping music and sometimes without, the system tended to create two different speaker models, one for the speaker with overlapping music and one without. In order to reduce the influence of overlapping music, source separation algorithms may help by removing this music. In addition, methods for speech overlap detection that were presented in the first sections of this chapter could be adopted to detect speech overlapped by music and handle it accordingly.

In the last years, the introduction of graphical processing units for parallelisation of computing tasks made it possible to speed up the processing times of offline diarization systems. The study in [49] describes a way to parallelise an offline speaker diarization system in order to simulate an online system. The results show that the online system achieves similar results as the original offline system. This, however, comes at the cost of higher amounts of required computing power, while a dedicated online system as the one presented in this section works more efficiently.

4.6 Chapter Summary

This chapter presented methods to address open problems in the field of research of speaker diarization. One notable shortcoming of most diarization systems is the inability to process overlapping speech, which, however, occurs frequently in natural conversations. Different systems for detecting overlap were introduced in this thesis; new features for overlap detection were proposed, using CNSC for signal

decomposition, exploiting lexical information, or as selected from a large set of energy, spectral, and voicing-related audio features. Furthermore, LSTM neural networks were considered as an alternative or in combination with HMMs.

The detected overlap segments were used to improve speaker diarization through overlap exclusion and overlap labelling. Compared to a baseline overlap detection system relying on MFCC features and an HMM classifier, the methods presented in this chapter achieved considerable performance improvements in overlap detection. The obtained results mark a big step in the direction of solving the problem of overlapping speech for speaker diarization as well as for other speech processing applications that are affected by this problem.

In the present thesis, for overlap detection, it was not differentiated between different types of overlapping speech. The only categorisation that was made was an analysis of overlap detection performance depending on the segment length. Beyond that, analysing the system performance with respect to different types of overlapping speech might give helpful insights. For example, it might be beneficial to exploit different models for different types of overlap in order to improve the detection performance.

In addition, a speaker diarization system capable of online processing was presented. At the initialisation, this system works without any known speaker models. These are just added when the speakers appear for the first time.

5

Robust Speech Recognition

This chapter presents developments and findings in the field of robust acoustic modelling. Methods for robust recognition in environments with non-stationary noise or reverberation are elaborated.

5.1 Introduction

Automatic speech recognition (ASR) under realistic acoustic conditions (e. g. involving room reverberation and interfering noise sources) is still a major research challenge. Additive noise causes a spectro-temporal masking of the speech signal, while reverberation can be regarded as a convolutive channel distortion. The different categories of approaches to address robustness did not change over the years [79, 227, 138]. System robustness can be achieved by several strategies at different levels: speech/feature enhancement, robust features, or robust acoustic models. On the one hand, the speech signal can be enhanced using de-noising algorithms [146, 181]. Monaural signal separation techniques such as non-negative matrix factorization (NMF) [208] are especially useful for cases where multi-channel audio with a specified microphone placement is not available. Furthermore, robust features such as those obtained with the relative spectral transform – perceptual linear prediction method [97], feature enhancement techniques [130], or feature transformations such as linear discriminant analysis (LDA) [92] can improve the system robustness. On the other hand, robust models and decoding methods are often employed, including multi-condition training [140] and/or discriminative training [180, 117], e. g. using the maximum mutual information (MMI) principle [173]. In addition, methods such as vector Taylor series can be applied to adapt the acoustic model to noisy speech [122]. Such approaches which address the robustness of the back-end of the recognition system were mostly developed for conventional systems using Gaussian mixture models (GMMs) for acoustic modelling.

Recently, deep neural networks (DNNs) gained popularity in speech recognition due to the improved acoustic modelling performance compared to GMMs [99], al-

though the underlying neural network (NN) methods had already been developed years ago [19]. In [200, 232], the potential of DNNs for robust recognition was demonstrated. In addition, recurrent neural networks (RNNs) using the long short-term memory (LSTM) architecture [100] have been used for various machine learning tasks in audio processing (e. g. emotion recognition [244] and especially speech recognition [241, 87, 191, 169, 65]) or handwriting recognition [88] in the last years. In the previous chapter of this thesis, LSTM networks were used for speech overlap detection. Another application of LSTM networks is as a de-noising auto-encoder for feature enhancement [233].

Previously, in the context of (robust) speech recognition, LSTM networks have been applied mostly in a double-stream HMM architecture, where they are combined with a GMM acoustic model. This approach was first proposed in [241] and uses LSTM networks for phoneme prediction. The predicted phoneme probabilities are used for decoding jointly together with the GMM. The system presented in [237] exploits even a third stream of observations, which consists of predictions from non-negative sparse coding. In the past, LSTMs were mostly compared to simple acoustic models. In addition, a detailed investigation of the potential for robust recognition is still missing. Furthermore, until now, LSTM networks have never been directly applied as an acoustic model, predicting context-dependent HMM states for the hybrid acoustic modelling approach. A hybrid system that employs LSTMs for HMM state prediction could make use of the LSTM topology to exploit long-range temporal context and of the modelling power of a large network to be able to accurately predict HMM states.

5.1.1 Contributions

The first contribution of this chapter is to combine the LSTM approach with a state-of-the-art discriminatively trained ASR system. In particular, this study wants to address the following research question: is the LSTM system capable of improving a state-of-the-art noise-robust GMM-HMM ASR system? The experimental results (in two recognition tasks, influenced by highly non-stationary noise or by reverberation) affirm this question.

Furthermore, this study investigates the application of LSTM RNNs for acoustic modelling in the hybrid NN-HMM system architecture. The proposed system employs LSTM networks to predict context-dependent HMM states and uses the network predictions for acoustic modelling. The experimental results show that, in the hybrid architecture, LSTMs outperform GMM acoustic models and can compete with DNNs.

After discussing the related work, the LSTM principle is introduced in the next section. Section 5.3 presents the systems for recognition in highly non-stationary noise, while Section 5.4 describes the system for recognition in reverberant environments. The chapter finishes with a short summary in Section 5.5. Section 5.3 is

based on the results published in [59, 69], and Section 5.4 is based on the results published in [72].

5.1.2 Related Work

The two most widespread methods for using NNs for ASR are the *hybrid* and the *tandem* setup. In the hybrid setup, the network predicts HMM states, and thus these predictions are interpreted as posterior probabilities that are directly used for HMM decoding in a Viterbi search [19]. This method was constantly refined and advanced over the years [184, 185, 189, 221, 209], and currently, it is very successfully applied to large-vocabulary continuous speech recognition [36, 99]. The recent success of hybrid systems can be attributed to the application of DNNs, the deep topology (multiple hidden layers) provides a high modelling power. Mostly, a long context window is used as input to the DNN. Because of their deep topology, DNN acoustic models can learn higher-level representations of the features by themselves. In this way, they also learn to process context information that is either introduced through feature frame stacking (for DNNs) or is inherently incorporated in the model topology (for LSTMs). Exploiting such context is helpful to improve noise robustness, for example in cases where a portion of frames within a longer window is spectrally masked by noise.

In the tandem setup [98], the predictions of the network are used as observations for a GMM acoustic model. Therefore, often, the network predicts phonemes (instead of HMM states) to keep the number of targets (and in turn, the number of GMM observations) small. One variant of this method is the extraction of bottleneck features: a network is designed with a relatively small intermediate layer, and the output of this layer is used (instead of the output of the last layer) as features for the GMM [91], after decorrelation. One advantage of tandem systems is that the additional technology is introduced in the front-end part of the system, and thus the back-end recognition system can remain unchanged.

Recent examples where recurrent networks were applied in a tandem system are [225, 169]. Deep RNNs with end-to-end training are also capable of being used for speech recognition on their own without an HMM framework [89], in the tandem setup [246], or in the hybrid NN-HMM setup [87]. Another possibility of combining an NN acoustic model with a GMM is a multi-stream HMM. In this topology, the model has access to two independent streams of observations. Multi-stream HMM systems were initially proposed to combine independent feature streams [13]. For example, in this way, GMMs can be fused with NNs [93] or with NMF-based sparse coding techniques [211] for increased robustness.

One shortcoming of conventional RNNs is that the amount of context they use decays exponentially over time (the well-known 'vanishing gradient problem' [101]). To overcome this problem, the LSTM concept has been introduced [100]. An LSTM RNN exploits a self-learnt amount of temporal context, which makes it especially

suited for a speech recognition task that involves reverberation and additive non-stationary noise. Context information is helpful when only a portion of frames within a longer window is masked by noise. Previously, LSTM networks were suggested for noise-robust spelling recognition in a tandem HMM-LSTM system [245]. The application of LSTM networks in a double-stream system was first introduced in [241] for conversational speech recognition, where LSTM phoneme predictions improved a simple triphone HMM system. One study that tries to compare LSTM networks with conventional RNN and DNN architectures (in the hybrid setup) for large-vocabulary speech recognition is presented in [191], where the introduction of a modified LSTM architecture leads to superior results of the proposed system, compared to RNN and DNN systems.

Building upon the first CHiME Speech Separation and Recognition Challenge [11], in its second instalment [224], a medium-vocabulary speech recognition track was introduced by using the Wall Street Journal (WSJ0) read speech corpus. Together with degradation introduced by room reverberation and highly non-stationary additive noise, this proved to be a challenging recognition scenario. Preceding the present study, a GMM-LSTM multi-stream system was used in combination with NMF speech enhancement in the very successful contributions to the 1st and 2nd CHiME challenge [235, 243, 60, 73]. An LSTM network was used to generate frame-wise phoneme predictions, largely improving the performance of the maximum likelihood (ML) trained HMM baseline system. The HMM system employed NMF speech enhancement in its front-end. However, up to now, the LSTM approach has never been combined with discriminatively trained HMM systems. Since in previous studies, it was always combined with an ML-trained GMM-HMM system, it is not clear whether the LSTM approach will also lead to such large improvements in combination with a state-of-the-art discriminatively trained GMM system.

5.2 Long Short-Term Memory Recurrent Neural Networks

LSTM networks were introduced in [100]. Compared to conventional RNNs, LSTM RNNs are able to exploit a self-learnt amount of temporal context. In a conventional RNN, the hidden state vector sequence $h = (h_1, \ldots, h_T)$ is computed as

$$h_t = \mathcal{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \tag{5.1}$$

where W is a weight matrix, x_t is the input vector at time step t, b_h denotes the hidden bias vector, and $\mathcal{H}$ is the hidden layer function, which is usually an element-wise application of a sigmoid function. The output of the network is determined by the following equation:

$$y_t = W_{hy}h_t + b_y. \tag{5.2}$$

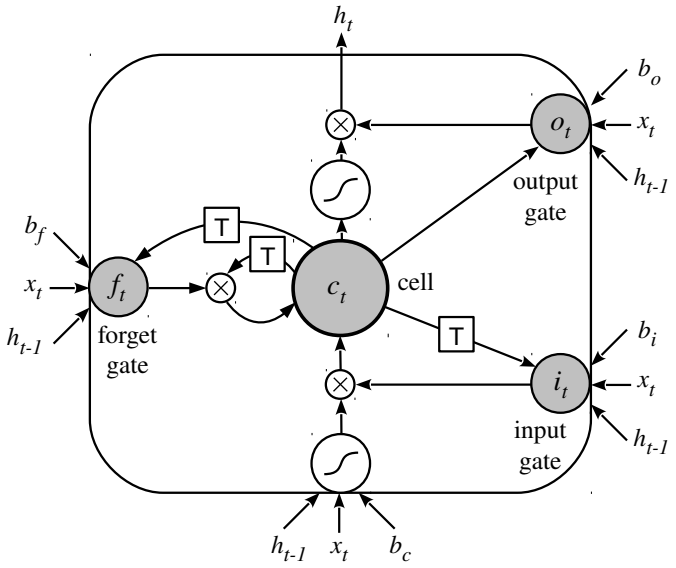


Figure 5.1: Long short-term memory (LSTM) block, containing a memory cell and the input, output, and forget gates [59]. 'T' denotes a delay of one time step.

In an LSTM network, the function $\mathcal{H}$ is replaced by so-called memory blocks. These memory blocks can store information in the cell variable c. In this way, the network can exploit long-range temporal context and thus, overcome the vanishing gradient problem, where the influence of previous inputs decreases exponentially over time, as in a conventional RNN.

Each memory block consists of a memory cell and three gates: *input*, *output*, and *forget* gate, as depicted in Figure 5.1. These gates control the behaviour of the memory block. The activation vectors of the gates are computed as

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{5.3}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{5.4}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \tag{5.5}$$

where σ is a sigmoid function, causing each gate either to be open or closed. The forget gate can reset the cell variable which leads to 'forgetting' the stored input c_t [74], while the input and output gates are responsible for reading input from x_t and writing output to h_t, respectively:

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{5.6}$$

$$h_t = o_t \otimes \tanh(c_t) \tag{5.7}$$

where $\otimes$ denotes element-wise multiplication and tanh is also applied in an element-wise fashion. Each memory block can be regarded as a separate, independent unit.

Therefore, the activation vectors i_t, o_t, f_t, and c_t are all of the same size as h_t, i. e. the number of memory blocks in the hidden layer. Furthermore, the weight matrices from the cells to the gates are diagonal, which means that each gate depends only on the cell within the same memory block.

In addition to LSTM memory blocks, the systems proposed in this study use bidirectional RNNs [198]. A bidirectional RNN can access context from both temporal directions, which makes it suitable for speech recognition, since whole utterances are decoded. This is achieved by processing the input data in both directions with two separate hidden layers. Both hidden layers are then fed to the output layer. The combination of bidirectional RNNs and LSTM memory blocks leads to bidirectional LSTM networks [86], where context from both temporal directions is exploited. A network composed of more than one hidden layer is referred to as a DNN [99]. By stacking multiple (potentially pre-trained) hidden layers on top of each other, increasingly higher level representations of the input data are created (deep learning). When multiple hidden layers are employed, the output of the network is (in the case of a bidirectional RNN) computed as

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t^N + W_{\overleftarrow{h}y} \overleftarrow{h}_t^N + b_y, \tag{5.8}$$

where $\overrightarrow{h}_t^N$ and $\overleftarrow{h}_t^N$ are the forward and backward state vector of the N-th (last) hidden layer, respectively. Furthermore, a softmax activation function is used at the output,

$$p(b_t^{(j)}|x_t) = \frac{\exp(y_t^{(j)})}{\sum_{j'=1}^{P} \exp(y_t^{(j')})}, \tag{5.9}$$

to generate probabilities at the output units $b_t^{(j)}, j = 1, \ldots, P$.

For network training, the systems employed in this study used online gradient descent by backpropagation through time, using mini-batch learning where weights were updated after processing a set of sequences. Training utterances were 'shuffled' (presented in random order) to improve generalisation in online learning. For the purpose of defining an error function for training, the output of the network (after applying the softmax function) and the targets are both regarded as a probability distribution over all possible output observations. Thus, the cross entropy between the targets and the output distribution was used as an error measure. In this way, the networks were trained discriminatively. For implementation, the publicly available CURRENNT toolkit was used [238][1].

[1] https://sourceforge.net/p/currennt, last accessed in April 2014

5.3 Recognition in Highly Non-Stationary Noise

In this section, a state-of-the art GMM system is combined with a deep bidirectional LSTM recurrent neural network in a double-stream architecture. Such networks use memory cells in the hidden units, enabling them to learn long-range temporal context and thus increasing the robustness against noise and reverberation. The network is trained to predict frame-wise phoneme estimates, which are converted into observation likelihoods to be used as an acoustic model. It is of particular interest whether the LSTM system is capable of improving a robust state-of-the-art GMM system, which is confirmed in the experimental results. In addition, LSTM networks are applied for acoustic modelling in the classical hybrid setup together with an HMM. The network predicts HMM states, and the predicted state posteriors are directly used for decoding within the HMM system. Experiments were conducted on the medium-vocabulary task of the 2nd CHiME speech separation and recognition challenge [224], which includes reverberation and highly variable noise. The performed experiments demonstrate the influence of different system components on the recognition performance and show that the proposed system strongly outperforms the challenge baseline as well as the best-performing challenge entry.

5.3.1 System Description

A flow chart of the evaluated ASR system is depicted in Figure 5.2. On the back-end side, a double-stream architecture is used for acoustic modelling. In addition to a GMM acoustic model, a deep bidirectional LSTM network is used as an acoustic model in the HMM framework. Both of them are always trained in a multi-condition fashion, using noise-free and noisy data. The GMM and LSTM can be combined in a double-stream HMM decoding architecture, where the HMM has access to both streams ob observations. The GMM acoustic model makes use of various feature transformations (as explained later). Since the LSTM is deep, such transformations (e. g. LDA + MLLT for decorrelation) are presumably not needed because the network can 'learn' useful feature representations on its own.

5.3.1.1 GMM-HMM-based Speech Recognition

The employed ASR system makes use of a state-of-the-art GMM-HMM, as it was described by Tachioka *et al.* in [213]. This system is implemented with the Kaldi speech recognition toolkit [171]. In addition to maximum-likelihood (ML) parameter estimation, it uses discriminative learning (DL) and various feature transformation (FT) methods. Discriminative training is performed using boosted maximum mutual information (bMMI) as proposed in [173]. The MMI principle aims at maximising the posterior probabilities of the correct utterances, given the trained models. By

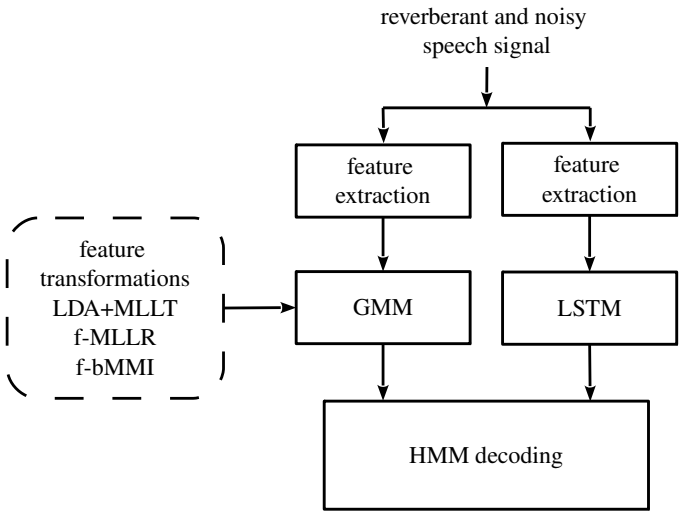


Figure 5.2: Block diagram of the evaluated system [59]. The central component is a multi-stream HMM fusing GMM and LSTM acoustic models. For the GMM stream, feature transformations (as explained in later sections) can be employed.

applying bMMI, a weight is introduced, strengthening the influence of hypotheses with a higher error. In addition to model-space bMMI, feature-space bMMI is applied as well.

Furthermore, techniques for feature transformation are employed. Feature transformations can improve the class separation and address the speaker variability in the training data. Channel variability, such as different channels and additive noise or reverberation, can also be compensated by feature transformations. LDA is applied on 'stacked' MFCC vectors (i. e. extracted from multiple signal frames using a sliding window centered around the current frame) and reduces these high-dimensional features to a smaller dimension. The necessary class labels are obtained by aligning the triphone HMM states. There are too few data to train full-covariance models, because of the high-dimensional acoustic feature space. Therefore, diagonal-covariance models, which do not consider correlations between features, are used instead. In this study, a maximum likelihood linear transform (MLLT) as described in [192] is used for decreasing the correlations between features. The combination of LDA and MLLT exploits context to reduce the influence of non-stationary noise, and correlations between feature dimensions that were introduced by noise are removed. To address the problem of large variations among speakers, speaker-adaptive training is applied: during the ML training procedure, feature-space maximum likelihood linear regression (MLLR), which is the same as constrained MLLR [51], is applied to estimate a speaker-dependent transform. The estimated transform is subsequently used during model re-estimation. First, a tight-beam decoding is performed to re-

estimate the speaker-dependent transform (the speaker identities are known), before doing a final decoding pass.

During decoding, the GMM acoustic likelihood $p_G(x_t|s_t)$ for observations given HMM states is computed as

$$p_G(x_t|s_t) = \sum_{m=1}^{M} c_{s_t m} \mathcal{N}(x_t; \mu_{s_t m}, \Sigma_{s_t m}), \tag{5.10}$$

modelling the continuous feature observations via a mixture of M Gaussians per state, where s_t denotes a context-dependent triphone state at time step t. The index m denotes the mixture component, c_{sm} is the weight of the m-th Gaussian associated with state s, and $\mathcal{N}(\cdot; \mu, \Sigma)$ represents a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ.

5.3.1.2 LSTM Acoustic Modelling

As an alternative to GMM acoustic modelling, an LSTM network is used. Thereby, two different approaches are employed. In the first approach, the network is trained to generate frame-wise phoneme estimates, as first proposed in [241]. The observation likelihoods are derived from these phoneme estimates. In the second approach, the LSTM directly predicts HMM states, and the observation likelihoods are computed from these predictions using Bayes' rule.

For a phoneme prediction LSTM, the input vectors x_t of the network correspond to the employed acoustic features, whereas the output y_t represents frame-wise activations for each phoneme (and therefore each output node corresponds to one class). In order to use phonemes as training targets, a forced alignment of the baseline HMM recogniser is created.

During decoding, a phoneme prediction $\hat{b}_t$ is derived from the network output activations,

$$\hat{b}_t = \arg\max_j \left(p(b_t^{(j)}|x_t) \right), j = 1, \ldots, P \tag{5.11}$$

leading to one phoneme prediction per frame. Here, P is the number of possible phonemes, and $p(b_t^{(j)}|x_t)$ are the outputs of the network. The process of LSTM decoding and generating the phoneme prediction is summarised in the function

$$\mathcal{L}(x_t) = \hat{b}_t. \tag{5.12}$$

These frame-wise phoneme predictions are used to obtain the likelihood $p(b_t|s_t)$ for the acoustic model in the following way: using development data, the frame-wise phoneme predictions are evaluated, and all confusions between true labels b

and predictions $\hat{b}$ are counted and stored in the phoneme confusion matrix C as row-normalised probabilities:

$$C(i,j) = p\big(\hat{b} = j | b = i\big). \tag{5.13}$$

Although the phoneme confusions are estimated on the development set in the current study, the performance generalises well to the test set. The likelihood $p(x_t|s_t)$ (observation given HMM state) is obtained through the mapping $b = m(s)$ from HMM states to phonemes. Since the LSTM works with monophones, triphone structures are ignored here, mapping triphone HMM states to the corresponding monophones. The acoustic likelihoods are therefore computed as

$$p_L(x_t|s_t) = C\big(m(s_t), \mathcal{L}(x_t)\big). \tag{5.14}$$

Thus, instead of directly predicting the probability $p(s_t|x_t)$ with the network and using Bayes' rule to obtain observation likelihoods, as in a typical hybrid system, the network converts the output scores $p(b_t|x_t)$ to discrete phoneme predictions $\hat{b}_t$ using Equation (5.11). These phoneme predictions are evaluated on the development set. By storing the confusions in C and normalising the rows of C, this matrix constitutes a discrete probability table for $p(\hat{b}_t|b_t)$. For HMM decoding, the likelihoods $p_L(x_t|s_t)$ are required, which are now approximated by

$$p(\mathcal{L}(x_t)|m(s_t)) = p(\hat{b}_t|b_t), \tag{5.15}$$

exploiting the surjective mapping from states to phonemes. Thereby, the confusions of the network are 'learnt' in the conditional probability table C and used to derive the observation likelihoods $p_L(x_t|s_t)$. These likelihoods are expected to have a high discriminative power. With this method, the RNN needs fewer output nodes (as compared to predicting state posteriors), which makes it easier to train.

Phoneme classification experiments in [84] support the choice of using bidirectional LSTM networks instead of other network architectures. In that study, bidirectional LSTMs are shown to perform better than feedforward networks or traditional RNNs without LSTM cells. To underpin this statement, in the experimental section of this thesis, additional results are shown where a feedforward network is employed for phoneme prediction instead of an LSTM.

In order to combine GMM acoustic modelling with LSTM phoneme predictions, a double-stream HMM system is employed. In every time frame t, the double-stream HMM has access to two independent information sources, $p_G(x_t|s_t)$ and $p_L(x_t|s_t)$, which are the acoustic likelihoods of the GMM and the LSTM predictions, respectively. In this case, the double-stream emission probability is computed as

$$p(x_t|s_t) = p_G(x_t|s_t)^{\lambda} \cdot p_L(x_t|s_t)^{2-\lambda}, \tag{5.16}$$

where the variable $\lambda \in [0, 2]$ denotes the stream weight of the GMM stream.

The second method for LSTM acoustic modelling that is considered in this study corresponds to the classical hybrid NN-HMM approach, in which the neural network is trained to predict HMM states s. The training targets are generated using a forced alignment of the HMM system, as in the case of phoneme predictions. From the posterior probabilities $p(s_t|x_t)$ resulting from the network predictions, the required state likelihoods are obtained by dividing by the state frequencies:

$$p(x_t|s_t) = \frac{p(s_t|x_t)}{p(s_t)}. \tag{5.17}$$

These state likelihoods represent the acoustic model and are used for decoding in the HMM. This corresponds to setting $\lambda = 0$ in Equation (5.16), and thus only the NN acoustic model is used.

One of the main differences between the two methods of using neural network predictions for acoustic modelling is the number of training targets. For phoneme predictions, the network has 40 output units (corresponding to the 40 phonemes), whereas the number of output units of a network that predicts context-dependent triphone HMM states is usually several thousands.

5.3.2 The CHiME Challenge

The experiments reported in this section are conducted on the medium-vocabulary speech recognition task of the 2nd CHiME challenge [224]. This database consists of utterances from the WSJ0 5 k vocabulary read speach corpus [165], convolved with real binaural impulse responses measured in a domestic environment and mixed with realistic noise backgrounds recorded in the same environment. The impulse responses were measured for a fixed position of 2 m in front of a head and torso simulator. The background noise contains a rich collection of sound sources from a lounge and kitchen such as electronic and kitchen appliances, noise produced by the inhabitants (such as footsteps, laughter or background speech), and noise from outside the house. Speech utterances are temporally placed in the background noise such that different signal-to-noise ratios (SNRs) from -6 to 9 dB, in steps of 3 dB, are obtained. The training set contains 7 138 utterances from 83 speakers summing up to 14.5 hours (forming the WSJ0 SI-84 training set), in clean, reverberated, and reverberated + noisy form. For the development set, 409 noisy utterances from 10 other speakers are provided at all six different SNRs, leading to a total number of 2 454 utterances (4.5 h in total). The test set includes 330 noisy utterances from 12 speakers at all SNRs (1 980 recordings or 4 h in total). All noisy utterances are also provided in an embedded form (not used in this study), where 10 s of surrounding background noise are included. The word error rate (WER) is used as an evaluation measure, counting the number of word substitutions, insertions, and deletions as

a fraction of the number of target words. For all evaluated systems, the WER is reported for each of the six different SNR values. In addition, the average WER across all SNRs is provided for better comparison of the systems.

5.3.3 Experimental Evaluation

5.3.3.1 Preprocessing and Feature Extraction

While the challenge data are stereophonic, only single-channel signals are considered in this study. These signals were obtained by averaging over both channels. For the employed database, this corresponds to a delay-and-sum beamforming, since the target speaker is located at a fixed position in front of the microphones (azimuth 90 degrees).

All features were extracted from frames of 25 ms and a frame shift of 10 ms. The baseline GMM-HMM system used standard MFCCs, i. e. 13 coefficients with their delta and delta-delta coefficients, whereas for the advanced GMM-HMM the features were processed using feature frame stacking and LDA projection (as described in the following section). The LSTM networks (both in the experiments for phoneme or state prediction) used log Mel filter bank coefficients (instead of MFCCs) that were also complemented by their delta and delta-delta coefficients. This follows other recent studies that use DNNs for speech recognition [99, 89, 150]. 26 logarithmic Mel filter bank coefficients (plus root-mean-square energy) covering the frequency range from 20 – 8 000 Hz were computed with the same frame size and shift as applied for the MFCCs. In addition, experiments were performed in which the LSTM network (in the phoneme recognition setup) uses MFCCs (same configuration as for the GMM-HMM system) as inputs.

5.3.3.2 Parametrisation

Parametrisation and training of GMM-HMM acoustic models in the system were the same as described in [213] and work as follows: 40 phonemes (including silence) were integrated in context-dependent triphone models with 2 500 states and a total number of 15 000 Gaussians. First, models were trained with clean training data applying the ML principle. Next, ML training was continued with reverberated training data, using the alignments and triphone tree structures from the clean models. Then, isolated noisy training data were used for training. In the experimental section, this basic system (using only ML training) is referred to as the ML GMM acoustic model. From this setup, an advanced system was created using discriminative training and feature transformations. First, another set of ML training iterations was performed after applying the described feature transformations, using the noisy training data. Here, the 13 static MFCC coefficients of nine consecutive frames were concatenated together and LDA was applied to reduce the

117-dimensional vector to 40 dimensions. The LDA used the 2 500 aligned triphone HMM states as classes. Subsequently, features were transformed using MLLT and models were re-estimated. Afterwards, a feature-space MLLR transform was estimated for speaker adaptation, leading to another set of model re-estimation iterations. Based on the resulting acoustic models, discriminative training was performed with the noisy training data, using model-space and feature-space bMMI. The language model weight used to generate the final hypothesis from the lattices was tuned for each system to minimise the average WER across all SNRs on the development set.

Parameters for the LSTM networks were estimated through multi-condition training, using the combination of the reverberated noise-free and noisy training sets. The inputs to the LSTM network were globally normalised in mean and variance. To this end, the global means and variances were computed from the reverberated noise-free and noisy training set features.

The bidirectional LSTM networks for phoneme prediction were parametrised as follows: in addition to the input and output layers, the employed bidirectional LSTM network was made of three hidden layers (making it a deep network), where 81, 128, and 90 hidden units were employed for the network with filter bank coefficients as input. These values correspond to the number of memory blocks in each of the two temporal directions (since the network is bidirectional). The number of input nodes corresponds to the length of the feature vector (81 in case of filter bank features), while the number of output nodes is equal to the number of phonemes, which is 40. In the case of using MFCC features as input to the LSTM, the size of the hidden layers was 78, 128, and 90, respectively. LSTM topologies were chosen according to previously performed experiments on similar databases.

For the state prediction networks, different configurations were tested, to investigate how the topology influences the recognition performance. Generally, these networks are equipped with broader layers, leading to more trainable parameters, which is necessary because of the larger output layer. The networks were tested with one, two, or three layers with a size of 150 memory blocks (per direction). In addition, one experiment was performed where the layer size was increased to 250.

All networks were trained through gradient descent with a learning rate of 10^{-5} and a momentum of 0.9. During training, zero mean Gaussian noise with a standard deviation of 0.6 was added to the inputs in order to further improve generalisation. All weights were randomly initialised from a Gaussian distribution with mean 0 and standard deviation 0.1. The average cross entropy error per sequence on the development set was evaluated after every fifth epoch in the training phase. Using an early stopping strategy, training was aborted as soon as no improvement on the development set could be observed during 25 epochs.

Table 5.1: WER [%] on the CHiME development set when combining different GMM acoustic models with the phoneme prediction LSTM. The GMMs were trained either with maximum likelihood (ML) training or using discriminative learning (DL) and feature transformation (FT). DNN results are included for comparison.

Acoustic model		**SNR [dB]**						
GMM	LSTM	-6	-3	0	3	6	9	**Mean**
ML	-	68.5	59.0	50.3	44.3	39.7	34.5	49.4
DL+FT	-	52.9	43.0	34.6	26.7	23.5	19.0	33.3
ML	✓	53.7	43.2	35.2	29.8	26.7	22.1	35.1
DL+FT	✓	**45.2**	**34.4**	**27.5**	**21.5**	**19.2**	**15.7**	**27.3**
DNN [214]		57.2	45.9	36.2	30.6	26.4	23.3	36.6
DL+FT	DNN	50.9	40.6	32.6	25.8	22.6	18.7	31.9

5.3.3.3 GMM versus LSTM Results

First, the effects of combining the employed LSTM method with the two different GMM acoustic models are studied: the standard system using only ML training or the advanced discriminatively trained system employing LDA, MLLT, and speaker adaptation. Experimental results for the four system combinations are displayed in Table 5.1. The ML-trained GMM acoustic model (first row, 49.4 % average WER) and the discriminatively trained system including all feature transformations (second row, 33.3 %) correspond to the systems described by Tachioka *et al.* in [213], except that beamforming (cf. Section 5.3.3.1) was applied, which brought an absolute improvement of about 7 % in average WER. Combining the two techniques for acoustic modelling in the double-stream system (GMM stream weight $\lambda = 1$) led to further large improvements. The simple GMM was improved upon by almost 30 % relatively (35.1 %). Furthermore, the discriminatively trained HMM could also vastly be improved upon (18 % relatively) by adding the LSTM phoneme predictions (27.3 %). The relative improvements are nearly the same for all SNRs. For comparison, results for a standard DNN, taken from [214], are also listed in Table 5.1. The DNN acoustic model had three hidden layers and 500 k parameters and is thus comparable to the LSTM employed in the present study. As it is not speaker-adapted (though it still uses the LDA+MLLT feature transformation), the DNN (36.6 %) is not able to beat the GMM. Beyond that, the performance is also weaker than the GMM-LSTM double-stream system. A phoneme prediction DNN (employed in the same way as the LSTM and described in more detail later in this section) performed significantly worse than the LSTM.

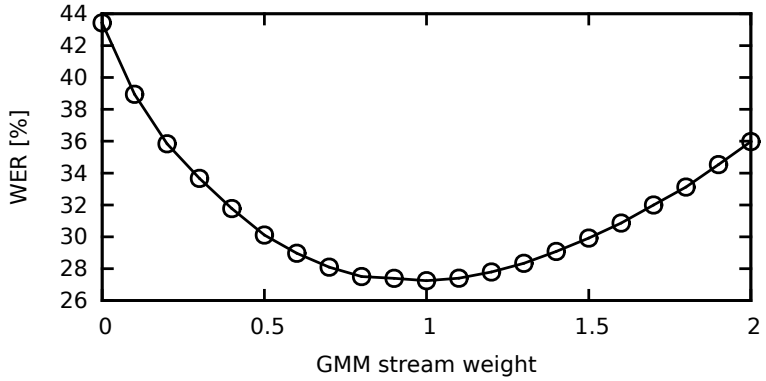


Figure 5.3: Average WER (development set) for different HMM decoding stream weights λ [59]. Values of 0 and 2 correspond to using only the LSTM or GMM acoustic model, respectively.

The stream weight λ in Equation (5.16) controls the trade-off between the influence of the GMM and LSTM acoustic model likelihoods. When setting $\lambda = 2$, the system uses only the GMM acoustic model (though with an exponent of 2). Accordingly, $\lambda = 0$ means that the system uses only the information from the LSTM stream. Figure 5.3 shows the average WER for different stream weights. Of particular interest are the results with $\lambda = 0$ and $\lambda = 2$. The GMM alone performed better than the LSTM. This might appear contrastive to the conclusion that DNN acoustic models perform better than GMMs [99]. However, only the advanced GMM (including DL+FT) beat the LSTM, while the LSTM approach outperformed the standard ML-trained GMM. In addition, the employed LSTM only models monophones. Further improvements are expected when modelling context-dependent HMM states. The best performance of the double-stream system was achieved with a stream weight of $\lambda = 1$ (27.3 %). Therefore, this value was used in all other reported experiments. These results show that, even if the LSTM acoustic model alone performs worse than the GMM, the GMM system can greatly benefit from the combination with the LSTM predictions in the double-stream setup.

In order to demonstrate the merits of the chosen LSTM architecture, different feedforward DNNs were trained for phoneme recognition. Table 5.2 shows the framewise phoneme error rate on the development set for these experiments. A layer size of 400 hidden units was chosen for the DNNs, with either three or four hidden layers. Feature frame stacking (incorporating seven neighbouring frames) was applied to exploit temporal context. The phoneme recognition results show that the DNN (51.8 %) was not able to reach the performance of the LSTM network (35.8 %). What can also be seen is that adding the fourth layer to the DNN (and thereby adjusting the number of parameters to the LSTM) brought no improvement (52.1 %). In that case, the missing pre-training or initialisation of the DNN becomes

Table 5.2: Framewise phoneme error rate on the CHiME development set, comparing an LSTM with different DNNs with or without feature frame stacking.

Network	**Layers**	**# Weights**	**Error** [%]
DNN	400-400-400	370 k	59.5
DNN (feature stacking)	400-400-400	490 k	51.8
DNN (feature stacking)	400-400-400-400	650 k	52.1
LSTM	81-128-90	660 k	35.8

noticeable. Compared to the DNN, the LSTM is better structured and thus easier to train. The DNN with 3 layers and feature frame stacking was also used to obtain the results in the last row in Table 5.1, where it could also be seen that the LSTM performs better than the comparable DNN.

The findings support the results presented in [87], where a similar effect was observed. In that study, it was suspected that the main reason for this discrepancy between framewise phoneme error and WER (in the test set experiments of the present study, it is shown that a well-tuned DNN and an LSTM perform similarly well in terms of WER) is that the frame-wise error does not take into account the language model. The LSTM might learn a word-level language model itself, which interferes with the language model during decoding. In fact, the experiments performed in the present study also showed that for LSTM decoding, much higher language model weights are necessary compared to DNN decoding.

To better understand what the LSTM is learning, an analysis of the amount of exploited context was performed. In [242], a methodology is proposed to perform such an analysis. From the sequential Jacobian [84],

$$J_{ji}^{tt'} = \frac{\partial y_t^{(j)}}{\partial x_{t'}^{(i)}} \tag{5.18}$$

which corresponds to the derivative of the network outputs $y_t^{(j)}$ with respect to network inputs $x_{t'}^{(i)}$ at different time steps (given as a relative position compared to time step t), the sensitivity can be computed. This is achieved by summing up the absolute magnitudes of the derivatives over all input units i, output units j, and all time steps t and normalising them:

$$S_{t'} = \frac{\sum_t \sum_j \sum_i |J_{ji}^{tt'}|}{\max \sum_t \sum_j \sum_i |J_{ji}^{tt'}|} \tag{5.19}$$

This sensitivity can be considered as a measure of the contribution of input nodes to the activity at the output of the network. Figure 5.4 shows the sensitivity (mean

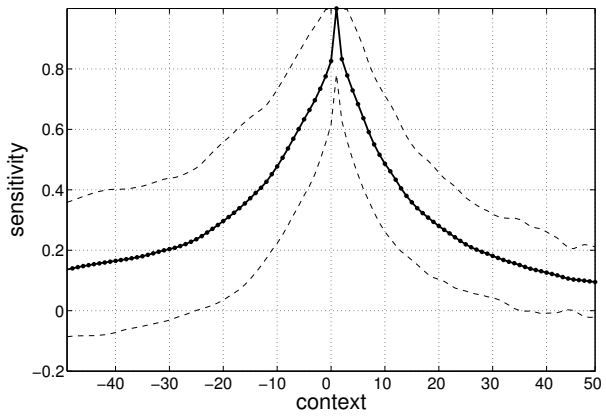


Figure 5.4: Sensitivity (mean ± standard deviation) of network outputs to input nodes of neighboring time frames [59].

± standard deviation over time steps t) of a randomly chosen sequence (with SNR of -6 dB) in the development set. In particular, the plot shows the average sensitivity of the outputs with respect to the inputs from ± 50 frames of context. For example, considering a sensitivity threshold of 0.2, the network exploited roughly 30 frames (300 ms) of past and future information. The standard deviation (dashed lines) shows that there was a higher variability in using past context. In comparison to standard DNNs, which usually exploit context of around 7–11 frames via feature frame stacking [99], the employed LSTM architecture has access to a much larger amount of context information. For a standard DNN, the amount of context could be increased with larger windows for feature frame stacking. This would, however, increase the number of trainable parameters of the network, and thus make it more difficult to train. An advantage of the LSTM topology is that the amount of exploitable context is independent of the number of parameters.

5.3.3.4 LSTM Input Features

While for the results in Table 5.1, the LSTM network used log Mel filter bank coefficients as input features, additionally, experiments were performed in which the LSTM used the same features as the GMM, namely MFCCs. The results can be seen in Figure 5.5. The network that uses filter bank coefficients as input performed consistently better than the one using MFCCs. Averaging over all SNRs, the relative improvement in combination with the basic model (ML) was 10 %, and with the advanced model (DL + FT) it was 8 %. This shows that since the system uses a deep network, it is favourable to let the network 'learn' better representations for the features instead of using the discrete cosine transform for approximate feature de-correlation, as is done in MFCC computation. These results support the previ-

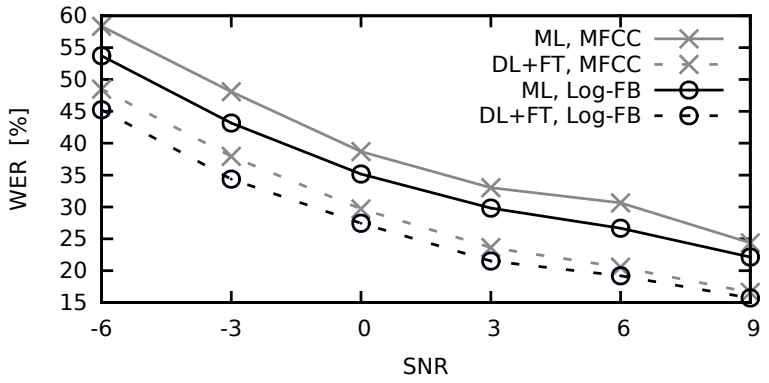


Figure 5.5: Influence of input (MFCC or logarithmic Mel filter bank (Log-FB) coefficients) to the phoneme prediction LSTM network on WER (development set), evaluated in combination with the two different GMM systems (ML or DL + FT).

ous finding that NN-based acoustic models perform better with filter bank features instead of MFCCs [150, 99].

5.3.3.5 Results for LSTM Hybrid Acoustic Modelling

Results for the evaluated state prediction networks, used as an acoustic model in the hybrid setup, are shown in Table 5.3. In order to compare them to the double-stream approach where a GMM is combined with a LSTM for phoneme prediction or to a conventional DNN, these results should be set in comparison with the results listed in Table 5.1. First, the experiments show that an LSTM network with one layer achieved a better performance (40.0 %) than the ML GMM (49.4 %) but fell back behind the advanced GMM (33.3 %). Adding a second layer led to a substantial improvement (31.5 %). This network (which is later employed in the test set experiments) had 1.4 million weights in total. Increasing the layer size of this network (3.2 million weights) or adding a third layer (2 million weights) brought no benefit. These results show the limits of increasing the size of the network. Compared to the best setup with a phoneme prediction LSTM (27.3 %), the best state prediction LSTM (31.4 %) could not quite reach the same performance (although the systems are difficult to compare, since the phoneme prediction LSTM is combined with a GMM). It did, however, perform better than a standard DNN (36.6 %).

5.3.3.6 Test Set Results

Finally, Table 5.4 shows the results for the CHiME challenge test set. Generally, the results show the same tendencies as for the development set.

From the proposed systems, the GMM-only system (with DL + FT) is included. Furthermore, results are reported for the two different LSTM approaches: either

Table 5.3: Results for hybrid acoustic modelling with different state prediction networks (WER [%] on the development set).

		SNR [dB]						
Network	**Layers**	-6	-3	0	3	6	9	**Mean**
LSTM	300	60.2	48.9	39.2	34.5	31.4	25.8	40.0
LSTM	300-300	49.0	**39.2**	31.5	**26.4**	23.1	19.7	31.5
LSTM	500-500	**48.9**	**39.2**	31.8	26.6	**22.5**	**19.6**	**31.4**
LSTM	300-300-300	50.4	40.7	**31.1**	**26.4**	23.2	20.0	32.0

Table 5.4: Test set evaluation (WER [%]) of the proposed ASR systems with NMF enhancement and LSTM phoneme predictions, and comparison to related approaches.

	SNR [dB]						
System	-6	-3	0	3	6	9	**Mean**
Other systems							
Baseline noisy GMM [224]	70.4	63.1	58.4	51.1	45.3	41.7	55.0
NMF, noisy GMM [111]	61.9	55.6	50.9	43.5	39.1	37.4	48.1
NMF, GMM (ML)+LSTM [60]	57.4	49.0	42.5	37.4	32.6	29.7	41.4
GMM (DL+FT) [213]	54.7	45.1	36.0	28.6	24.4	21.4	35.0
Blind source extraction [153]	42.2	38.4	32.7	29.2	26.9	23.7	32.2
Bin. mask., GMM (DL+FT) [214]	44.1	35.5	28.1	21.2	17.4	14.8	26.9
DNN [232]	42.1	31.7	24.7	19.4	16.4	14.3	24.8
RDNN [232]	38.1	29.1	23.0	17.9	15.0	13.6	22.8
LSTM enh., GMM (DL+FT) [234]	**35.6**	**27.1**	**22.4**	17.5	16.1	14.3	22.2
Systems proposed in this study							
GMM (DL+FT)	46.4	36.2	28.5	21.6	17.9	15.7	27.7
GMM (DL+FT) + LSTM	37.1	27.2	22.5	**16.7**	**13.9**	**11.8**	**21.5**
LSTM (states)	40.3	32.2	25.0	19.8	16.8	15.8	25.0

with the state prediction network as an acoustic model alone, or with the phoneme prediction network in combination with the GMM in the double-stream setup. The results for the official CHiME challenge baseline (multi-condition ML-trained GMM-HMM using MFCCs) are shown in the first row of the table (55.0 % average WER). This was improved with NMF enhancement exploiting long-context speech and noise models by Hurmalainen *et al.* [111] by 13 % relatively. In the original contribution to the challenge [60], which was the basis for the present study, an NMF speech enhancement approach was used, together with an earlier version of the LSTM

double-stream system, in combination with the official challenge baseline. This system reduced the WER to 41.4 %. An alternative recognition system for the challenge was provided by Tachioka *et al.* in [213], which, compared to the official baseline, used LDA, MLLT, speaker adaptation, and discriminatively trained GMM-HMMs, resulting in a WER of 35.0 %. This result was surpassed (8 % relatively) by the approach proposed by Nesta *et al.* in [153]. Their system worked mainly on the front-end side, exploiting blind source extraction, and using the challenge baseline recogniser. Including binary feature masking (bin. mask.) in the front-end of the system in [213] improved the result by 23 %, relatively, which was the challenge entry with the best results [214]. In [232], a well-tuned DNN and a recurrent DNN were evaluated on the CHiME task. This DNN is different from the DNN presented in [214] and included in Table 5.1, which was a preliminary version. For example, the systems presented in [232] made use of multiple passes of alignment and model retraining. These systems outperformed the best GMM baseline. The application of LSTM networks for feature enhancement was proposed in [234]. Together with an advanced GMM acoustic model, this resulted in an average WER of 22.2 %. The WER improvement obtained through the LSTM enhancement is especially noticeable at lower SNR values.

The systems elaborated in the present study are also based on the GMM system described in [213]. First, performing the simple beamforming method as described in Section 5.3.3.1 led to a relative improvement of 21 %, down to an average WER of 27.7 %. The GMM-LSTM system brings a larger improvement, yielding a WER of 21.5 %. Compared to the official challenge baseline, this is a relative improvement of 61 %. The best challenge entry is beaten by 21 % relatively. Notably, the best proposed system also surpasses the DNN and recurrent DNN results presented in [232], as well was the LSTM feature enhancement approach proposed in [234]. The LSTM system for state prediction achieved an accuracy of 25.0 % on the test set, which makes it comparable to the DNN evaluated in [232].

5.3.4 Conclusions

This section presented a system for noise-robust ASR that combines GMM acoustic modelling with phoneme predictions from a deep bidirectional LSTM network. In addition, LSTM networks for state prediction were employed as an acoustic model alone.

In particular, the focus was on the following research question: when a state-of-the art discriminatively trained GMM-HMM system including feature transformations is used instead of a simple baseline, can the LSTM predictions still lead to an improvement? The results (cf. Table 5.1) reveal that the LSTM brings large improvements to both GMM systems. Overall, the experimental results show that the novel combination of a state-of-the-art GMM and an LSTM is highly efficient. The system achieved large improvements in WER and outperformed all entries to

the 2nd CHiME challenge while being compliant with the challenge guidelines. On the test set, the challenge baseline, a standard HMM system, had an average WER of 55.0 %, whereas with the best proposed system, a WER of 21.5 % was obtained.

In addition, LSTM networks were proposed as an acoustic model alone. In this system, bidirectional LSTM networks are trained with HMM states as training targets, and the resulting predictions are converted into state likelihoods for decoding in the HMM framework in the hybrid setup. The experimental results show that the performance of this system is comparable to that of a standard DNN, while it is slightly worse than the GMM-LSTM double-stream approach.

5.4 Recognition in Reverberant Environments

While in the previous section, the application scenario for the speech recognition system was an environment primarily influenced by highly non-stationary noise, the scenario in this section is characterised by high reverberation. Reverberation severely degrades the performance of automatic speech recognition. The REVERB challenge [125] addresses the problem of reverberated speech by providing a testbed for speech enhancement and speech recognition methods in a reverberant environment. In this section, it is shown how the LSTM principle can be applied to speech recognition in a reverberant environment. Due to its improved capability of context modelling, an LSTM network is especially suited for this task, because reverberation has a large impact on the context of a speech utterance.

The LSTM is trained with phonemes as targets, and the predictions are converted into observation likelihoods and used as an acoustic model in a double-stream system, in combination with the GMM. Furthermore, the system is compared to a de-reverberation method called correlation shaping [78], which works with eight-channel recordings. This method is based on linear prediction and reduces the length of the equalised speaker-to-receiver impulse response. Using de-reverberation as a front-end of the GMM in combination with the LSTM predictions leads to substantial improvements of the WER for the official REVERB challenge database.

5.4.1 System Description

Figure 5.6 shows an overview of the proposed system. In addition to a standard GMM-HMM system, the HMM can make use of phoneme predictions from an LSTM network in a double-stream architecture. This LSTM network predicts phonemes and the predictions are converted to observation likelihoods for HMM decoding. Compared to the baseline GMM-HMM provided by the organisers of the REVERB challenge, a slightly improved system is used in this study. This system uses a different method for model adaptation and the main difference is that it uses a trigram language model instead of a bigram.

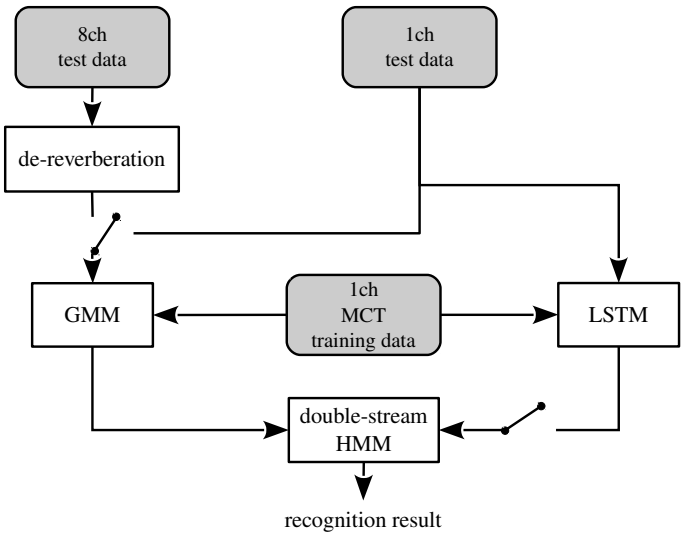


Figure 5.6: System overview: a double-stream HMM system combining GMM with LSTM, and de-reverberation using 8-channel recordings [72].

The GMM is trained either with clean or multi-condition training data, while the LSTM uses multi-condition training data in all experiments. Furthermore, a de-reverberation method (correlation shaping) is applied, processing the 8-channel recordings, and the GMM is either fed with the 1-channel reverberated test data or with the processed test data from the 8-channel recordings.

5.4.1.1 HMM-GMM Recognition System

In addition to the REVERB baseline recognition system, which is implemented with the hidden Markov model toolkit [252], experiments were performed with a (slightly improved) re-implementation with the Kaldi toolkit [171].

The baseline recogniser was a GMM-HMM system that employed tied-state HMMs with ten Gaussian components per state and was trained according to the maximum-likelihood criterion. As features, standard MFCCs (computed every 10 ms from windows of 25 ms) including delta and delta-delta coefficients were used. Two methods were utilised to address the reverberation in the audio recordings. First, multi-condition training was employed by training the recogniser not only with clean training data, but also with the reverberated version of the training data. Second, constrained MLLR adaptation (in batch processing) was used to adapt the features to each test condition. The WSJ0 bigram language model was used during decoding.

A re-implementation using the Kaldi toolkit of this system was used for the experiments in this study. Instead of constrained MLLR, the Kaldi system employed basis feature space MLLR [172] for adaptation. This method performs well even on small amounts of adaptation data and thus was used for utterance-based batch

processing instead of full batch processing. This implicates, however, that the implementation is not capable of online processing since it always waits for the end of the current utterance. The biggest improvement that was made compared to the baseline system is the introduction of a trigram language model instead of the bigram language model that was used in the baseline.

The Kaldi HMM system was tested in similar configurations as the challenge baseline system. First, a clean triphone recogniser was trained with the WSJCAM0 training set. Then, the reverberated training set was used to train a multi-condition acoustic model. For this model, the bases for MLLR adaptation were estimated, and finally, the trigram language model was used for decoding with this model. In the case of using front-end de-reverberation, the employed method was always only applied on the test data, while the original acoustic model (trained with unenhanced data) was used.

5.4.1.2 LSTM Double-Stream Recognition

In addition to GMM acoustic modelling, an LSTM network was used to generate frame-wise phoneme estimates. From these phoneme estimates, the observation likelihoods for the acoustic model were derived. These were used together with the GMM in a double-stream architecture. The general approach of using LSTM networks for phoneme prediction in the double-stream architecture was the same as described in Section 5.3.1.2.

Instead of MFCCs, the LSTM used Mel filter bank features, complemented by their delta coefficients. In the previous section of this thesis, it was already shown that this leads to a small performance gain. 26 logarithmic Mel filter bank coefficients (plus root-mean-square energy) covering the frequency range from 20 – 8 000 Hz were used, computed with a frame size of 25 ms and frame shift of 10 ms. Thus, in total, the dimension of features for the LSTM was 54. Features for the LSTM were extracted from the one-channel recordings of the REVERB challenge database. As an additional preprocessing step, per-utterance peak normalisation of the waveforms of the audio recordings was considered. To this end, the recordings were amplified to set the largest occurring absolute value to -3 dB below the maximum amplitude. This was necessary because the recordings from the REAL partition of the dataset are badly adjusted.

The topology of the tested bidirectional LSTM network was as follows: as the dimension of the feature vector was 54, this was also the size of the input layer. Three hidden layers were employed, where two systems were tested, with 100 or 200 LSTM blocks. The number of output units corresponds to the number of phonemes, which was 45 in the implemented system. For training the networks, the multi-condition training set was employed. The networks were trained through online gradient descent with a learning rate of 10^{-5} and a momentum of 0.9. During training, zero mean Gaussian noise with standard deviation of 0.6 was added to the inputs

in order to further improve generalisation. All weights were randomly initialised from a Gaussian distribution with mean 0 and standard deviation 0.1. After every training epoch, the average cross entropy error per sequence on a validation set was evaluated. Training was aborted as soon as no improvement on the validation set could be observed during 10 epochs. This validation set was a held-out part of the multi-condition training set, consisting of the utterances from 10 speakers. The stream weight for double-stream decoding was set to $\lambda = 1.2$.

5.4.1.3 De-reverberation

The proposed method of using LSTM acoustic models to increase the robustness of the recognition system is compared to a method for multi-channel de-reverberation called correlation shaping (CS) [78, 72].

In short, this approach works as follows: the goal of the method is to decrease the length of the equalised speaker-to-receiver impulse response. This is achieved by reducing the long-term correlation in the linear prediction residual of the reverberant speech recordings. The CS method is implemented as a multi-input single-output linear filter and thus processes multi-channel audio recordings.

5.4.2 The REVERB Challenge

The goal of the 2014 REVERB challenge [125] was to evaluate methods for speech enhancement and robust speech recognition in reverberant environments. Thus, there are two tasks in the challenge, enhancement and recognition. The contribution of this thesis is limited to the recognition track, where the task is to recognise read medium vocabulary (5 k) speech in different reverberant environments (reverberation times T60 ranging from 0.25 to 0.7 s). There are eight different environments, whereof six (called the SIM condition) are simulated by convolving the WSJCAM0 corpus [186] (which is a British English version of the WSJ corpus [165]) with measured room impulse responses. These impulse responses were measured in three different rooms, each at a near (50 cm) and far (200 cm) microphone distance. Additionally, stationary noise from the same rooms are added at an SNR of 20 dB. The other two conditions (called the REAL condition) correspond to recordings from the MC-WSJ-AV corpus [139]. This database contains real recordings from a reverberated room, measured at two distances (near = 100 cm and far = 250 cm). For all data (SIM and REAL), eight-channel recordings from a microphone array are available. In addition, it is also possible to evaluate one-channel systems. In this case, only the recording from the first microphone is considered. For training the recognition system, the WSJCAM0 training set containing 7 861 utterances from 92 speakers is provided. In addition, a multi-condition training set is available, which is created similarly like the SIM data from the WSJCAM0 training set. Test experiments are performed using data from the eight different environments, where

Table 5.5: Phoneme classification error [%] on the REVERB development set for different LSTM configurations with and without waveform normalisation.

Normalisation	Layers	# Weights	Data SIM	Data REAL
-	100-100-100	170 k	25.91	67.79
✓	100-100-100	170 k	25.55	51.39
✓	200-200-200	600 k	24.97	52.35

the six conditions from the SIM data together have 1 484 and 2 176 utterances in the development and test set, respectively, each from 20 speakers. The REAL data consists of 179 and 372 utterances (development and test) from five/ten speakers. Systems are evaluated using the WER, counting the number of word substitutions, insertions, and deletions as a fraction of the number of target words.

5.4.3 Experimental Evaluation

First of all, different configurations of the LSTM recognition system were tested and evaluated in terms of frame-wise phoneme classification performance on the development set. The results are listed in Table 5.5. A smaller and a larger LSTM network were considered (the number of weights is included in the table), and the influence of the audio normalisation was investigated. The results show that the normalisation had a positive effect on the results for the REAL data (51.39 %), while the SIM results were unaffected (25.55 %). The recordings from the REAL partition were badly adjusted and all of them had a low amplitude. This problem was solved by the normalisation, but the phoneme recognition rates for the REAL data are still far worse than for the SIM data. Increasing the number of LSTM units in the hidden layers to 200 brought a small improvement for the SIM data (24.97 %). It was decided to use the larger network in the other experiments. The reason for this is because the LSTM was validated with a small partition of the original multi-condition training (MCT) data (using a forced alignment of the development data for system training was not allowed in the challenge rules), which is comparable to the SIM data.

Experimental results for combining the Kaldi GMM recognition system in the double-stream setup with the LSTM predictions are given in Table 5.6. Results are reported for four different configurations of the GMM system, with increasing complexity: using the simple GMM, employing MCT or basis feature space MLLR for adaptation to the test conditions (or their combination), or finally, employing the trigram language model. These steps gradually improved the recognition per-

Table 5.6: REVERB development set: influence of combining the GMM with LSTM predictions, in terms of WER [%].

GMM					+LSTM	
Adapt	**MCT**	**Language model**	**SIM**	**REAL**	**SIM**	**REAL**
-	-	bigram	50.61	88.50	38.46	79.90
-	✓	bigram	27.85	53.00	21.07	47.20
✓	✓	bigram	22.07	45.52	17.38	42.70
✓	✓	trigram	16.85	38.33	12.98	36.14

Table 5.7: REVERB test set: influence of combining the GMM with LSTM predictions, in terms of WER [%].

GMM					+LSTM	
Adapt	**MCT**	**Language model**	**SIM**	**REAL**	**SIM**	**REAL**
-	-	bigram	49.95	88.50	36.80	79.81
-	✓	bigram	27.53	53.78	20.79	48.97
✓	✓	bigram	22.36	46.14	17.77	43.83
✓	✓	trigram	17.26	39.76	13.75	36.78

formance of both the SIM and REAL data. For the SIM condition, including LSTM predictions led to a substantial improvements for each GMM configuration. Apart from that, the improvements with the REAL data were smaller. Here, the mismatch between training and test data had a larger influence on the LSTM recognition performance. Finally, experimental results with the test set are shown in Table 5.7 for the same system configurations. Overall, the results are comparable to the development set results, and the same tendencies are visible. The improvements obtained by adding LSTM predictions were roughly the same as for the development set experiments.

To give a detailed coverage of the results on the test set, Table 5.8 includes test set results for different system configurations, broken down into the eight different recording conditions. Firstly, the results of the REVERB challenge baseline recogniser are also included (first row). The Kaldi implementation of the GMM (row two) already achieved substantial improvements compared to the baseline system. The results also include experiments where de-reverberation using CS was applied (row three). Results for the LSTM system were similar to CS for the SIM data, but for the REAL data, a large improvement was obtained with the employed de-reverberation method. It could furthermore be observed that, while the relative

Table 5.8: Test set results in terms of WER [%] for selected systems and for another state-of-the-art system, for all eight test conditions. The GMM recogniser is improved with correlation shaping (CS) de-reverberation and/or LSTM predictions.

	SIM							Real		
	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
System	near	far	near	far	near	far		near	far	
Baseline	16.2	18.7	20.5	32.5	24.8	38.9	25.3	50.1	47.6	48.9
GMM	10.2	12.3	13.0	23.3	15.3	29.5	17.3	40.6	39.0	39.8
GMM+CS	10.9	11.5	10.7	15.6	11.4	19.1	13.2	**28.0**	28.3	28.2
GMM+LSTM	**8.3**	10.0	10.6	18.7	12.2	22.7	13.8	36.4	37.2	36.8
GMM+CS+LSTM	8.5	**9.7**	**9.4**	**13.7**	**9.6**	**16.3**	**11.2**	28.3	**28.0**	**28.1**
Enh.+LSTM [239]	5.1	5.7	6.0	8.6	6.7	10.1	7.0	17.0	22.3	19.6

improvement from the LSTM predictions was similar for all room conditions, the employed de-reverberation technique worked better with higher reverberation times. This is due to the fact that CS punishes long-term reverberation energy. Thus, better de-reverberation was observed under long impulse responses. The combination of de-reverberation and LSTM prediction led to further improvements for the SIM dataset. For comparison, Table 5.8 includes results of a recognition system that includes several state-of-the-art techniques to address robustness [239]. In this system, the eight-channel waveforms are processed using sum-and-delay beamforming and LSTM feature enhancement in the front-end part. The back-end part consists of a GMM using feature transformations and discriminative training, in combination with LSTM phoneme predictions (as also employed in this study). This system is able to achieve further improvements in all testing conditions, at the cost of a substantially increased system complexity.

5.4.4 Conclusions

This section presented an LSTM-based system for the recognition of reverberated speech. An LSTM network was employed for phoneme prediction in addition to the GMM acoustic model, which increased the robustness of the system. Experiments were performed according to the official REVERB challenge guidelines with the provided databases. The results showed that the proposed methods are highly effective for the recognition of reverberated speech.

Further improvements are possible with a full integration of all system components. In the current version, speech de-reverberation is not applied on the multi-condition training set, which might bring another small improvement. In addition, the input to the LSTM network is also unenhanced. However, it is not yet fully confirmed in the literature, whether speech enhancement is still relevant for deep neural network based systems; this has to be shown in future work, especially also

for LSTM systems. A detailed comparison of LSTMs (used as an acoustic model in a hybrid system) and similar DNNs without LSTM cells is also to be done in the future.

5.5 Chapter Summary

This chapter addressed the problem of speech recognition in adverse conditions, namely corrupted by highly non-stationary additive noise and by room reverberation. The main part of the evaluated recognition systems consisted of an LSTM RNN, which was deployed in two different configurations. It was shown that network predictions are capable of improving a state-of-the-art GMM acoustic model. Used as an acoustic model alone (for state predictions in the hybrid setup), the LSTM was furthermore able to keep up with the GMM and with other acoustic models based on neural networks. This is especially noticeable since the employed LSTM did not require methods for pre-training or initialisation that are used in other DNN methods.

In a highly reverberant environment, improvements obtained with the LSTM were comparable to a dedicated de-reverberation method, and the combination of both methods resulted in the best system performance.

Due to the enhanced memory of the LSTM topology, this type of RNN can exploit long-range temporal context, which is especially helpful in noisy and reverberant environments. It was shown how much context is analysed by an LSTM, which is larger than the amount of context that is usually used in standard DNNs (through feature frame stacking).

Future work should focus on a better understanding of the performance differences between LSTMs and other network architectures. For example, it is not yet clear, whether methods for pretraining and initialisation (which are applied for standard DNNs) are able to improve LSTM networks. Furthermore, the interaction between speech or feature enhancement and robust acoustic modelling using NNs should be studied in more detail.

6

Summary

The goal of this thesis was to advance the state of the art in different fields of computational auditory scene analysis. The focus was on pattern recognition aspects and thus on the back-end of a system that potentially uses methods for source separation and localisation in its frond-end. This goal was achieved by presenting solutions for different problems, ranging from general acoustic scene classification to robust automatic speech recognition.

Applications of such audio analysis systems can be found in smart homes or in robotic assistants, for example. Such systems have to cope with changing environments and a wide range of different sound sources. On the one hand, this constitutes a difficult challenge for the system. The system is required to adapt to changing external conditions and to previously unseen sound sources. Solutions are desired to address these issues. On the other hand, a system that is utilised in such a rich environment can make use of all kinds of information that is conveyed in the audio signal. Not only human speech, but also other sounds can contain important information. Methods are required to extract the relevant information from different sound sources and to analyse these sounds. Each of the main chapters of this thesis made a contribution towards these goals, addressing the objectives and research questions that were defined in the introduction of this thesis.

In Chapter 2, methods for the **recognition of acoustic scenes and events** were presented. The goal of this system is first to classify an acoustic scene as a whole. Window-based classification is employed to derive a decision for a long-term recording. The suitability of different audio features for this task was investigated, showing that beyond MFCCs, other energy, spectral, and voicing-related features are capable of improving the classification performance. Experiments were performed within the challenge on detection and classification of acoustic scenes and events, where the proposed system was able to beat a baseline derived from speech processing methods and most of the other participating systems, leading to a ranking in the upper third of all participants. In addition, a methodology was proposed to learn new acoustic events in an acoustic event classification system. Here, the diffi-

culty is to create a model for a new class with limited amounts of training data. It was shown that the proposed method (using MAP adaptation of one of the existing models) prevails over conventional learning methods.

A system for **acoustic gait-based person identification** was proposed in Chapter 3. This system analyses the step sounds of persons walking in a corridor. Two different methods for classification were considered. First, static classification was performed building upon a large set of audio features. The contribution of different features to the final result was analysed, and it was shown how a feature selection method can help to choose the optimal set of features. Furthermore, the capability of audio analysis to enhance video processing was demonstrated through multimodal fusion. Second, a dynamic classifier was employed for this task. This system models the distinct gait cycles, which leads to large performance improvements compared to static classification. With this system, two thirds out of 155 subjects (from a publicly available database) were identified correctly when walking in the same condition as for system enrolment.

Different ideas to improve **speaker diarization** were introduced in Chapter 4. The largest portion of this chapter aimed at detecting overlapping speech, where several methods were developed: using the outcome of a signal separation technique, analysing the suitability of different low-level audio features for this task, exploiting lexical information, and using a classifier that analysis long-range temporal context. Each of these methods contributed to improving the overlap detection performance compared to a state-of-the-art system. Furthermore, it was shown how the results of an overlap detection module can be used to improve a speaker diarization system. Nevertheless, the overlap problem is far from being solved. The detection of overlapping speech is a difficult problem, and solutions can contribute to improving all kinds of speech processing applications where multiple speakers are involved. In addition, an algorithm for online speaker diarization was proposed.

The focus of Chapter 5 was on **robust speech recognition**. The goal was to improve recognition in environments with highly non-stationary noise and high reverberation. Therefore, an acoustic model exploiting large amounts of temporal context was applied. This model was either combined with conventional acoustic models or used alone. It was analysed how the proposed architecture performs in comparison to other state-of-the-art methods. Experiments were performed using the databases of the CHiME challenge and the REVERB challenge, demonstrating the improved capability of the proposed system to cope with reverberation and additive noise.

In summary, methods were proposed to address different aspects of an audio recognition system. For different recognition tasks, it was shown how the choice and design of correct set of audio features can improve the system performance. It was also demonstrated across multiple chapters, that methods which exploit temporal context are highly performant in audio recognition tasks.

Directions for future research in the addressed topics were already discussed to some extent in the respective chapters of this thesis. Future systems for the rich transcription of an audio recording (such as a speaker diarization system) should also aim at incorporating other methods from automatic speaker analysis (e. g. gender, emotion, health state [195]), and not only identification. In fact, considering speaker traits or states (e. g. in a multi-task learning system) could help to improve the identification performance in a similar way like handling of overlapping speech. Furthermore, the evaluation of such systems should always be analysed depending on these influencing factors, for example alcohol intoxication [61].

Emerging topics, such as acoustic scene recognition or acoustic gait-based person identification, open possibilities for different applications in the future. On the one hand, these methods can complement computer vision systems in order to improve scene analysis, e. g. in surveillance scenarios. On the other hand, such methods can be deployed on their own, for example where cameras are not available. In the future, further research is required to improve the robustness and performance of these methods.

In this thesis, the focus was only on audio recognition, whereas the separation of multiple simultaneous audio sources was not addressed. In the future, a better coupling of these two tasks (source separation and recognition), for example using exemplar-based approaches [73], could improve the performance in both of these tasks.

A system that combines all of the methods proposed in this thesis is capable of providing a detailed analysis of an acoustic scene. Starting from a coarse classification of the scene as a whole, persons can be identified using their step sounds or voice, followed by a transcription of the spoken contents. Such a system can open a wide range of possible applications.

Acronyms

ALIAS ..... Adaptable Ambient Living Assistant
AMI ..... Augmented Multi-party Interaction
ASR ..... automatic speech recognition
bMMI ..... boosted maximum mutual information
CHiME ..... Computational Hearing in Multi-Source Environments
CNSC ..... convolutive non-negative sparse coding
CS ..... correlation shaping
DER ..... diarization error rate
DL ..... discriminative learning
DNN ..... deep neural network
Err ..... overlap detection error
ESV ..... energy, spectral, and voicing-related
FN ..... false-negative detection
FP ..... false-positive detection
FT ..... feature transformation
GAID ..... gait from audio, image and depth
GEI ..... gait energy image
GMM ..... Gaussian mixture model
HMM ..... hidden Markov model
LDA ..... linear discriminant analysis
LLD ..... low-level descriptor
LLK ..... log-likelihood
LSTM ..... long short-term memory
MAP ..... maximum a posteriori
MCT ..... multi-condition training
MFCC ..... Mel-frequency cepstral coefficient
ML ..... maximum likelihood
MLLR ..... maximum likelihood linear regression
MLLT ..... maximum likelihood linear transform

MMI ..... maximum mutual information
N ..... Negative samples (in a detection problem)
NIST ..... National Institute of Standards and Technology
NMF ..... non-negative matrix factorization
NN ..... neural network
NSC ..... non-negative sparse coding
OIP ..... overlap insertion penalty
P ..... Positive samples (in a detection problem)
Pre ..... precision
Rec ..... recall
REVERB .... reverberant voice enhancement and recognition benchmark
RNN ..... recurrent neural network
SMILE ..... speech and multimedia interpretation by large-space extraction
SNR ..... signal-to-noise ratio
SVM ..... support vector machine
TN ..... true-negative detection
TP ..... true-positive detection
VAD ..... voice activity detection
WSJ ..... Wall Street Journal

Mathematical Symbols

Recognition of Acoustic Scenes and Events

a_{ij} ..... HMM transition probability from state i to state j
A ..... set of HMM transition probabilities a_{ij}
b ..... SVM offset
b_{ik} ..... HMM observation probability for state i and observation k
B ..... set of HMM observation probabilities b_{ik}
c ..... additive constant for Bakis length modelling
$\mathrm{E}_m(x^2)$ ..... expected value of the squared observation vector x^2 for mixture m
f ..... factor for Bakis length modelling
$f(x)$ ..... output of a linear SVM for feature vector x
γ ..... normalising factor for MAP adaptation
k ..... possible HMM class
$\hat{k}$ ..... predicted HMM class
λ ..... HMM parameters
K ..... number of possible HMM observations
μ_i ..... expected value of the feature distribution for class i
μ_m ..... original mean of mixture component m
$\hat{\mu}_m$ ..... MAP-adapted mean of mixture component m
$\bar{\mu}_m$ ..... mean of observation data for mixture component m
n_i ..... number of samples belonging to class i
N ..... number of HMM states
N_m ..... occupation likelihood of adaptation data for mixture component m
N_{tr} ..... number of training recordings

N_w ..... number of windows per recording
O ..... HMM observation sequence
O_t ..... HMM observation at time step t
π_n ..... starting probability for HMM state n
Π ..... set of HMM starting probabilities π_n
q_t ..... HMM state at time step t
σ_i ..... standard deviation of the feature distribution for class i
σ_m^2 ..... original variance of mixture component m
$\hat{\sigma}_m^2$ ..... MAP-adapted variance of mixture component m
s_n ..... HMM state
S ..... set of possible HMM states s_n
τ ..... MAP adaptation weighting factor
t_{ij} ..... t-statistic for two different classes i and j
$\bar{t}(k)$ ..... average length of recordings for class k
T ..... number of time steps in an observation
v_k ..... possible HMM observation
V ..... set of possible HMM observations v_m
w ..... SVM normal vector to the hyperplane
w_m ..... original mixture weight of mixture component m
$\hat{w}_m$ ..... MAP-adapted mixture weight
x ..... feature vector

Acoustic Gait-based Person Identification

α ..... significance level
$a(f_n)$ ..... recognition accuracy for feature f_n
c ..... class index
$\delta(\cdot)$ ..... Kronecker delta function
f_n ..... audio feature with index n
F ..... set of audio features
$l(t)$ ..... true label for instance t
$m(c, f_n)$ ..... model for class c using feature f_n
N ..... number of features
T ..... number of test instances
x_t ..... feature vector for instance t
X ..... set of feature vectors x_t

Speaker Diarization

c_w . . . . . word class (speech or overlap)
δ_{ER} . . . . . threshold for the CNSC energy ratio ER
δ_{ET} . . . . . threshold for the CNSC total energy ET
d_f . . . . . discriminant value of feature f
$\mathrm{dur}(seg)$ . . . . . duration of a segment
$D(\cdot \mid\mid \cdot)$ . . . . . Kullback-Leibler divergence
$E_t(s)$ . . . . . estimated energy for speaker s in frame t
ER_t . . . . . CNSC energy ratio for frame t
ET_{nt} . . . . . normalised total CNSC energy for frame t
ET_t . . . . . total CNSC energy for frame t
θ . . . . . threshold for LSTM overlap detection
f . . . . . audio feature index
f_0 . . . . . fundamental frequency
f_{ij} . . . . . audio frames belonging to speaker i and classified as speaker j
$\mathcal{F}$ . . . . . LSTM output function
H . . . . . NSC activation matrix
$\hat{H}$ . . . . . objective for NSC activation matrix
I . . . . . identity matrix
I_s . . . . . speaker-specific rows in NSC activation matrix H for speaker s
λ . . . . . NSC sparsity parameter
l . . . . . length of an overlap segment
μ_p . . . . . mean of Gaussian distribution p
M . . . . . dimensionality of NSC data matrix (spectral components)
N . . . . . dimensionality of NSC data matrix (time frames)
$N_{corr}(seg)$ . . . number of correctly detected speakers for segment seg
$N_{ref}(seg)$ . . . . number of speakers for segment seg in the reference
$N_{sys}(seg)$ . . . . number of hypothesised speakers for segment seg
$o(t)$ . . . . . LSTM output at time step t
$\hat{o}(t)$ . . . . . LSTM target at time step t
p . . . . . column shift index for CNSC
p_f . . . . . probability distribution of feature f for overlap frames
P . . . . . convolutional range in CNSC
P_c . . . . . diarization cluster purity

q_f ..... probability distribution of feature f for all frames
r ..... regularisation factor for CNSC total energy
R ..... number of NSC bases
σ_p^2 ..... variance of Gaussian distribution p
s ..... speaker index
s_{lex} ..... overlap score derived from lexical cues
$\hat{s}_i$ ..... speaker with the i^{th} highest energy
seg ..... speech segment
S ..... number of speakers
$\hat{S}$ ..... number of speakers hypothesised by the system
τ ..... MAP adaptation weighting factor
t ..... time step index
t_{ET} ..... factor for threshold δ_{ET}
T_{sp} ..... set of all speech frames in a recording
w ..... discrete random variable for the word
W ..... NSC basis matrix
$\hat{W}$ ..... objective for NSC basis matrix
W^G ..... global CNSC speaker basis
W_p ..... basis matrix for CNSC for column shift of p
W_s ..... CNSC basis for speaker s
x_{ft} ..... feature vector at time step t
X ..... data matrix for NSC
$\hat{X}$ ..... NSC recomposition for data matrix
X_f ..... whole set of feature vectors
X_{ft} ..... set of feature vectors up to time step t

Robust Speech Recognition

$b_t^{(j)}$ ..... LSTM output element j at time step t after softmax
$\hat{b}_t$ ..... discrete LSTM phoneme prediction at time step t
b_x ..... bias vector for variable x
c_{sm} ..... weight of the m-th Gaussian mixture component for state s
c_t ..... LSTM cell variable at time step t
C ..... LSTM phoneme confusion matrix
$\frac{\partial y}{\partial x}$ ..... derivative of y with respect to x

f_t ..... LSTM forget gate activation vector at time step t
h ..... hidden state vector sequence of a neural network
$\overleftarrow{h_t^N}$ ..... backward state vector of the N-th hidden layer
$\overrightarrow{h_t^N}$ ..... forward state vector of the N-th hidden layer
$\mathcal{H}$ ..... neural network hidden layer activation function
i_t ..... LSTM input gate activation vector at time step t
$J_{ji}^{tt'}$ ..... sequential Jacobian
λ ..... GMM stream weight for HMM double-stream decoding
$\mathcal{L}$ ..... LSTM prediction function
μ_{sm} ..... mean vector of the m-th Gaussian mixture component for state s
$m(s)$ ..... mapping function from HMM states s to phonemes
M ..... number of mixture components in a GMM
N ..... number of hidden layers in a deep neural network
$\mathcal{N}(\cdot; \mu, \Sigma)$ .... multivariate Gaussian distribution with mean vector μ and covariance matrix Σ
o_t ..... LSTM output gate activation vector at time step t
p_G ..... GMM acoustic likelihood
p_L ..... LSTM acoustic likelihood
P ..... number of LSTM output units
σ ..... sigmoid function
Σ_{sm} ..... covariance matrix of the m-th Gaussian mixture component for state s
s_t ..... HMM state at time step t
$S_{t'}$ ..... average network output sensitivity to input at relative position t'
t ..... time frame/step index
T ..... sequence length
W_{xy} ..... weight matrix from x to y
x_t ..... input feature vector at time step t
y_t ..... neural network output at time step t

List of Figures

List of Tables

References

[1] M. Adda-Decker, C. Barras, G. Adda, P. Paroubek, P. B. de Mareüil, and B. Habert, "Annotation and analysis of overlapping speech in political interviews," in *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008, ELRA, pp. 3105–3111.

[2] D. T. Alpert and M. Allen, "Acoustic gait recognition on a staircase," in *Proc. World Automation Congr. (WAC)*, Kobe, Japan, 2010, IEEE.

[3] M. U. B. Altaf, T. Butko, and B.-H. Juang, "Person identification using biometric markers from footstep sounds," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013, ISCA, pp. 2934–2938.

[4] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Philadelphia, PA, USA, 1996, IEEE, pp. 1137–1140.

[5] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[6] H. Aronowitz, Y. A. Solewicz, and O. Toledo-Ronen, "Online two speaker diarization," in *Proc. Odyssey Speaker and Language Recognition Workshop*, Singapore, 2012, ISCA.

[7] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[8] J. H. Bach and J. Anemüller, "Detecting novel objects in acoustic scenes through classifier incongruence," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, ISCA, pp. 2206–2209.

[9] R. Bakis, "Continuous speech recognition via centisecond acoustic states," *The Journal of the Acoustical Society of America*, vol. 59, no. S1, pp. S97, 1976.

[10] A. Bannat, J. Blume, J. Geiger, T. Rehrl, F. Wallhoff, C. Mayer, B. Radig, S. Sosnowski, and K. Kühnlenz, "A multimodal human-robot-dialog applying emotional feedbacks," in *Proc. Int. Conf. on Social Robotics (ICSR)*, Singapore, 2010, pp. 1–10.

[11] J. P. Barker, E. Vincent, N. Ma, H. Christensen, and P. D. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, Elsevier, 2013.

[12] A. Belouchrani and M. G. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, 1998.

[13] J. A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 89–100, 2005.

[14] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[15] R. E. Bland, "Acoustic and seismic signal processing for footstep detection," M.S. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, 2006.

[16] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped Speech Detection for Improved Diarization in Multi-Party Meetings," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, IEEE, pp. 4353–4356.

[17] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: improving speaker diarization by automatically identifying and excluding overlapped speech," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Brisbane, Australia, 2008, ISCA, pp. 32–35.

[18] K. Boakye, O. Vinyals, and G. Friedland, "Improved Overlapped Speech Handling for Speaker Diarization," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Florence, Italy, 2011, ISCA, pp. 941–944.

[19] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*, Kluwer Academic Publishers, 1994.

[20] S. Bozonnet, N. Evans, and C. Fredouille, "The LIA-Eurecom RT09 speaker diarization system: Enhancements in speaker modelling and cluster purification," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, 2010, IEEE, pp. 4958–4961.

[21] S. Bozonnet, R. Vipperla, and N. Evans, "Phone Adaptive Training for Speaker Diarization," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012, ISCA, pp. 494–497.

[22] S. Bozonnet, D. Wang, N. Evans, and R. Troncy, "Linguistic influences on bottom-up and top-down clustering for speaker diarization," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, IEEE, pp. 4424–4427.

[23] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1994.

[24] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method," *The Journal of the Acoustical Society of America*, vol. 92, no. 3, pp. 1394–1402, 1992.

[25] F. Brugnara, D. Falavigna, D. Giuliani, and R. Gretter, "Analysis of the characteristics of talk-show TV programs," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012, ISCA, pp. 1388–1391.

[26] J. P. Campbell Jr, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[27] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," *Machine Learning for Multimodal Interaction*, pp. 28–39, Springer, 2006.

[28] B. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Jeju Island, Korea, 2004, ISCA, pp. 612–615.

[29] I. Chen, S.-S. Cheng, and H.-M. Wang, "Phonetic subspace mixture model for speaker diarization," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, ISCA, pp. 2298–2301.

[30] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*. Lansdowne, VA, USA, 1998.

[31] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[32] M. Chum, A. Habshush, A. Rahman, and C. Sang, "IEEE AASP scene classification challenge using hidden Markov models and frame based classification," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, Springer, 1995.

[34] D. Cunado, M. S. Nixon, and J. N. Carter, "Using gait as a biometric, via phase-weighted magnitude spectra," in *Proc. Int. Conf. Audio-and Video-based Biometric Person Authentication (AVBPA)*, Crans-Montana, Switzerland, 1997, Springer, pp. 93–102.

[35] J. E. Cutting and L. T. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues," *Bulletin of the psychonomic society*, vol. 9, no. 5, pp. 353–356, Springer, 1977.

[36] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, 2005, IEEE, pp. 886–893.

[38] R. de Carvalho and P. Rosa, "Identification system for smart homes using footstep sounds," in *Proc. Int. Symp. on Industrial Electronics*, Bari, Italy, 2010, IEEE, pp. 1639–1644.

[39] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *Journal of New Music Research*, vol. 30, no. 1, pp. 39–58, Taylor & Francis, 2001.

[40] G. R. Doddington, "Speaker recognition - identifying people by their voices," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[41] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, John Wiley & Sons, 2012.

[42] B. Elizalde, H. Lei, G. Friedland, and N. Peters, "An i-vector based approach for audio scene detection," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[43] N. Evans, S. Bozonnet, D. Wang, C. Fredouille, and R. Troncy, "A comparative study of bottom-up and top-down approaches to speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 382–392, 2012.

[44] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 483–487.

[45] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (ACM-MM)*, Florence, Italy, 2010, pp. 1459–1462.

[46] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, "The Sheffield wargames corpus," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013, ISCA, pp. 1116–1120.

[47] C. Fredouille and N. Evans, "The influence of speech activity detection and overlap on the speaker diarization for meeting room recordings," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Antwerp, Belgium, 2007, ISCA, pp. 2953–2956.

[48] C. Fredouille, S. Bozonnet, and N. Evans, "The LIA-EURECOM RT '09 speaker diarization system," in *NIST Rich Transcription Evaluation Workshop*, Melbourne, Florida, USA, 2009.

[49] G. Friedland, "Using a GPU, online diarization= offline diarization," Tech. Rep., International Computer Science Institute, Berkeley, CA, USA, 2012.

[50] G. Friedland, O. Vinyals, Y. Huang, and C. Müller, "Prosodic and other long-term features for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 985–993, 2009.

[51] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, Elsevier, 1998.

[52] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Brighton, UK, 2009, ISCA, pp. 2583–2586.

[53] J. Garofolo, J. G. Fiscus, and W. M. Fisher, "Design and preparation of the 1996 Hub-4 broadcast news benchmark test corpora," in *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, USA, 1997, pp. 15–21.

[54] J. Geiger, T. Leykauf, T. Rehrl, F. Wallhoff, and G. Rigoll, "The robot ALIAS as a gaming platform for elderly persons," in *Proc. German Ambient Assisted Living Congr. (AAL)*, Berlin, Germany, 2013, VDE Verlag, pp. 380–386.

[55] J. Geiger, J. Schenk, F. Wallhoff, and G. Rigoll, "Optimizing the number of states for HMM-based on-line handwritten whiteboard recognition," in *Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)*, Kolkata, India, 2010, IEEE, pp. 107–112.

[56] J. Geiger, I. Yenin, T. Rehrl, F. Wallhoff, and G. Rigoll, "Display of emotions with the robotic platform ALIAS," in *Proc. German Ambient Assisted Living Congr. (AAL)*, Berlin, Germany, 2013, VDE Verlag, pp. 248–253.

[57] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013, IEEE.

[58] J. T. Geiger, B. Schuller, and G. Rigoll, "Recognising acoustic scenes with large-scale audio feature extraction and SVM," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[59] J. T. Geiger, F. Weninger, J. F. Gemmeke, M. Wöllmer, B. Schuller, and G. Rigoll, "Memory-enhanced neural networks and NMF for robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1037–1046, 2014.

[60] J. T. Geiger, F. Weninger, A. Hurmalainen, J. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM+TUT+KUL approach to the 2nd CHiME challenge: Multi-stream ASR exploiting BLSTM networks and sparse NMF," in *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, Vancouver, Canada, 2013, pp. 25–30.

[61] J. T. Geiger, B. Zhang, B. Schuller, and G. Rigoll, "On the influence of alcohol intoxication on speaker recognition," in *Proc. AES Int. Conf. on Semantic Audio*, London, UK, 2014, pp. 1–7.

[62] J. Geiger, F. Wallhoff, and G. Rigoll, "GMM-UBM based open-set online speaker diarization," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, ISCA, pp. 2330–2333.

[63] J. T. Geiger, F. Eyben, N. Evans, B. Schuller, and G. Rigoll, "Using Linguistic Information to Detect Overlapping Speech," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013, ISCA, pp. 690–694.

[64] J. T. Geiger, F. Eyben, B. Schuller, and G. Rigoll, "Detecting Overlapping Speech with Long Short-Term Memory Recurrent Neural Networks," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013, ISCA, pp. 1668–1672.

[65] J. T. Geiger, J. F. Gemmeke, B. Schuller, and G. Rigoll, "Investigating NMF speech enhancement for neural network based acoustic models," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*. 2014, ISCA.

[66] J. T. Geiger, M. Hofmann, B. Schuller, and G. Rigoll, "Gait-based person identification by spectral, cepstral and energy-related audio features," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 458–462.

[67] J. T. Geiger, M. Kneissl, B. Schuller, and G. Rigoll, "Acoustic gait-based person identification using hidden Markov models," in *Proc. Personality Mapping Challenge & Workshop (MAPTRAITS), Satellite of the Int. Conf. on Multimodal Interaction (ICMI)*, Istanbul, Turkey, 2014, pp. 25–30, ACM.

[68] J. T. Geiger, R. Vipperla, N. Evans, B. Schuller, and G. Rigoll, "Speech Overlap Detection and Attribution Using Convolutive Non-Negative Sparse Coding: New Improvements and Insights," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Bucharest, Romania, 2012, EURASIP, pp. 340–344.

[69] J. T. Geiger, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*. 2014, ISCA.

[70] J. T. Geiger, M. A. Lakhal, B. Schuller, and G. Rigoll, "Learning new acoustic events in an HMM-based system using MAP adaptation," in *Proc. Annu.*

Conf. of the Int. Speech Communication Association (INTERSPEECH), Florence, Italy, 2011, ISCA, pp. 293–296.

[71] J. T. Geiger, R. Vipperla, S. Bozonnet, N. Evans, B. Schuller, and G. Rigoll, "Convolutive Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012, ISCA, pp. 2154–2157.

[72] J. T. Geiger, E. Marchi, B. W. Schuller, and G. Rigoll, "The TUM system for the REVERB challenge: Recognition of reverberated speech using multi-channel correlation shaping dereverberation and BLSTM recurrent neural networks," in *REVERB Workshop*, Florence, Italy, 2014, 8 pages.

[73] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[74] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, MIT Press, 2000.

[75] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley, "IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events," Tech. Rep., Electrical Engineering department, Queen Mary University of London, London, UK, 2013.

[76] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013, IEEE.

[77] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Marrakech, Morocco, 2013.

[78] B. W. Gillespie and A. E. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, 2003, IEEE, pp. 673–676.

[79] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, Elsevier, 1995.

[80] D. Graff, "An overview of broadcast news corpora," *Speech Communication*, vol. 37, no. 1-2, pp. 15–26, Elsevier, 2002.

[81] A. Gravano, S. Benus, J. Hirschberg, S. Mitchell, and I. Vovsha, "Classification of discourse functions of affirmative words in spoken dialogue," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Antwerp, Belgium, 2007, ISCA, pp. 1613–1616.

[82] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, no. 3, pp. 601–634, Elsevier, 2011.

[83] A. Gravano and J. Hirschberg, "A corpus-based study of interruptions in spoken dialogue," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012, ISCA, pp. 855–858.

[84] A. Graves, *Supervised sequence labelling with recurrent neural networks*, Ph.D. thesis, Department of Informatics, Technische Universität München, Munich, Germany, 2008.

[85] A. Graves, S. Fernández, and J. Schmidhuber, "Multidimensional recurrent neural networks," in *Proc. Int. Conf. on Artificial Neural Networks (ICANN)*, Porto, Portugal, 2007, Springer.

[86] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, Elsevier, 2005.

[87] A. Graves, N. Jaitly, and A.-R. Mohamed, "Speech recognition with deep recurrent neural networks," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Olomouc, Czech Republic, 2013, IEEE, pp. 273–278.

[88] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[89] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 6645–6649.

[90] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *Int. Conf. on Language Resources, Evaluation and Corpora (LREC)*, Istanbul, Turkey, 2012, ELRA, pp. 114–118.

[91] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottleneck features for LVCSR of meetings," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, HI, USA, 2007, IEEE, pp. 754–757.

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

[93] A. Hagen and A. Morris, "Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR," *Computer Speech and Language*, vol. 19, no. 1, pp. 3–30, Elsevier, 2005.

[94] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The Weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[95] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[96] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, Elsevier, 2010.

[97] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, San Francisco, CA, USA, 1992, IEEE, pp. 121–124.

[98] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000, IEEE, pp. 1635–1638.

[99] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[100] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, MIT Press, 1997.

[101] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A Field*

Guide to Dynamical Recurrent Networks, S. C. Kremer and J. F. Kolen, Eds. IEEE Press, 2001.

[102] M. Hofmann, S. Bachmann, and G. Rigoll, "2.5D gait biometrics using the depth gradient histogram energy image," in *Proc. Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, Washington, DC, USA, 2012, IEEE, pp. 399–403.

[103] M. Hofmann, S. Schmidt, A. Rajagopalan, and G. Rigoll, "Combined face and gait recognition using alpha matte preprocessing," in *International Conference on Biometrics*, New Delhi, India, 2012, IAPR, pp. 390–395.

[104] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The TUM Gait from Audio, Image and Depth (GAID) Database: Multimodal Recognition of Subjects and Traits," *Journal of Visual Communication and Image Representation (JVCI), Special Issue on Visual Understanding and Applications with RGB-D Cameras*, vol. 25, no. 1, pp. 195–206, Elsevier, 2014.

[105] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, MIT Press, 2004.

[106] Q. Huang and S. Cox, "Using high-level information to detect key audio events in a Tennis game," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, ISCA, pp. 1409–1412.

[107] M. Huijbregts, D. van Leeuwen, and F. D. Jong, "Speech overlap detection in a two-pass speaker diarization system," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Brighton, UK, 2009, ISCA, pp. 1063–1066.

[108] M. Huijbregts, D. van Leeuwen, and C. Wooters, "Speaker diarization error analysis using oracle components," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 393–403, 2012.

[109] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Antwerp, Belgium, 2007, ISCA, pp. 1857–1860.

[110] M. Huijbregts, *Segmentation, diarization and speech transcription: surprise data unraveled*, Ph.D. thesis, Centre for Telematics and Information Technology, University of Twente, Twente, The Netherlands, 2008.

[111] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Compact long context spectral factorisation models for noise robust recognition of medium vocabulary speech," in *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, Vancouver, Canada, 2013, pp. 13–18.

[112] A. Itai and H. Yasukawa, "Footstep recognition with psyco-acoustics parameter," in *Proc. Asia Pacific Conf. on Circuits and Systems (APCCAS)*, Singapore, 2006, IEEE, pp. 992–995.

[113] A. Itai and H. Yasukawa, "Footstep classification using simple speech recognition technique," in *Proc. Int. Symp. on Circuits and Systems (ISCAS)*, Seattle, WA, USA, 2008, IEEE, pp. 3234–3237.

[114] A. K. Jain, R. M. Bolle, and S. Pankanti, *Biometrics: personal identification in networked society*, Springer, 1999.

[115] F. Jelinek, *Statistical methods for speech recognition*, MIT press, 1997.

[116] H. Jiang, J. Bai, S. Zhang, and B. Xu, "SVM-based audio scene classification," in *Proc. Int. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Wuhan, China, 2005, IEEE, pp. 131–136.

[117] H. Jiang, "Discriminative training of hmms for automatic speech recognition: A survey," *Computer Speech & Language*, vol. 24, no. 4, pp. 589–608, Elsevier, 2010.

[118] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, "Lexical, prosodic, and syntactic cues for dialog acts," in *Proc. ACL/COLING Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada, 1998, pp. 114–120.

[119] K. Kalgaonkar and B. Raj, "Acoustic doppler sonar for gait recogination," in *Proc. of Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, London, UK, 2007, IEEE, pp. 27–32.

[120] O. Kalinli, S. Sundaram, and S. Narayanan, "Saliency-driven unstructured acoustic scene classification using latent perceptual indexing," in *Proc. Int. Workshop on Multimedia Signal Processing (MMSP)*, Rio de Janeiro, Brazil, 2009, IEEE.

[121] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[122] D. Y. Kim, C. Kwan Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Communication*, vol. 24, no. 1, pp. 39–49, Elsevier, 1998.

[123] S. Kim, S. H. Yella, and F. Valente, "Automatic detection of conflict escalation in spoken conversations," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012, ISCA, pp. 1167–1170.

[124] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, Elsevier, 2010.

[125] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013, IEEE.

[126] A. Klapuri and M. Davy, Eds., *Signal processing methods for music transcription*, Springer, 2006.

[127] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, Elsevier, 1997.

[128] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel, "Temporal ica for classification of acoustic events in a kitchen environment," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Lisbon, Portugal, 2005, ISCA, pp. 2689–2692.

[129] J. D. Krijnders and G. A. ten Holt, "A tone-fit feature representation for scene classification," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[130] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1692–1707, 2010.

[131] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, C. Lévy, H. Li, J. Mason, and J.-Y. Parfait, "ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013, ISCA, pp. 2768–2772.

[132] K. Laskowski, M. Heldner, and J. Edlund, "On the Dynamics of Overlap in Multi-Party Conversation," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012.

[133] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Information Processing Systems (NIPS)*, Denver, CO, USA, 2000, MIT Press, pp. 556–562.

[134] L. Lee and W. E. L. Grimson, "Gait analysis for recognition and classification," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG)*, Washington, D.C., USA, 2002, IEEE, pp. 148–155.

[135] C. J. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, Elsevier, 1995.

[136] S. C. Levinson, *Pragmatics*, Cambridge University Press, 1983.

[137] D. Li, J. Tam, and D. Toub, "Auditory scene classification using machine learning techniques," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[138] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[139] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, San Juan, PR, USA, 2005, IEEE, pp. 357–362.

[140] R. Lippmann, E. Martin, and D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, 1987, IEEE, pp. 705–708.

[141] D. Liu and F. Kubala, "Online speaker clustering," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, 2003, IEEE, pp. 572–575.

[142] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 20, no. 1-2, pp. 61–79, Springer, 1998.

[143] K. Mäkelä, J. Hakulinen, and M. Turunen, "The use of walking sounds in supporting awareness," in *Proc. of Int. Conf. on Auditory Displays (ICAD)*, Boston, MA, USA, 2003, pp. 144–147.

[144] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1338–1353, 2000.

[145] K. Markov and S. Nakamura, "Never-ending learning system for on-line speaker diarization," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Kyoto, Japan, 2007, IEEE, pp. 699–704.

[146] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[147] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *Proc. A Speaker Odyssey - The Speaker Recognition Workshop*, Crete, Greece, 2001, ISCA, pp. 175–180.

[148] N. Miura, A. Nagasaka, and T. Miyatake, "Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification," *Machine Vision and Applications*, vol. 15, no. 4, pp. 194–203, Springer, 2004.

[149] M. H. Moattar and M. M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Communication*, vol. 54, no. 10, pp. 1065–1103, Elsevier, 2012.

[150] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[151] M. P. Murray, "Gait as a total pattern of movement: Including a bibliography on gait," *American Journal of Physical Medicine & Rehabilitation*, vol. 46, no. 1, pp. 290–333, 1967.

[152] J. Nam, Z. Hyung, and K. Lee, "Acoustic scene classification using sparse feature learning and selective max-pooling by event detection," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[153] F. Nesta, M. Matassoni, and R. F. Astudillo, "A flexible spatial blind source extraction framework for robust speech recognition in noisy environments," in *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, Vancouver, Canada, 2013, pp. 33–38.

[154] T. Nguyen, H. Sun, S. Zhao, S. Khine, H. Tran, T. Ma, B. Ma, E. Chng, and H. Li, "The IIR-NTU speaker diarization systems for RT 2009," in *NIST Rich Transcription Evaluation Workshop*, Melbourne, Florida, USA, 2009.

[155] NIST consortium, "The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan," http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, 2009, Online; accessed March 2014.

[156] M. Nixon, T. Tan, and R. Chellappa, *Human identification based on gait*, Springer, 2006.

[157] W. Nogueira, G. Roma, and P. Herrera, "Sound scene identification based on MFCC, binaural features and a support vector machine classifier," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[158] H. Okuno, T. Ogata, K. Komatani, and K. Nakadai, "Computational auditory scene analysis and its application to robot audition," in *Proc. Int. Conf. on Informatics Research for Development of Knowledge Society Infrastructure (ISKCS)*, Washington, DC, USA, 2004, IEEE, pp. 73–80.

[159] E. Olivetti, "The wonders of the normalized compression dissimilarity representation," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[160] M. Otero, "Application of a continuous wave radar for human gait recognition," in *Proc. SPIE Conf. on Signal Processing, Sensor Fusion, and Target Recognition XIV*, Orlando, FL, USA, 2005, International Society for Optics and Photonics, pp. 538–548.

[161] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Kyoto, Japan, 2007, IEEE, pp. 683–686.

[162] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time differences," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Pittsburgh, PA, USA, 2006, ISCA, pp. 2194–2197.

[163] K.-M. Park, J.-S. Park, J.-H. Bae, and Y.-H. Oh, "Online speaker diarization for multimedia data retrieval on mobile devices," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 08, World Scientific, 2012.

[164] K. Patil and M. Elhilali, "Multiresolution auditory representations for scene classification," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[165] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop on Speech and Natural Language (HLT)*, Harriman, NY, USA, 1992, pp. 357–362.

[166] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL, USA, 2002, IEEE, pp. 1938–1941.

[167] W. D. Penny and S. J. Roberts, "Variational bayes for 1-dimensional mixture models," Tech. Rep., Department of Engineering Science, Oxford University, Oxford, UK, 2000.

[168] T. Pfau, D. P. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Madonna di Campiglio, Italy, 2001, IEEE, pp. 107–110.

[169] C. Plahl, M. Kozielski, R. Schlüter, and H. Ney, "Feature combination and stacking of recurrent and non-recurrent neural networks for LVCSR," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 6714–6718.

[170] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods*, B. Schölkopf, A. J. Smola, and R. Soentpiet, Eds., pp. 185–208. MIT press, 1999.

[171] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Honolulu, HI, USA, 2011, IEEE.

[172] D. Povey and K. Yao, "A basis method for robust estimation of constrained MLLR," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, IEEE, pp. 4460–4463.

[173] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, IEEE, pp. 4057–4060.

[174] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, Elsevier, 1994.

[175] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[176] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[177] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[178] D. R. Reddy, "Speech recognition by machine: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 501–531, 1976.

[179] T. Rehrl, J. Blume, J. Geiger, A. Bannat, F. Wallhoff, S. Ihsen, Y. Jeanrenaud, M. Merten, B. Schönebeck, S. Glende, and C. Nedopil, "ALIAS: Der anpassungsfähige Ambient Living Assistent," in *Proc. German Ambient Assisted Living Congr. (AAL)*, Berlin, Germany, 2011, VDE Verlag.

[180] W. Reichl and G. Ruske, "Discriminative training for continuous speech recognition," in *Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)*, Madrid, Spain, 1995, ISCA, pp. 537–540.

[181] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, IEEE, pp. 1833–1836.

[182] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, Elsevier, 2000.

[183] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, 2005, IEEE, pp. 950–953.

[184] G. Rigoll, "Maximum mutual information neural networks for hybrid connectionist-HMM speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 175–184, 1994.

[185] G. Rigoll and D. Willett, "A NN/HMM hybrid for continuous speech recognition with a discriminant nonlinear feature extraction," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, WA, USA, 1998, IEEE, pp. 9–12.

[186] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, MI, USA, 1995, IEEE, pp. 81–84.

[187] R. V. Rodríguez, R. P. Lewis, J. S. D. Mason, and N. W. D. Evans, "Footstep recognition for a smart home environment," *International Journal of Smart Home*, vol. 2, no. 2, pp. 95–110, SERSC, 2008.

[188] G. Roma, W. Nogueira, and P. Herrera, "Recurrence quantification analysis features for auditory scene classification," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[189] J. Rottland and G. Rigoll, "Tied posteriors: An approach for effective introduction of context dependency in hybrid NN/HMM LVCSR," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Istandbul, Turky, 2000, IEEE, pp. 1241–1244.

[190] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013, ISCA, pp. 1477–1481.

[191] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*. 2014, ISCA.

[192] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000, IEEE, pp. 1129–1132.

[193] B. Schuller, F. Pokorny, S. Ladstätter, M. Fellner, F. Graf, and L. Paletta, "Acoustic geo-sensing: Recognising cyclists' route, route direction, and route progress from cell-phone audio," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 453–457.

[194] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, 2004, IEEE, pp. 574–577.

[195] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "Introduction to the special issue on broadening the view on speaker analysis," *Computer Speech and Language, Special Issue on Broadening the View on Speaker Analysis*, vol. Vol. 28, pp. 343–345, Elsevier, 2014.

[196] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011-the first international audio/visual emotion challenge," in *Int.*

Audio/Visual Emotion Challenge and Workshop, Memphis, TN, USA, 2011, SSPNET, pp. 415–424.

[197] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language - state-of-the-art and the challenge," *Computer Speech and Language*, vol. 27, no. 1, pp. 4–39, Elsevier, 2013.

[198] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[199] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[200] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 7398–7402.

[201] B. She, "Framework of footstep detection in in-door environment," in *Proc. Int. Congress on Acoustics (ICA)*, Kyoto, Japan, 2004, pp. 715–718.

[202] Y. Shoji, T. Takasuka, and H. Yasukawa, "Personal identification using footstep detection," in *Proc. Int. Symp. on Intelligent Signal Processing and Communication Systems (ISPACS)*, Seoul, South Korea, 2004, IEEE, pp. 43–47.

[203] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Lisbon, Portugal, 2005, ISCA, pp. 1781–1784.

[204] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)*, Aalborg, Denmark, 2001, ISCA, pp. 1359–1362.

[205] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, USA, 1997, pp. 15–21.

[206] M. Sinclair and S. King, "Where are the challenges in speaker diarization?," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 7741–7745.

[207] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarisation system," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Lisbon, Portugal, 2005, ISCA, pp. 2437–2440.

[208] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[209] J. Stadermann and G. Rigoll, "Comparing NN paradigms in hybrid NN/HMM speech recognition using tied posteriors," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, St. Thomas, U.S. Virgin Islands, USA, 2003, IEEE, pp. 89–93.

[210] H. Sun and B. Ma, "Study of overlapped speech detection for NIST SRE summed channel speaker recognition," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Florence, Italy, 2011, ISCA, pp. 2345–2348.

[211] Y. Sun, J. F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves, "Using a DBN to integrate sparse classification and GMM-based ASR," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, ISCA, pp. 2098–2101.

[212] J. Suutala and J. Roning, "Combining classifiers with different footstep feature sets and multiple samples for person identification," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, 2005, IEEE, pp. 354–357.

[213] Y. Tachioka, S. Watanabe, and J. R. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 6935–6939.

[214] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," in *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, Vancouver, Canada, 2013, pp. 19–24.

[215] A. Temko and C. Nadeu, "Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, 2005, IEEE, pp. 502–505.

[216] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, "Acoustic event detection and classification," in *Computers in the Human Interaction Loop*, A. Waibel and R. Stiefelhagen, Eds., pp. 61–73. Springer, 2009.

[217] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," *Multimodal Technologies for Perception of Humans*, pp. 311–322, Springer, 2007.

[218] L. ten Bosch, N. Oostdijk, and L. Boves, "On temporal aspects of turn taking in conversational dialogues," *Speech Communication*, vol. 47, no. 1, pp. 80–86, Elsevier, 2005.

[219] S. E. Tranter and D. A. Reynolds, "Speaker diarisation for broadcast news," in *Proc. Odyssey Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, ISCA, pp. 337–344.

[220] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[221] E. Trentin and M. Gori, "A survey of hybrid ANN/HMM models for automatic speech recognition," *Neurocomputing*, vol. 37, no. 1, pp. 91–126, Elsevier, 2001.

[222] A. Tritschler and R. A. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in *Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)*, Budapest, Hungary, 1999, ISCA, pp. 679–682.

[223] C. Vaquero, O. Vinyals, and G. Friedland, "A hybrid approach to online speaker diarization," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, ISCA, pp. 2638–2641.

[224] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 126–130.

[225] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, IEEE, pp. 4085–4088.

[226] R. Vipperla, J. T. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection and attribution using convolutive non-negative sparse coding," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, IEEE, pp. 4181–4184.

[227] T. Virtanen, R. Singh, and B. Raj, *Techniques for noise robustness in automatic speech recognition*, Wiley, 2012.

[228] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley, 2006.

[229] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 12, pp. 1505–1518, 2003.

[230] W. Wang, "Convolutive non-negative sparse coding," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, Hong Kong, China, 2008, IEEE, pp. 3681–3684.

[231] B. L. Welch, "The generalization of Student's problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, Oxford Journals, 1947.

[232] C. Weng, D. Yu, S. Watanabe, and B.-H. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, IEEE, pp. 5569–5573.

[233] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich feature enhancement approach to the 2013 CHiME challenge using BLSTM recurrent neural networks," in *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, Vancouver, Canada, 2013, pp. 86–90.

[234] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments," *Computer Speech and Language*, vol. 28, no. 4, pp. 888–902, Elsevier, 2014.

[235] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments," in *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, Florence, Italy, 2011, pp. 24–29.

[236] F. Weninger and B. Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, IEEE, pp. 337–340.

[237] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative matrix factorization for highly noise-robust ASR: to enhance or to recognize?," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, IEEE, pp. 4681–4684.

[238] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT – the Munich Open-Source CUDA RecurREnt Neural Network Toolkit," *Journal of Machine Learning Research*, vol. 15, 2014, 5 pages, to appear.

[239] F. Weninger, S. Watanabe, J. Le Roux, J. R. Hershey, Y. Tachioka, J. T. Geiger, B. W. Schuller, and G. Rigoll, "The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement," in *REVERB Workshop*, Florence, Italy, 2014, 8 pages.

[240] M. Włodarczak, J. Simko, and P. Wagner, "Temporal entrainment in overlapped speech: Cross-linguistic study," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012, pp. 615–618.

[241] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "A multi-stream ASR framework for BLSTM modeling of conversational speech," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, IEEE, pp. 4860–4863.

[242] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of BLSTM neural networks for enhanced emotion classification in dyadic spoken interactions," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, IEEE, pp. 4157–4160.

[243] M. Wöllmer, F. Weninger, J. Geiger, B. Schuller, and G. Rigoll, "Noise robust ASR in reverberated multisource environments applying convolutive NMF and long short-term memory," *Computer Speech and Language, Special Issue on Speech Separation and Recognition in Multisource Environments*, vol. 27, no. 3, pp. 780–797, 2013.

[244] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Brisbane, Australia, 2008, ISCA, pp. 597–600.

[245] M. Wöllmer, F. Eyben, B. Schuller, Y. Sun, T. Moosmayr, and N. Nguyen-Thien, "Robust in-car spelling recognition-a tandem BLSTM-HMM approach," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Brighton, UK, 2009, ISCA, pp. 2507–2510.

[246] M. Wöllmer, B. Schuller, and G. Rigoll, "Probabilistic ASR feature extraction applying context-sensitive connectionist temporal classification networks," in

Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, 2013, IEEE, pp. 7125–7129.

[247] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," *Multimodal Technologies for Perception of Humans*, pp. 509–519, Springer, 2008.

[248] M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *Proc. Int. Conf. on Multimedia and Expo (ICME)*, Amsterdam, The Netherlands, 2005, IEEE.

[249] C. Yam, M. Nixon, and J. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognition*, vol. 37, no. 5, pp. 1057–1072, Elsevier, 2004.

[250] S. H. Yella and H. Bourlard, "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, IEEE, pp. 7746–7750.

[251] S. H. Yella and F. Valente, "Speaker diarization of overlapping speech based on silence distribution in meeting recordings," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012, ISCA, pp. 490–493.

[252] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtcho, and P. Woodland, "The HTK book (for HTK version 3.4)," Tech. Rep., Cambridge University Engineering Department, Cambridge, UK, 2006.

[253] J. S. Yun, S. H. Lee, W. T. Woo, and J. H. Ryu, "The user identification system using walking pattern over the ubifloor," in *Proc. Int. Conf. on Control, Automation, and Systems (ICCASS)*, Gyeongju, Korea, 2003, pp. 1046–1050.

[254] M. Zelenák and J. Hernando, "The detection of overlapping speech with prosodic features for speaker diarization," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Florence, Italy, 2011, ISCA, pp. 1041–1044.

[255] M. Zelenák, C. Segura, and J. Hernando, "Overlap detection for speaker diarization by fusing spectral and spatial features," in *Proc. Annu. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, ISCA, pp. 2302–2305.

[256] M. Zelenák, C. Segura, J. Luque, and J. Hernando, "Simultaneous speech detection with spatial features for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 436–446, 2012.

[257] Z. Zhang, P. O. Pouliquen, A. Waxman, and A. G. Andreou, "Acoustic micro-Doppler radar for human gait imaging," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. EL110–EL113, 2007.

[258] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "HMM-based acoustic event detection with AdaBoost feature selection," *Multimodal Technologies for Perception of Humans*, pp. 345–353, Springer, 2008.

[259] Y. Zhu, T. Tan, and Y. Wang, "Biometric personal identification based on Iris patterns," in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, Barcelona, Spain, 2000, IEEE, pp. 801–804.

[260] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, IEEE, pp. 17–20.

[261] J. Žibert, N. Pavesić, and F. Mihelič, "Speech/non-speech segmentation based on phoneme recognition features," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 47–47, Hindawi Publishing Corp., 2006.