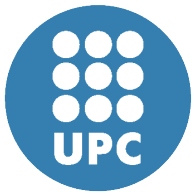
DETECTION AND HANDLING OF OVERLAPPING SPEECH FOR SPEAKER DIARIZATION

MARTIN ZELENÁK

Dissertation presented for the degree of Doctor of Philosophy



TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
Barcelona, Spain

Advisor: Prof. Javier Hernando Pericás

October 2011

Táto práca je venovaná mojej rodine . . .

ABSTRACT

For the last several years, speaker diarization has been attracting substantial research attention as one of the spoken language technologies applied for the improvement, or enrichment, of recording transcriptions. Recordings of meetings, compared to other domains, exhibit an increased complexity due to the spontaneity of speech, reverberation effects, and also due to the presence of overlapping speech.

Overlapping speech refers to situations when two or more speakers are speaking simultaneously. In meeting data, a substantial portion of errors of the conventional speaker diarization systems can be ascribed to speaker overlaps, since usually only one speaker label is assigned per segment. Furthermore, simultaneous speech included in training data can eventually lead to corrupt single-speaker models and thus to a worse segmentation.

This thesis concerns the detection of overlapping speech segments and its further application for the improvement of speaker diarization performance. We propose the use of three spatial cross-correlation-based parameters for overlap detection on distant microphone channel data. Spatial features from different microphone pairs are fused by means of principal component analysis, linear discriminant analysis, or by a multi-layer perceptron.

In addition, we also investigate the possibility of employing long-term prosodic information. The most suitable subset from a set of candidate prosodic features is determined in two steps. Firstly, a ranking according to mRMR criterion is obtained, and then, a standard hill-climbing wrapper approach is applied in order to determine the optimal number of features.

The novel spatial as well as prosodic parameters are used in combination with spectral-based features suggested previously in the literature. In experiments conducted on AMI meeting data, we show that the newly proposed features do contribute to the detection of overlapping speech, especially on data originating from a single recording site.

In speaker diarization, for segments including detected speaker overlap, a second speaker label is picked, and such segments are also discarded from the model training. The proposed overlap labeling technique is integrated in Viterbi decoding, a part of the diarization algorithm. During the system development it was discovered that it is favorable to do an independent optimization of overlap exclusion and labeling with respect to the overlap detection system.

We report improvements over the baseline diarization system on both single- and multi-site AMI data. Preliminary experiments with NIST RT data show DER improvement on the RT '09 meeting recordings as well.

The addition of beamforming and TDOA feature stream into the baseline diarization system, which was aimed at improving the clustering process, results in a bit higher effectiveness of the overlap labeling algorithm. A more detailed analysis on the overlap exclusion behavior reveals big improvement contrasts between individual meeting recordings as well as between various settings of the overlap detection operation point. However, a high performance variability across different recordings is also typical of the baseline diarization system, without any overlap handling.

It is a capital mistake to theorize before one has data.

— Sherlock Holmes in "Scandal in Bohemia"
by Sir Arthur Conan Doyle

ACKNOWLEDGMENTS

I have to admit that doing a Ph. D. over a period of four years was sometimes much more difficult than I have imagined. It was not only the investigation task itself, but also the dealing with doubts such as if one is on the right track, or if the work has actually sense that formed the imaginary rocks in the road that needed to be overpassed. I do not remember exactly who it was who said that the most challenging about doing a doctorate is not the research, but rather the problem of motivation, since it is a highly individual work usually faced with a lot of failures. I believe it is correct. Therefore, I find it important when one knows that he is not alone in this struggle.

First of all, I would like to thank my parents who were always believing in me, supporting me, and offering their help when it was necessary. The same applies for my sister Daniela, although being in a foreign country at a young age herself, she found her ways to show me that she is there for me. I appreciate it a lot.

I would certainly not be writing this thesis if it was not for my advisor Javier Hernando. I am very grateful for his guidance, for helping me not to lose sight of the objectives, and also for being understanding in these years. Thank you Javier. I am also very obliged to Gregor, friend and professor from my Alma Mater in Bratislava, who was standing at the beginning of my trip to Barcelona.

Some of my colleagues were crucial for doing this work. Especially Jordi and Carlos with whom I was working on speaker diarization and the extraction of spatial information from speech in meeting rooms. There are also other colleagues and friends that were important to me during this time at the UPC, particularly Mawo, Adolfo, Henrik, David, Carlos H., Tarás, and not to forget Lefty.

Finally, I cannot omit to express my gratitude to Monika who was walking a big part of the way with me. It was actually thanks to her, during a lunch conversation in winter 2010 in Bratislava's old town, that I got inspired to optimize the exclusion and labeling of overlapping speech independently.

Many thanks to all of you again, and also to everyone I forgot to mention and would have deserved it.

Martin Zelenák

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BIC	Bayesian Information Criterion
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CMS	Cepstral Mean Subtraction
DER	Diarization Error Rate
DPE	Diarization Posterior Entropy
DOA	Direction of Arrival
EM	Expectation Maximization
FF	Frequency Filtering
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
GCC-PHAT	Generalized Cross-Correlation with Phase Transform Weighting
GLR	Generalized Likelihood Ratio
HES	Harmonic Enhancement and Suppression
HER	Harmonic Energy Ratio
HPS	Harmonic Product Spectrum
HMM	Hidden Markov Model
ICA	Independent Component Analysis
JFA	Joint Factor Analysis
LPC	Linear Predictive Coding
LPCRE	Linear Predictive Coding Residual Energy

LDA	Linear Discriminant Analysis
MAP	Maximum A Posteriori Probability
MFCC	Mel Frequency Cepstral Coefficient
ML	Maximum Likelihood
MLP	Multi-layer Perceptron
MSG	Modulation Spectrogram
mRMR	minumum Redundancy Maximum Relevance
OD	Overlap Detection
OIP	Overlap Insertion Penalty
PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
PPF	Pitch Prediction Feature
RMSE	Root-Mean-Squared Energy
ROC	Receiver Operating Characteristic
SAD	Speech Activity Detection
SAPVR	Spectral Autocorrelation Peak-Valley Ratio
SF	Spectral Flatness
SVM	Support Vector Machine
TDOA	Time Delay of Arrival
UBM	Universal Background Model
ZCR	Zero-Crossing Rate

1 INTRODUCTION

1.1 SPEAKER OVERLAP CHALLENGE IN SPEAKER DIARIZATION

Since the beginning of the digitalization era we can observe a much easier access to various multimedia documents such as television shows, the news, lectures, meeting recordings, and many others. This continuous offer created a growing demand for the application of automatic human language technologies in order to allow for effective search and access of audio information sources. These technologies make it possible to extract from spoken documents meta-data that provides contextual information beyond words. The simplest example is to break up the signal into speech and non-speech segments by so-called Speech Activity Detection (SAD). For other purposes one may desire to have more details, such as the locations of music, narrow-band speech, or to know the gender of the speakers. Such tasks are generally known as audio diarization, i. e., marking and categorizing of audio sources within a spoken document [1].

Definition and applications of speaker diarization

A specific kind of diarization is *speaker diarization*. Given a speech recording, this task aims to answer the question: *"Who spoke when?"* Speaker diarization can vary according to the amount of prior knowledge that is provided, but in general it is assumed that nothing is known in advance. In further reading when referring to diarization normally speaker diarization is meant.

Diarization systems can be primarily used in three application domains: broadcast news audio, meeting room data, and telephone conversations. Their application is often a very useful preprocessing step for other audio technologies, such as Automatic Speech Recognition (ASR), speaker identification, speaker localization, etc. For instance, given the output of a speaker diarization system, ASR can carry out unsupervised speaker adaptation by joining segments from the same speakers, which can significantly improve transcription performance. Furthermore, the readability of automatic transcriptions can also be improved by structuring the audio stream into speaker turns, and eventually, coupled together with speaker identification, by providing the identity of speakers. This kind of information is of interest in indexation of multimedia documents [2].

In the early years most research in speaker diarization concentrated mainly on the broadcast news domain [3]. Over time, however, there started to be a strong interest in the meeting domain as well [4]. Meeting domain brings more difficulties for speaker diarization. Not only

is the speech completely spontaneous, with possibly large amount of silences for any speaker, but the recordings with different types of microphones positioned at various room locations lead to different signal qualities. Furthermore, the use of distant microphones makes the effect of room reverberation significant. All of these factors hinder speaker diarization. The spontaneity of speech also raises the importance of another issue, the one of *overlapping speech*.

Overlapping speech — a challenging problem for speaker diarization

It is a well known fact that people sometimes tend to speak at the same time, i. e., simultaneously. It is a normal part of human conversation behavior. For that reason, audio recordings of meetings commonly include regions of overlapping speech. This factor, however, poses a burden for a lot of spoken language technologies, speaker diarization being no exception. According to some studies [5, 6, 7], a portion of the performance degradation on real meeting data can be directly associated with the occurrence of speaker overlaps. Nevertheless, this specific issue became of interest to the scientific community only recently and the number of related works that have been published in the literature is thus far rather limited. Dealing with overlapping speech still remains a challenging problem.

1.2 OBJECTIVES

This thesis addresses the issues related to the occurrence of simultaneous speech in meeting recordings. The motivation is to improve speaker diarization performance, since conventional diarization systems suffer from this common conversation phenomenon. However, the investigation of overlapping speech may also be useful for other speech processing tasks such as speech, or speaker recognition.

There are several objectives this work attempts to meet. In the first place it is necessary to acquaint ourselves with the state of the art regarding overlapping speech, its detection and also further processing.

Thesis objective is to develop an overlap detection system in order to assist speaker diarization.

One of the main goals is the development of a robust overlap detection system. This system should work with distant channel data without any constraints about microphone configuration or the recording room. Our interest is to research and propose new features which may be useful for this task.

For instance, we aim at exploring the possibilities of employing spatial-based information for the detection of simultaneous speech since (smart) meeting rooms are normally equipped with microphone arrays. The availability of multi-channel data provides the option to estimate features that are in some way related to spatial location. Another option is to investigate the potential of higher-level information. "Higher" in this case refers to speech information which is above the level of short-term spectral or cepstral features, such as prosody.

The other main goal of this thesis is to apply the detected overlapping speech in the UPC speaker diarization system in order to reduce diarization error. This should be achieved by both recovering missed speaker time, as well as by improving the clustering. We seek to implement a novel technique for the assignment of extra speaker labels in speaker overlap segments. Different overlap detection systems will be examined according to the quality of their hypotheses for diarization improvement.

Finally, since our general intention is contribute to the research in human language processing, we participate in the organization of the Albayzin evaluation campaign. Our responsibility is the speaker diarization section. Such evaluations help comparing recent approaches in a particular field and generally stimulate the investigation progress.

1.3 ORGANIZATION OF THE THESIS

The rest of this thesis is organized as follows. Chapter 2 intends to give the reader a brief overview about the state-of-the-art techniques related to speaker diarization and overlapping speech. The basic idea of the chapter organization is to separate the topics on the detection of overlapping speech from speaker diarization and its improvement by the use of detected speaker overlap. Moreover, both topics are divided into a more theoretical part and into an experimental part. The first addresses the system design and acoustic features, and the second describes and interprets the obtained experimental results. Since the design of a system is often closely interrelated with experiments, in some cases the border of such a division is blurred.

Chapter 3 addresses the construction of the overlap detection system. A large part is dedicated to the discussion of different features which might be suited for this task. The UPC speaker diarization system and the techniques on how to improve its performance, given the knowledge about simultaneous speech in recordings, are explained in Chapter 4.

The audio data coming from AMI Meeting corpus, which is used throughout the work, is described in Chapter 5. In addition, this chapter introduces an alternative data corpus consisting of NIST RT meeting recordings. Experimental results of overlapping speech detection and speaker diarization are presented in Chapters 6 and 7, respectively. General discussion and conclusions are given in Chapter 8.

Finally, Appendix A reports on the Albayzin 2010 speaker diarization evaluation organized under the FALA 2010 workshop. The task, data, and submitted systems are described and the results are discussed.

2

STATE OF THE ART

This chapter begins with a brief introduction to the general concepts of acoustic modeling, classification, and segmentation in order to set up the framework in which speaker diarization and overlap detection operate. Then, an overview of the field of speaker diarization is given, with the focus on the most popular approaches and recent advances. In the end, the topic of simultaneous speech is discussed from various perspectives. Overlapping speech can be viewed as a separation problem, but in practice, when real (not artificially overlapped) audio recordings are used, it is necessary to firstly determine the locations of such segments. After reviewing the most successful approaches for overlap detection, the relationship between overlapping speech and speaker diarization performance is discussed, together with previous attempts of handling this issue in the given context.

2.1 ACOUSTIC CLASSIFICATION AND SEGMENTATION

Definition of segmentation and acoustic classification

The goal of segmentation is to divide an acoustic waveform, or a sequence of acoustic features, into certain segments that demarcate acoustic (or phonetic, linguistic etc.) units defined beforehand. The goal of classification is to perform an identification or "labeling" of these segments. For instance, in the case of Automatic Speech Recognition (ASR), the acoustic units can be defined as phonemes or words. Segmentation and acoustic classification are processes which can be performed sequentially or in parallel. The methods for acoustic classification (also referred to as decoding) can be basically divided into heuristic-based approaches, distance-based approaches and probabilistic approaches [8]. The first two are mentioned only for completeness, in the following, only the third concept is considered. Satisfactory management of this process is especially important in classifiers which are dealing with a large set of classes or spontaneous speech.

Hidden Markov Models

The probabilistic approach is usually based on the use of Hidden Markov Models (HMMs). The HMMs are one of the most commonly applied probabilistic finite-state machines. They have the ability of modeling sequences of states that cannot be observed directly, since they are *hidden*, but only through sequences of statistically related observations. These models are created for acoustic realization of every analyzed unit, e. g., every word in a system dictionary. Alternatively, Markov models can be constructed for smaller units (phonemes), and then words and phrases are modeled by their concatenation. The

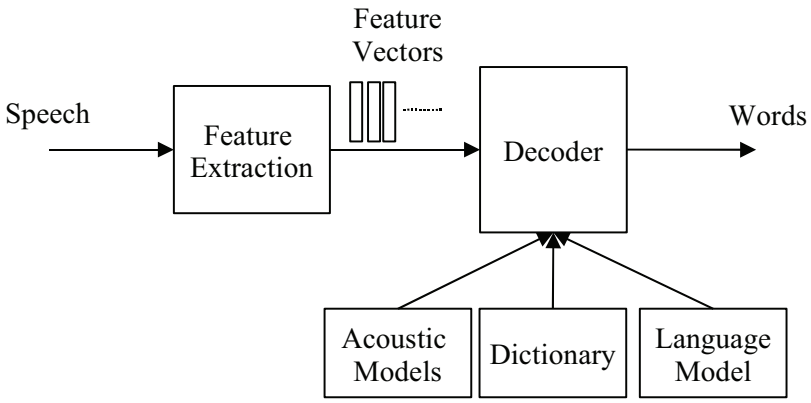


Figure 1: Architecture of a typical HMM-based recognizer [9].

architecture of a typical HMM-based recognizer (ASR system) is given in Figure 1.

Given a sequence of acoustic observations, i. e., feature vectors, $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$ extracted from an input audio signal, the decoder tries to find the sequence of words $\mathbf{w} = w_1, w_2, \ldots, w_L$ which most likely have generated Y. Since it is difficult to model such probability $P(\mathbf{w} \mid \mathbf{Y})$ directly using the generative HMMs, Bayes' rule is applied to transform the task into

$$\hat{\mathbf{w}} = \arg\max_{w} p(\mathbf{Y} \mid \mathbf{w}) P(\mathbf{w}), \tag{2.1}$$

which is an equivalent problem. The likelihood of the observation sequence given a word sequence $p(\mathbf{Y} \mid \mathbf{w})$ is determined by an *acoustic model*. $P(\mathbf{w})$ is the prior probability of observing a particular word sequence and is normally determined by a *language model*.

A typical structure of a left-right phone model is illustrated in Figure 2. HMM makes a transition from its current state to one of its connected states every time step. For first-order Markov chains used to model stochastic processes it is assumed that the condition in any state only depends on the previous state and observations are conditionally independent of all other observations given the state that generated it. The probability of making a particular transition from state i to state j is given by the transition probability matrix $\mathbf{A} = \{a_{ij}\}$. The parameter set $\mathbf{B} = \{b_j(\cdot)\}$ holds the emission probability functions associated with each state of the model, $b_j(\mathbf{y_k})$ is the probability of emitting observation $\mathbf{y_k}$ on entering the state j. The most common approach for modeling the feature distribution is by using continuous density Gaussian Mixture Model (GMM), that is

Gaussian Mixture Models

$$b_{\Theta_t}(\mathbf{y}) = \sum_{m=1}^{M} w_m N(\mathbf{y}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{2.2}$$

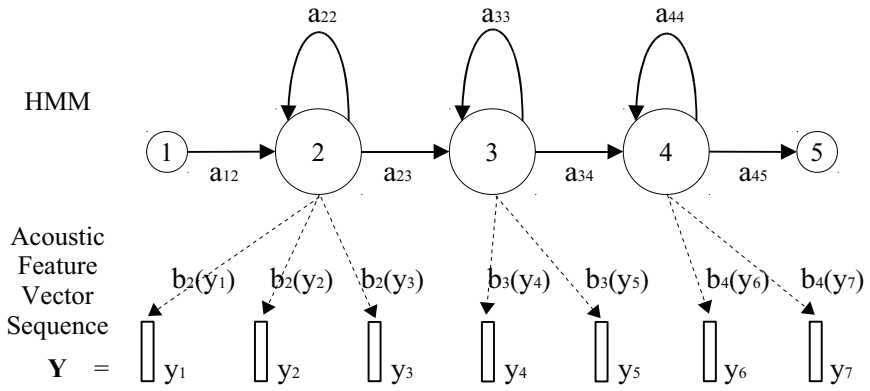


Figure 2: An example of a left-right HMM used for a phoneme [9].

where w_m is the weight of the m^{th} mixture component, and $N(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. While the model supports full covariance matrices, these are usually used in their diagonal form. Furthermore, the model's mixture weights w_m sum to unity. The number of mixtures is usually a subject of a trade-off between the model accuracy and the generalization on unseen data.

Given a model λ and an observation sequence $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$, there are three problems which need to be addressed to effectively use HMMs in real applications [10, 11]:

The three practical problems with the use of HMMs

- **The likelihood problem.** How do we estimate the likelihood of the model that generates the observations, i. e., $p(\mathbf{Y} \mid \lambda)$?
- **The learning problem.** How do we find a new model estimate $\hat{\lambda} = \{\mathbf{A}, \mathbf{B}\}$ which maximizes the likelihood $p(\mathbf{Y} \mid \lambda)$?
- **The decoding problem.** How do we find the state sequence $\boldsymbol{\Theta} = \{\Theta_1, \ldots, \Theta_T\}$ that generates $\mathbf{Y}$ with the highest probability?

The likelihood can be efficiently estimated in a recursive manner by computing *forward-* and *backward-* probability variables. For the learning problem no analytical method has been presented so far that would ensure finding the global maximum of the probability of model λ generating the sequence $\mathbf{Y}$, $p(\mathbf{Y} \mid \lambda)$. Nevertheless, iterative procedures were suggested which choose $\hat{\lambda}$ so that this probability is maximized at least locally on the training data.The most popular solution to this problemis a particular version of the Expectation Maximization (EM) technique suitable for HMMs, known as the *Baum-Welch* re-estimation algorithm.

The decoding problem is addressed with the *Viterbi algorithm*, one of the most widely applied decoding approaches. The goal of uncovering

Viterbi decoding

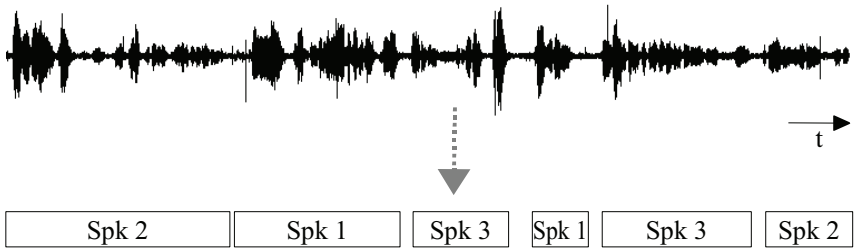


Figure 3: The objective of speaker diarization.

the best word sequence can be approximated by finding the most likely sequence of hidden states (called Viterbi path),

$$\hat{\mathbf{w}} \approx \arg\max_{w}\{P(w)\max_{\Theta} p(\mathbf{Y}, \mathbf{\Theta} \mid \mathbf{w})\}. \tag{2.3}$$

Two operations are involved. First, the estimate of the highest probability along a path of length T through the states of the HMM is found, and then the single states $\hat{\Theta}_1, \ldots, \hat{\Theta}_T$ of the best path are determined. For more details on these algorithms refer to one of [10, 11, 12, 13].

Speaker modeling

When the task of a classification system is to distinguish among different speakers (e. g., speaker identification, verification), we are faced with the question what is the best way to model the voice of a speaker. The most widely applied approach to speaker representation is based on GMMs and was presented by Reynolds in [14]. A GMM can be considered a one-state HMM. In applications where there is strong prior information on the spoken text, additional temporal knowledge can be incorporated by using multiple-state HMMs as the basis for the likelihood function.

In some situations the amount of training data for particular acoustic classes, such as speakers, for instance, is limited. A common solution how to deal with this problem is adaptation. Basic idea of adaptation is to derive the speaker's model by updating well-trained parameters in a so-called Universal Background Model (UBM) using the speaker's training speech and a form of Bayesian adaptation. Comprehensive explanation can be found in [15]. This adaptation, Maximum A Posteriori Probability (MAP) estimation of Gaussian mixtures, was originally introduced by Gauvain and Lee in [16].

2.2 SPEAKER DIARIZATION

Speaker diarization, tracking and indexing

Speaker diarization task consists of segmenting a conversation involving multiple speakers into speaker-homogeneous parts and grouping together all the segments that correspond to the same speaker. The objective of speaker diarization is illustrated in Figure 3. The first part of the process is also referred to as speaker segmentation or *speaker*

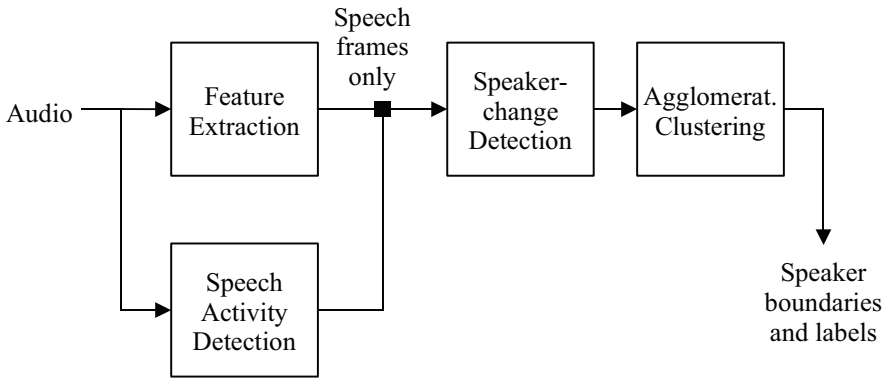


Figure 4: Basic concept of a speaker diarization system.

change detection, and the second step is known as *clustering*. The most common condition for speaker diarization is that the number of speakers and speaker characteristics are a priori unknown to the system and need to be determined automatically. For completeness, we can also mention the *speaker tracking* task which follows (or tracks) a speaker identity in an audio recording. The basic difference between diarization and tracking is that in the latter the voice characteristics of a particular speaker must be known beforehand, similarly to speaker identification or verification. Another frequently occurring term in this field is *indexing*. It can be understood as performing diarization on an audio database and eventually associating the time stamps with true speaker identities in order to have better overview and search options in recordings.

Sequential vs. integrated approach to diarization

We can find in the literature a lot of diverse approaches to the speaker diarization problem, however, there are two predominant strategies. The *step-by-step* strategy deals with the main steps successively, first finding the speaker turns, and second, re-grouping the segments coming from one speaker during the clustering phase [17, 18, 19]. A limitation of this method is that it is not only difficult to correct the errors made in the segmentation later on, but these errors degrade the performance of the subsequent clustering step. The basic concept of a speaker diarization system with individual subtasks is depicted in Figure 4.

An alternative approach, referred to as *integrated* strategy, is to optimize the segmentation and clustering jointly [20, 21]. Both steps are carried out simultaneously in an iterative procedure which uses a set of GMMs or an ergodic HMM. The main disadvantage of the integrated approach lies in the need to learn these models using very short segments, even though the speaker models get refined along the process.

Mixed strategies are also proposed where the classical step-by-step segmentation and clustering is applied first, and then the segment boundaries and clusters are refined jointly [22, 23, 24]. Fusion of both techniques can be found in [25]. The two steps, independently of the strategy, are discussed later in this chapter.

2.2.1 *Acoustic Features*

In almost any kind of pattern recognition system, one of the basic steps is the extraction of features from raw data. Features are measurable characteristics which are important to the distinction between different classes. They should not only have possibly low inter-class similarity, but also low intra-class variability. In the context of speaker recognition, features obtained from the speech signal attempt to reflect the discriminative speaker information. Since speaker diarization and recognition are closely related, commonly used features are very similar.

A standard in the field is to extract short-term low-level acoustic features derived from speech spectrum. The spectrum of the speech is closely related to the physiology of the human vocal tract, an important discriminating factor. By far the most popular are the Mel Frequency Cepstral Coefficients (MFCCs), which showed to perform well in speaker recognition tasks [26, 27] and, somehow ironically, in speech recognition as well. Static MFCCs are also the most widely applied features in the majority of state-of-the-art speaker diarization systems.

Short-term cepstral features

Apart from MFCCs, other used parameters are the linear predictive coding (LPC) coefficients, frequency filtered (FF) filter-bank energies [28], linear frequency cepstral coefficients, and perceptual linear predictive (PLP) parameters. In speaker recognition, for instance, first- and second-order time derivatives (also called delta and delta-delta coefficients) are usually also obtained to assist the recognition. However, speaker diarization systems often do not use deltas, especially not for acoustic change detection, since this practice empirically turned out not to be very successful.

Variable channel or background conditions can sometimes seriously degrade the performance of automatic systems. To compensate for these variations, several normalization techniques have been proposed in feature, score, or decision domain. We will focus on feature normalization used for speaker-related tasks. Feature warping normalization introduced by [29] was applied for diarization in [30]. Here, the distribution of a cepstral feature stream is warped to a standardized distribution over a specified time interval. Another technique called feature mapping [31] maps features from different channels into a

common channel-independent feature space using previously learned linear transformation.

Long-term features

Speech signal conveys several different levels of information on which the humans rely in order to recognize others by voice. This information reaches from cues related to physical traits to cues related to learned habits and talking style. Automatic human language processing area was dominated by systems using acoustic information. Since several years, an increased effort could be observed to combine low-level features with higher-level information. For instance in [32], wide ranging approaches using pronunciation models, prosodic dynamics, pitch gestures, phone streams, and conversational interactions were explored. The potential contribution of prosodic information to automatic processing of meeting/broadcast data was suggested in several works, such as [33, 34]. A successful application of long-term prosody-based features in combination with conventional parameters for speaker diarization was eventually presented by Žibert and Mihelič in [35] and Friedland et al. in [36]. In a related work, Imseng and Friedland [37] use the prosodic features for the initialization of agglomerative clustering. Pitch, energy, peak-frequency centroid and peak-frequency bandwidth are examples of features considered for speaker segmentation by [38]. Furthermore, three new features related to the cross-correlation of the signal power spectrum are investigated, namely, temporal feature stability, spectral shape, and white noise similarities.

Modulation spectrogram provides an alternative representation of the speech signal with a focus on temporal structure, it represents a filtered version of a spectrogram. It was observed that modulation spectrogram features also carry speaker-specific information and together with MFCCs can aid the speaker diarization task [39].

Location-related features

When speech is recorded in multi-channel environment , it is possible to extract complementary discriminative information which reflects the time-delays of signal between microphones. There are several works addressing this topic. A technique that segments audio recording according to speakers based on their locations was proposed by Lathoud and McCowan [40]. In this paper, speaker locations are obtained by estimating Time Delay of Arrival (TDOA) values from cross-correlation peaks between paired microphones within an array. The same technique was used in combination with a MFCC-based system to improve diarization performance in [41]. Pardo, Anguera, and Wooters [42] considered this approach, i. e., combining MFCC and TDOA feature streams, also for the general case when the location of microphones is unknown. In [43], the use of Direction of Arrival (DOA) information was explored to assist the speaker change detection. Exploiting this spatial information led to significant improvement compared to results achieved with close-talking microphones. In addition, spatial

information can aid the initialization of speaker clusters, as discussed in [44].

Speaker factors

Recently, Joint Factor Analysis (JFA) methods have demonstrated very good results addressing issues such as channel- or speaker-variability compensation. Moreover, JFA is among the state-of-the-art techniques for speaker and language recognition. An effective factor analysis scheme for speaker diarization was firstly proposed by Castaldo et al. in [45] and later extended by Kenny et al. [46]. The main idea is to exploit prior knowledge about the speaker space to find a low dimensional vector of speaker factors that summarize the distinctive speaker characteristics.

2.2.2 *Speaker Segmentation*

Acoustic change detection, in general, aims to timestamp an audio stream according to the changes in acoustic conditions. For speaker segmentation the focus is on detecting speaker turns in a recording. The literature offers several methods addressing this problem that can be roughly categorized into three groups: *metric-*, *model-*, and *silence-based* algorithms.

Silence-based segmentation

Silence-based segmentation chops the audio stream in the silence locations either using an energy threshold [47] or a decoder-guided technique [48, 49, 3]. In the later case, the silences marking segment boundaries are detected by a recognition system, usually putting constraints on their minimum duration. As there is no clear relationship between silences in a recording and speaker turns, such techniques are seldom used for diarization.

Model-based segmentation

Model-based segmentation performs a Maximum Likelihood (ML) classification with trained models (GMMs) corresponding to a closed set of acoustic classes [18, 47]. The boundaries between assigned classes become the segmentation change points. Examples of acoustic classes can be telephone/wideband channel, male/female voice, music/speech/silence, or their combinations. Pre-trained models can face a robustness problem, though. Model-based techniques are playing a major role in the integrated diarization strategy where segmentation and clustering are performed at the same time, searching for optimal acoustic change points without any previous knowledge of the acoustic classes [20, 50].

Metric-based segmentation

Probably the most popular approach is *metric-based segmentation* [27, 19, 51]. Here, two neighboring windows of a relatively small size are moved over the audio signal. The similarity between data in these two windows is determined by a distance function. Acoustic change points are put to locations where distance function local maxima exceed some threshold value. Various metric-based algorithms differ according to the kind of distance function they apply, the length of

the two windows, their time shift, or the way the similarity values are evaluated. The most common distance metric is based on the Bayesian Information Criterion (BIC). However, various other forms of distance measures such as symmetric Kullback-Leibler (KL) divergence, Generalized Likelihood Ratio (GLR), or Gish distance exist in literature.

A novel approach concerning distance metrics for speaker diarization was investigated in [52]. Here, a dissimilarity measure based on an one-class Support Vector Machine (SVM) is used in both speaker turn detection and clustering, and the obtained results are very competitive to standard methods.

2.2.2.1 *Bayesian Information Criterion*

Bayesian Information Criterion definition

BIC was originally introduced by Schwarz [53, 54] and its popularity lies in its simplicity and effectiveness. BIC value informs how well a model fits some data. Consider modeling the acoustic data $\mathbf{X} = \{x_i \in \mathbb{R}^d; i = 1, \ldots, N\}$ using a model M. For the model M we assume that $\#(M)$ parameters are chosen to maximize the likelihood and let $L(\mathbf{X}|M)$ denote this maximum value. BIC is a likelihood criterion penalized by the number of parameters in the model and is defined as follows:

$$\mathrm{BIC}(\mathbf{X}, M) = \log L(\mathbf{X}|M) - \lambda \frac{1}{2} \#(M) \log(N), \tag{2.4}$$

where the penalty weight λ is a free tuning parameter (but only $\lambda = 1$ corresponds to the strict definition of BIC) and N is the number of observations in the acoustic segment. For the purpose of speaker segmentation, BIC was firstly proposed by Chen and Gopalakrishnan in [17, 55] and later also in [56]. Considering that the feature sequence $\mathbf{X}$ is drawn from an independent multivariate Gaussian process, a change at time i is resolved with a hypothesis that consecutive segments $\mathbf{X}_i : x_1 \ldots x_i$ and $\mathbf{X}_j : x_{i+1} \ldots x_N$ are better modeled with models M_i and M_j, respectively, than the two segments jointly by a single model M. It can be viewed as a model selection problem—the data is modeled by one or two Gaussians—what is determined by computing the difference between the BIC values for the hypotheses:

$$\begin{aligned} \Delta\mathrm{BIC}(i) = \log L(\mathbf{X}|M) - \big(\log L(\mathbf{X}_i|M_i) + \log L(\mathbf{X}_j|M_j)\big) \\ - \frac{\lambda}{2} \#(M_i, M_j) \log(N), \end{aligned} \tag{2.5}$$

where $\#(M_i, M_j) = \#(M) - \big(\#(M_i) + \#(M_j)\big)$ denotes the difference in the number of parameters between model M and models M_i,M_j. The two-model hypothesis is favored if ΔBIC is positive, the ML changing point can be expressed as $\hat{t} = \arg\max_i \Delta\mathrm{BIC}(i)$. Several works addressed the fine tuning of the penalty weight parameter [57, 58, 49], or it was discarded totally [59].

Modifications of BIC

In the majority of implementations the search for the segment

boundaries is carried out iteratively with a growing window [60]. Such algorithm is robust, but unfortunately computational complexity is quite high. Various modifications were proposed in order to address this issue. For instance, a two-pass mechanism called DISTBIC [57, 58] first makes a rough selection of change points with a faster GLR metric, and in the second pass BIC is used for refinement. Alternate distance measures seeking to reduce the computational load include XBIC [61], which was shown to be faster while having similar performance.

Since it is a known issue that BIC does not perform well on short segments, a MAP adaptation of speaker models that allows to detect shorter speaker changes was suggested in [62]. Recently, so-called MultiBIC segmentation scheme was introduced in [63] where two change points are assumed to be present in a window of data instead of the usual one. It is supposed to considerably reduce the number of undetected short segments.

2.2.2.2 *Generalized Likelihood Ratio*

GLR was introduced by Willsky and Jones [64] for change detection and constitutes a likelihood ratio test between two hypotheses. Given two adjacent portions of parameterized audio signal $\mathbf{X}_1$ and $\mathbf{X}_2$ (similar to BIC), the first hypothesis assumes that both portions $\mathbf{X} = \mathbf{X}_1 \bigcup \mathbf{X}_2$ are modeled by one Gaussian process, i. e., both are generated by the same speaker. The alternate hypothesis, on the other hand, considers that segments originate from different speakers and therefore two Gaussians are a better representation of the data. The distance measure is computed as the log-value of the likelihood ratio between the two hypotheses:

$$d_{GLR} = -\log \frac{L(\mathbf{X}|M(\mu, \Sigma))}{L(\mathbf{X}_1|M_1(\mu_1, \Sigma_1)) \cdot L(\mathbf{X}_2, M_2(\mu_2, \Sigma_2))}, \tag{2.6}$$

where $M(\mu, \Sigma)$ denotes a Gaussian process with mean μ and covariance Σ trained from $\mathbf{X}$ (by EM algorithm). A low value of d_{GLR} signifies the modeling with one Gaussian. In contrast, a high value of d_{GLR} indicates that the second hypothesis should be preferred and suggests a speaker change on the border between the two segments.

DISTBIC — Algorithm combining GLR and BIC

For segmentation, GLR is usually used together with BIC in a two-step process proposed by Delacourt, Kryze, and Wellekens [57, 58]. First, the most likely speaker changes are detected, and then they are validated or discarded during a second pass (previously listed as the DISTBIC algorithm). A variation of GLR, called Gish distance, was used in [27, 65] and proved efficiency for the identification task.

2.2.2.3 *Kullback-Leibler Distance*

The average discriminating information between two classes is known as the Kullback-Leibler number and was initially defined in [66]. Later

introduced by Siegler et al. [19] and Campbell [26] for speaker-related topics, KL measures the dissimilarity between two distributions of random variables. Considering X and Y being random variables, KL is formulated as follows:

$$KL(X,Y) = E_X\{\log P_X - \log P_Y\}, \tag{2.7}$$

where E_X denotes the expectation computed with the probability density function P of X (see [19]), reflecting the distribution of samples in an acoustic segment. A symmetrical measure is obtained as

$$KL2(X,Y) = KL(X,Y) + KL(Y,X). \tag{2.8}$$

Symmetrical Kullback-Leibler distance — Cross-Entropy

KL2 is sometimes also referred to as cross-entropy [67]. It is based on the fact, that an utterance is expected to have a large likelihood with respect to its own model, but a small likelihood for a different model. For Gaussian variables X and Y, KL2 can be rewritten as

$$KL2(X,Y) = \frac{\sigma_X^2}{\sigma_Y^2} + \frac{\sigma_Y^2}{\sigma_X^2} + (\mu_X - \mu_Y)^2 \left(\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2}\right) - 1. \tag{2.9}$$

Compared to GLR which requires a model to be trained for each segment plus another model for them both, KL distance only requires one model for each segment [67]. A comparison between KL and Gish distance was presented in [47].

2.2.3 *Clustering*

A loose definition of clustering could be: "The process of organizing objects into groups whose members are similar in some way" [68]. For adaptation of acoustic models, for instance, it is enough to group together speakers which are acoustically similar. However, speaker diarization, in general, intends to arrive to an accurate distinction between speakers and tries to aggregate all the speech segments during the clustering process that belong to a particular speaker. Although in some cases the number of speakers or even their identity is known, in the following we only consider blind clustering, where there is no initial information at all. Ideally, we arrive to the final number of clusters equal to the number of speakers [69]. In such case each speaker is not assigned a true identity but rather a unique identifier. It is an identification task to link each identifier to a speaker identity.

State-of-the-art systems rely on hierarchical clustering

Most state-of-the-art systems rely on a hierarchical clustering[1] scheme [27, 19]. The optimal number of speaker clusters is determined by a subsequent splitting, or merging, of clusters in an iterative process until a stopping criterion is met.

1 Besides hierarchical clustering there also exist exclusive, overlapping, and probabilistic clustering in the literature.

Depending on the strategy, we differentiate between *bottom-up* (agglomerative) and *top-down* (divisive) clustering. As implied from previous statements, two crucial parts of a clustering mechanism are:

- *distance measure* between clusters to ascertain their acoustic similarity;
- *stopping criterion* to know when to stop the algorithm.

When the number of speakers is unknown, a distance threshold usually defines the stopping condition.

Agglomerative clustering is more frequent than divisive clustering

During the *agglomerative clustering* process, in each iteration a matrix is usually defined that holds the distances between all possible pairs of clusters, and the closest pair is merged together [15]. For instance, Chen and Gopalakrishnan [17, 55] suggested BIC for this distance. Cluster pairs assigned for merging have the highest ΔBIC value and when all pairs have $\Delta\text{BIC} < 0$ the merging finishes. BIC with some modifications was also employed in [56, 3].

Another from the presented metrics, the KL2, was used by [19] as a cluster distance measure and a stopping criterion. Combinations of the distances in segmentation and merging stages are assessed in several publications, KL2 and GLR were applied in [67] for iterative merging until the cluster purity was maximized. Distances adapted to the multi-dimensional Gaussian mixture case were introduced in [51, 70]. Particular interest was given to systems which derive speaker models for each cluster by means of a MAP adaptation of a UBM [71, 30].

Some other works, for instance by Ajmera et al. [69, 72] or Wooters et al. [73], integrate segmentation with clustering using model-based schemes and use BIC as the stopping criterion [72, 73]. The acoustic signal is initially segmented, and then iterative ML decoding is performed using adaptive GMM models. These approaches make also use of a hierarchical HMM to introduce temporal constraints to segment lengths.

A two-pass clustering was introduced in [22]. In the first step, the data is equally segmented and an agglomerative clustering is performed using a GLR distance matrix until the desired number of clusters is reached. In the second step, an integrated model-based approach of decoding and retraining follows until the likelihood converges.

Systems that rely on the *divisive clustering* scheme start with only one initial cluster which is then iteratively split until the algorithm stops on the optimal number of clusters [74, 3].

Combined and other approaches

Bottom-up systems sometimes suffer from merging instability and stopping criteria difficulties. On the other hand, top-down systems are particularly prone to poor model initialization, which can lead to large variations in performance. A number of works tried to combine both approaches in different ways. For instance, one possibility is

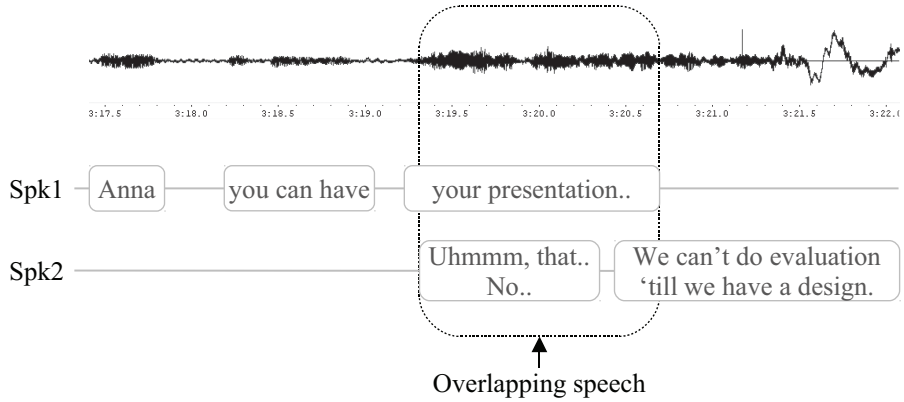


Figure 5: Overlapping speech example. Sample taken from the AMI Meeting corpus, recording IS1000d at 00:03:17 h.

that a top-down system is initialized with the segmentation output of a bottom-up system [75]. Alternatively, after matching hypothesized outputs, the common segments are associated together and a re-segmentation of the data where the two systems differed follows [75]. Only recently, another integrated solution was proposed where bottom-up and top-down systems are fused "at the heart" of the segmentation and clustering stage in [76].

Interesting work which shows how linguistic patterns can be used to identify the current, previous, or next speaker in order to improve and enrich the diarization output was presented in [77].

2.3 OVERLAPPING SPEECH

Overlapping speech refers to situations when two or more speakers speak simultaneously so that a listener hears a mixture of their voices. This kind of behavior is very natural for humans and occurs quite commonly in conversations [78]. An illustration of a speaker overlap is given in Figure 5.

Overlapping speech was identified in several publications as one of the major challenges for spoken language applications [33, 79]. We can distinguish several types of simultaneous speech which affect the flow of discourse in different ways. In a meeting environment an overlap can be categorized as one of the following [80]:

Categorization of speaker overlaps

- Floor grabber, to try to usurp floor from another speaker (*well I*);
- Backchannels, to encourage a speaker to continue (*right*, *uhhuh*);
- Interruptions (*so that's-*);
- Question, to determine the further discussion content (*and the new machines are faster?*);

- Statement (*it's easier just to buy new disks*);
- Other.

In addition, some overlaps may happen accidentally or part of a joint action (people trying to help a speaker to recall something) [33].

2.3.1 *Cocktail Party Problem and Source Separation*

The *cocktail party effect* describes the ability of human listeners to focus one's attention on a single talker among a mixture of conversations and background noises, for instance, during a loud crowded party. This phenomenon enables most people to talk in noisy places.

Cocktail party problem was first formulated by C. Cherry in 1953

Wang and Brown state in [81] that in "cocktail party"-like situations, in which all voices are equally loud, speech is intelligible for normal-hearing listeners even when as many as six interfering talkers are present. Binaural hearing is important to this effect, because it was observed that with interfering noise the understanding ability of listeners using only one ear was much more decreased. It still was not lost, though. An interesting observation was reported by Kashino and Hirahara [82]. When listeners were asked to guess the number of people speaking simultaneously in a recording (2–10 speakers), they mostly answered that they heard three speakers. It follows that even though humans can isolate and concentrate on a single source, they have problems to correctly determine the number of concurrent sources.

Since many years the perceptual segregation of sounds has been the subject of extensive research. A practical realization of this problem via computer analysis of microphone recordings is known as *source separation*. Even though the cocktail party problem is not very difficult to deal with for humans, it is non-trivial for machines.

Blind Source Separation with Independent Component Analysis

A popular statistical approach to separation through the use of multiple microphones is Blind Source Separation (BSS). BSS refers to the problem when there is no prior knowledge of the mixed signal, i. e., the mixing process is unknown [83]. When the number of sources to estimate is no more than the amount of sensors and independence of the source signals is assumed, a powerful tool for BSS is Independent Component Analysis (ICA) [84]. The difficulty of separating recorded speech signals is due to the delays and reflections in a real environment. Therefore, the mixing process is not linear as assumed in the basic BSS, but rather convolutive. Various solutions were proposed using iterative algorithms based on minimizing cross-channel correlation (termed adaptive decorrelation filtering) [85, 86], taking the problem into Z-domain and applying information maximization principle [87], or maximizing non-Gaussianity [84].

Computational Auditory Scene Analysis

The source separation problem is also the focus of study in Computational Auditory Scene Analysis (CASA), which draws inspiration from mechanisms of human auditory function. CASA seeks to exploit psychoacoustic features of speech to perform the separation. The general assumption is that although two speakers might speak simultaneously, there is little overlap in the time-frequency plane (spectrogram) if the speakers are different. CASA-based separation techniques partition the audio spectrogram so that each partition belongs to the speech of one of the overlapping speakers. This partitioning typically uses grouping cues such as pitch, onset times, offset times, and continuity [88, 89], or is based on modulation frequencies as in [90].

Assuming that $s(t)$ is the estimated source signal, the basic difference between BSS and CASA algorithms can be explained as follows. In the first case, the unmixing has the form:

$$s(t) = \alpha_1 m_1(t) + \alpha_2 m_2(t) + \cdots + \alpha_k m_k(t), \tag{2.10}$$

where $m_i(t)$ are simultaneous signals recorded with different microphones. The unmixing coefficients α_i are constant over time and chosen to optimize some property of the set of the recovered sources. In the later case, the basic principle of the refiltering method presented in [91] is to construct the sources by selectively reweighing (masking) the frequency subbands with automatically learned masking functions. Denoting the masking signals $\alpha_i(t)$ and subband signals of the original input $b_i(t)$, the source $s(t)$ can be obtained as follows

$$s(t) = \alpha_1(t) b_1(t) + \alpha_2(t) b_2(t) + \cdots + \alpha_n(t) b_n(t). \tag{2.11}$$

A related method to CASA was inspired by the pioneer work on source separation based on speech periodicity and harmonic selection in [92]. Relying on harmonic structure within speech, [93] proposed a Harmonic Enhancement and Suppression (HES) system for the separation of two speakers. Using the pitch estimate of the stronger speaker, his speech is recovered by enhancing its harmonic frequencies and formants. Speech of the other speaker is then obtained from the residual signal when the first speaker's harmonics are suppressed.

However, these methods have various limitations. For example, ICA has difficulties with one or more of the conditions of conversational speech. Many have problems in the presence of reverberation and nearly all source separation algorithms assume that the number of speakers in known [94]. A comparison of these techniques for segregation of speech from concurrent sounds concludes that BSS outperforms CASA for the majority of noise conditions [94]. Nevertheless, unmixing algorithms (BSS), in general, cannot operate on single-channel recordings, whereas CASA-based techniques can.

2.3.2 *Overlap Detection*

The source separation algorithms listed in Section 2.3.1 work with the assumption that the processed data already includes overlapping speech. However, when working with real (not artificially mixed) data, where overlap regions are not annotated, such segments need to be detected in the first place. The goal of overlap detection is to identify accurate temporal locations where several speakers are speaking concurrently.

Multi-speaker SAD on personal channel microphones

Several algorithms were published in the literature that detect overlapping speech as a result of multi-speaker Speech Activity Detection (SAD) on personal close-talking microphones. For instance, given a multi-channel meeting recording, Pfau et al. [95] applied speech/non-speech detection for every individual participant's channel to create preliminary hypotheses. Then, for regions where more than one channel was hypothesized as active, cross-correlation analysis was used to correct the false overlap regions. These were caused mainly due to crosstalk between nearby speakers. It was expected that the cross-correlation would be higher for crosstalk than for real overlaps. Since the SAD produced output for each channel separately, the system was also able to identify the regions of speaker overlap.

Inspired by Pfau's work, Wrigley et al. [96, 97] proposed to use a classifier based on an ergodic HMM (eHMM) to detect four classes: speaker alone, speaker+crosstalk, crosstalk alone, and silence. During the classification of multi-channel meeting data, each channel was classified by a different eHMM in parallel. This allowed for the application of a set of dynamic transition constraints so that only legal combinations of channel classifications were possible. Wrigley et al. considered a number of candidate features among which were MFCCs, energy, Zero-Crossing Rate (ZCR), Pitch Prediction Feature (PPF), kurtosis, so-called "fundamentalness", Spectral Autocorrelation Peak-Valley Ratio (SAPVR), and cross-channel correlation. He found kurtosis, fundamentalness, and cross-channel correlation metrics to be the best performing features.

Fundamentalness is based on the amplitude (AM) and frequency modulation (FM) extracted from the output of a bandpass filter analysis [98]. It is defined as having maximum value when FM and AM magnitudes are minimum, which corresponds to situation when the minimum number of components is present in the response area of the filter. When more speakers are active, interference of more (than one) fundamental components introduces modulation, thus decreasing the fundamentalness measure.

Signal amplitude kurtosis (will be discussed in Section 3.2) and SAPVR were applied by the authors of [99, 100, 101] for spotting usable speech segments in the context of speaker identification and speech

recognition. SAPVR metric relies on changes to the harmonic structure of the signal and is computed from the autocorrelation of the signal spectrum. The motivation for using SAPVR for overlap detection was that a single speaker should have a strongly periodic autocorrelation whereas in case of multiple speakers the autocorrelation function should be flatter due to the overlapping harmonic series. However, the performance on real meeting data reported in [97] was rather poor.

Lewis and Ramachandran [102] developed the PPF for speaker count labeling. The basic principle is that the distances between estimated successive pitch peaks should be more regular in single-speaker speech than in simultaneous speech. Similarly to SAPVR, the experiments were only performed on synthetically overlapped single-speaker data and Wrigley [97] summarizes that PPF was "not robust for real acoustic mixtures", giving "mediocre results".

In [103], a multi-pitch tracking algorithm was employed for a similar task of classifying pre-segmented multi-speaker audio into: local speech, crosstalk, overlapping speech, and non-speech.

A simple and efficient algorithm by Laskowski et al. for segmenting multi-speaker meeting data can be found in [104]. Without the necessity of training any model, joint maximum cross-correlation of personal microphone pairs was used to detect speech for each microphone wearer. Obviously, two simultaneously active speakers would mean speaker overlap. In another work, Laskowski and Schultz [105] proposed an algorithm which combines multi-speaker SAD with the idea of overlapping speech states. In contrast to the approach by Wrigley et al. [96] their eHMM had 2^K states, specifying every combination of speech and non-speech for each of K participants.

Approaches utilizing microphone array far-field data

A few algorithms for speaker overlap detection make use of distant-channel data. Lathoud and McCowan [40] suggested to segment the audio according to speakers using microphone-pair time delays (TDOAs) and showed the possibility to detect two simultaneous talkers by modeling short-term turns for each speaker combination. However, a constraining condition was that the number of speakers had to be known beforehand and it was assumed that their location will not change during the meeting. Since all the possible overlap combinations have to be modeled explicitly, this strategy suffers from an explosion of overlap classes.

In a later work, Lathoud et al. [106] presented two alternative strategies which produce individual segmentations for each participant and handle overlapping speaker combination implicitly. Here, the need to define all possible combinations of active speakers is avoided and the computational load is linear to the number of speakers. The segmentation strategies are based on speech/silence ratio or steered response power, confined to particular physical regions. Again, the number and location of meeting participants has to be fixed.

Otterson [107], in his thesis, also investigated the possibility to detect overlapping speech using multi-microphone location features derived from delays. He firstly tried the combination of the location features with MFCCs using a GMM classifier. When the GMM posterior probabilities were fed into a Multi-layer Perceptron (MLP), he observed a change of the performance (compared to the original results with GMM) towards higher detection precision, but lower recall (for overlap detection evaluation metrics see Section 3.7).

In the proposal of Yamamoto et al. [108], a spatial correlation matrix is calculated from a microphone array input. The eigenvalue distribution of this spatial correlation matrix reflects information on the number and relative power of sound sources. In an experiment on one meeting recording, overlapping speech could be detected by applying support vector regression to the set of input eigenvalues.

Even though Képesi et al. [109] did not aim at detecting overlapping speech in the first place, their Position-Pitch (PoPi) extraction algorithm makes it theoretically possible. This method decomposes real two-channel recordings into a 2D PoPi plane, where all acoustic sources are represented by their fundamental frequency and their position. Pitch and DOA candidates representing position are jointly extracted from multiple correlation peaks.

Single distant microphone methods

Boakye and Trueba-Hornero focused in their respective theses [110, 111] and publications [112, 113] on developing an overlap detection system for monaural recordings. The detection framework was relying on an eHMM segmenter, which segmented the signal into overlap, non-overlap, and non-speech class. Their approach was also inspirational for this thesis. A number of features were tested and assessed according to their suitability to discriminate overlapping speech, e. g., Diarization Posterior Entropy (DPE), Linear Predictive Coding Residual Energy (LPCRE), Modulation Spectrogram (MSG), Harmonic Energy Ratio (HER), and Spectral Flatness (SF).

Entropy, in the information theory, is a measure of uncertainty, or information content. The idea behind DPE was explained by the authors in [112] as follows. In the diarization process the system produces likelihoods which describe the expectation that particular frame or segment belongs to every cluster. Intuitively, in a non-overlap segment there should be a clear "winner" and the rest will have significantly lower scores. Thus, score entropy will be low. In a multiple-speaker region the scores will be more equally distributed, resulting in a higher entropy. DPE for frame $\mathbf{y}$ can be expressed as

$$H_{DP} = \sum_{k} p(C_k|\mathbf{y}) \log \frac{1}{p(C_k|\mathbf{y})}, \tag{2.12}$$

where the posterior probability of a particular cluster C_k, $p(C_k|\mathbf{y})$, is computed applying the Bayes' rule as

$$p(C_k|\mathbf{y}) = \frac{p(\mathbf{y}|C_k)p(C_k)}{\sum_k p(\mathbf{y}|C_k)p(C_k)}. \tag{2.13}$$

The prior cluster probability $p(C_k)$ is estimated according to the amount of assigned speaker time in the diarization output. Entropy analysis in the context of overlap detection, but in the time domain in this case, was recently also investigated by [114].

SF metric is defined as a log ratio of geometric and arithmetic mean of spectral magnitudes. Spectrum of speech signal turns out to be less flat in segments with more frequencies (overlapping speech) than in segments with few or no frequencies. LPCRE feature, inspired by [115], is based on the fact that Linear Predictive Coding (LPC) can model well one voice, but suffers if more voices are present. In such cases more energy is expected to be present in the residual signal [112]. Both SF and LPCRE are also discussed in Section 3.2.

Further overlap detection strategies [116, 117], integrated into speaker diarization systems are discussed in the following section.

2.4 OVERLAPPING SPEECH IN SPEAKER DIARIZATION

Simultaneous speech poses a difficult challenge for spoken language applications

As already stated in the previous section, several works consider overlapping speech as a challenging problem for automatic human language technologies, including speaker diarization [33, 79]. The presence of overlap can be especially strong in meeting environment where the discourse is less structured and more spontaneous (compared to broadcast news, for instance). Nevertheless, Shriberg et al. [79] observed that its amount is not necessarily dependent on the number of people being present, and that two people in a telephone conversation can produce significant overlap too. In the following, the focus is on techniques relying on distant channel data.

Previously, overlapping speech was discarded from speaker diarization evaluations of meeting data. In NIST RT '06s[2] evaluation, overlap was included in the main metric for the first time. The detection techniques researched by the participating labs did not bring any success in decreasing the overall error [7, 116]. A related error analysis published in [6] reports that approximately 22% of the error could be accounted to overlapping speech. Otterson and Ostendorf [5] suggest that in a conventional speaker diarization system overlaps cause errors in at least two ways:

Two sources of error by overlaps in conventional speaker diarization

2 The Rich Transcription (RT) evaluation series, organized by the National Institute of Standards and Technology (NIST), began in 2002 and promotes advances in the state of the art in several automatic speech processing technologies to produce more readable and useful transcriptions. One of the research tasks defined under Metadata Extraction is speaker diarization [118].

1. Clustering assigns segments to only one speaker, which means that during an overlap, the speech of the interfering speaker(s) will be denoted as missed.

2. Speaker models can be corrupted when overlapping speech is included in their training data, possibly causing less precise clustering (and higher speaker error).

Both increase the Diarization Error Rate (DER). This fact consequently leads to two open questions:

1. Would the impact of overlaps be decreased if they would be detected and *excluded* before clustering?

2. And, would the *assignment of a second label* for overlap segments lead to an improvement in the diarization score?

Dealing with the overlapping speech issue can be considered in two levels. The first is the detection of segments where the overlaps occur (see Section 2.3.2). Then, given the knowledge of the overlap locations, the second is the handling of such segments in order to improve diarization (e. g., by assigning more than one speaker label). An interesting assessment of the performance gain by handling overlap regions, assuming oracle overlap detection, is given in [5].

A straightforward option would be to pre-process detected overlaps with a source separation algorithm, and work with the separated signals individually. The potential advantage would be that speaker-specific characteristics could be isolated and employed in the clustering. However, this approach faces robustness issues of separation techniques with real overlapping speech (in meetings, for instance), and also seems rather complicated. To the knowledge of the author no comparable strategy was proposed in the literature in relation to speaker diarization.

Explicit modeling of concurrent speaker combinations

A completely different approach is to model overlapping combinations of previously detected speakers. Actually, in this way the overlap detection and the identification of which speakers are involved are accomplished simultaneously. Such approach was proposed by Leeuwen and Huijbregts in [116]. Their system starts with a standard diarization segmentation. Once finished, a new HMM is constructed with single-speaker states from detected individual clusters and also overlap states for every speaker pair. The overlap states are trained with the speech from both clusters of a particular pair. In addition, the HMM topology is altered so that transitions between the single and the overlap states are only allowed if the overlap state includes the speaker of the single state. Meeting data is then resegmented with the extended model. The authors reported that even though overlap was detected with this approach, it did not lead to a reduction of the diarization error on NIST RT data.

A global segmentation in terms of all possible single- and multiple-speaker classes was explored in [40] (already discussed in Section 2.3.2). Single-speaker models were used to generate dual-speaker models and were jointly integrated into a GMM/HMM framework. Limiting factors of this work were that the number of speakers had to be known beforehand and such strategy suffers from a huge number of overlap classes.

When a known number of speakers is confined to certain physical regions, it is possible to produce parallel individual speaker segmentations for every speaker based on his location instead of a global segmentation [106]. In this way, all overlapping speaker combinations are handled implicitly.

Having identified accurate locations of overlapping speech, it is possible to assist diarization without any further extensive processing. The algorithm presented by Boakye et al. [112, 113] detected overlapping speech with an HMM-based system utilizing various features (see Section 2.3.2) on single distant-channel recordings from the AMI corpus. In a following step, the detected overlap segments were excluded from speaker clustering process of the diarization system in order to obtain purer clusters.

Posterior assignment strategies

With the goal to assign correct speaker/cluster labels to the diarization output, Trueba-Hornero [111] explored four posterior labeling strategies for two-speaker overlap situations:

- *Random Selection* — the second label is chosen on a random basis from the remaining speaker candidates. This strategy may serve as a baseline for assessing other more complex schemes.

- *Most Talkative Speaker* — this strategy assigns the overlapping label to the most talkative speaker in the diarization output. In case that the most talkative speaker has already been the choice of the diarization system, the second most talkative speaker is selected. This strategy was also applied for an oracle experiment in [119].

- *Overlap Patterns* — the assignment of the missing speaker label is based on the analysis of the identified overlapping speech pattern (i. e., flow, interruption, floor-grabbing pattern). This strategy is very similar to the *Nearest Neighbor* scheme, where the second label is set according to the nearest neighboring speaker, as in [5, 117].

- *Diarization Posteriors* — speaker labels for an overlap region are assigned according to the two highest posterior probability scores, produced by a diarization system, in that particular region [112, 113].

The most successful strategy according to [111] was the one based on diarization posterior probabilities.

Experiments involving the exclusion of overlapping speech from diarization process can also be found in [5, 114, 117]. Huijbregts et al. [117] assume in their algorithm that speaker overlap is likely to occur around speaker-turn points. Accordingly, an ad hoc overlap model is trained in one diarization pass and the overlap GMM is added to the original HMM. The objective is to pool all the suspected overlapping speech in Viterbi decoding into one cluster, and not contaminating the others during a second diarization pass. At the same time, overlapping speech is also detected for posterior assignment of extra speaker labels. The application of this technique improved results on NIST RT data.

3 DETECTION OF OVERLAPPING SPEECH

The detection of overlapping speech concerns with the identification of speech segments with more simultaneously active speakers in a meeting recording. In our approach we do not determine the exact number of involved speakers, but rather focus on differentiating between single-speaker speech and overlapping speech. This chapter begins with the introduction of the general concept of our overlap detection system and its relation towards subsequent speaker diarization. Then, features assuming to convey discriminating information on speaker overlap are discussed. Before proposing the novel spatial-based and long-term prosody-based features for this task, baseline short-term parameters are presented. Baseline features are derived from speech spectrum or temporal course of the signal. Finally, the modeling and the decoding framework are described, followed by the definition of evaluation metrics for the detection of overlapping speech.

3.1 OVERLAP DETECTION SYSTEM ARCHITECTURE

Overlap detection process consists of the usual stages which can be found in almost every pattern recognition system: feature extraction and decoding/classification. The system diagram is given in Figure 6. The input is formed by a number of distant microphones in a microphone array. When only one channel is needed, e. g., for baseline or prosodic features, normally the first channel from the first array is used.

Features are categorized into three groups, spectral- and temporal-based parameters (Section 3.2), prosodic features (Section 3.5), and cross-correlation-based spatial parameters (Section 3.3). Since spatial parameters are produced for every microphone pair, the spatial feature extraction is coupled with a so-called microphone data fusion block for dimensionality reduction and unification purposes. If necessary, the different feature streams are synchronized after feature extraction in order to form feature vectors at a common frame rate.

HMM-based decoder relying on multiple feature streams

The detection of speaker overlap is achieved by Viterbi decoding of given feature streams. The system considers non-speech (e. g., silence, noise), single-speaker speech, and overlapping speech class to classify the signal and produce an output hypothesis. The HMM-based decoding framework will be explained in Section 3.6 more in detail.

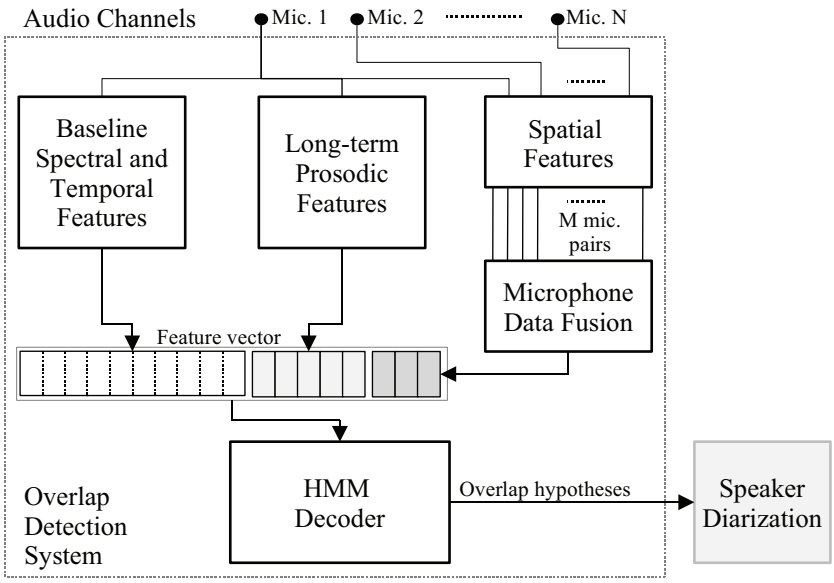


Figure 6: Overlap detection system diagram.

Finally, the hypothesized start times and durations of overlapping speech are provided as an input to the diarization system. The system works offline, however, the design is potentially open for live processing as well.

3.2 BASELINE SPECTRAL AND TEMPORAL FEATURES

Baseline features for overlap detection involve short-term parameters, derived from the speech spectrum or the temporal course of the signal, which were previously proposed in the literature. In the following, their definition and a brief description is given. A subset of these parameters is later selected for the construction of a baseline system (see Section 6.1).

Mel Frequency Cepstral Coefficients

Mel cepstrum

Mel Frequency Cepstral Coefficients (MFCCs) are a representation of the short-term power spectrum of a sound, based on a linear cosine transform of log power spectrum on a nonlinear frequency scale. The power spectrum is obtained by applying triangular-shaped filter bank to the spectral magnitudes of the signal. The frequency bands of the filter bank are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

MFCCs are a highly popular feature kind for several speech processing tasks and are well suited as baseline parameters [102]. In addition to phonetic information, these features also capture physiology characteristics of the talker and thus could also provide information whether multiple speakers are speaking.

In our experiments, MFCCs were extracted every 10 ms over 30 ms Hamming windows and normalized by Cepstral Mean Subtraction (CMS). We use static MFCCs together with their first-order derivatives (deltas).

Linear Predictive Coding Residual Energy

LPC is one of the most effective methods for the analysis of acoustic signals. In this model the speech is produced as the output of a linear, slowly time-varying system excited by either a quasi-periodic glottal impulses (voiced speech) or random noise (unvoiced speech). The linear system representing the vocal tract is described by an all-pole system function that can be expressed in the form

$$H(z) = \frac{Y(z)}{E(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{3.1}$$

where $E(z)$ is the excitation input, $Y(z)$ is the speech output of the model, the excitation gain is denoted as G, and p is the order of the model. The filter coefficients $\{a_k\}$ encode the formants, i. e., the vocal tract resonances.

In linear prediction analysis the set of predictor coefficients $\{a_k\}$ is efficiently computed directly from the speech signal. The output of a linear predictor is

$$\hat{y}[n] = \sum_{k=1}^{p} a_k y[n-k], \tag{3.2}$$

and the prediction error $d[n] = y[n] - \hat{y}[n]$, also termed residual, is defined as the amount by which the model fails to predict the original signal. Residual signal can be obtained by filtering $Y(z)$ with a prediction error filter $A(z)$, which will be inverse to the system filter $H(z) = G/A(z)$.

Linear prediction error increases with overlapping speakers

It is assumed that LPC of a reasonably chosen order can model the spectrum of a single speaker quite well, but will fail for a region with multiple speakers [115, 112]. Consequently, more energy is left in the residual signal (prediction error) in the later case. In our system, Linear Predictive Coding Residual Energy (LPCRE) of a 12^{th}-order LPC was computed over a 25 ms window. In the example illustrated in Figure 7, the LPCRE features corresponding to overlapping speech exhibit higher values compared to single-speaker speech segments.

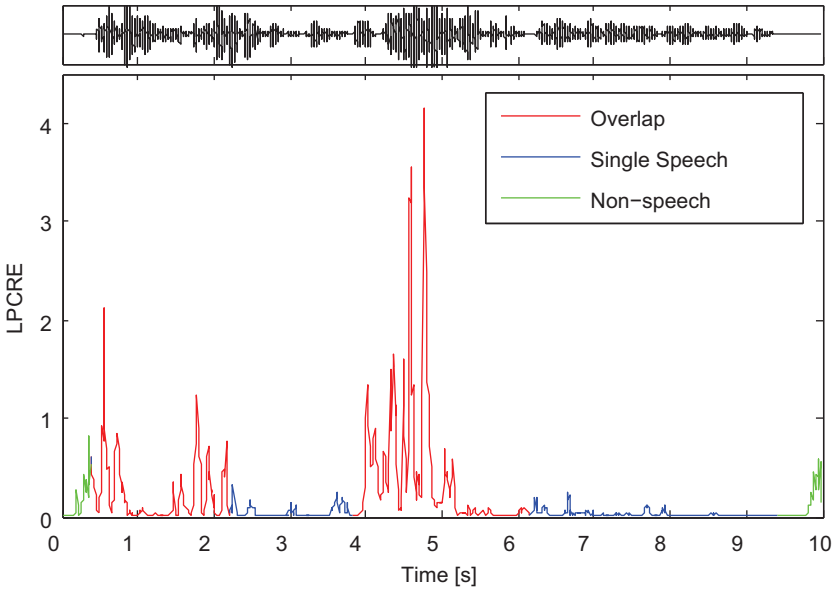


Figure 7: LPC residual energy of a sample audio signal (top) containing simultaneous speech. Sample taken from the recording IS1004d (AMI corpus) at approx. 00:16:32 h.

Spectral Flatness

Another spectral-based feature is the Spectral Flatness (SF), which was originally applied for discrimination between speech and non-speech [101], but can eventually convey information about the number of speakers speaking as well [113]. It is derived from the signal spectrum and yields high values for signals with power equally distributed across all frequency bands, such as noise. SF is defined as the ratio between the geometric and the arithmetic mean of N spectral magnitudes:

$$M_{SF} = 10 \log_{10} \frac{\sqrt[N]{\prod_{i=0}^{N-1} |X(i)|}}{\sum_{i=0}^{N-1} |X(i)|}. \tag{3.3}$$

Higher number of active frequency bands makes the spectrum flatter

The frequency domain structure of overlapping voiced speech can differ from that of single-speaker speech. Harmonic frequencies of overlapping speakers can introduce more elevated energy bands resulting in a flatter speech spectrum in comparison with single-speaker situation. This effect, however, is very much dependent on the pitch differences and relative energy concentrations [110].

In our experiments, SF was extracted over 30 ms Hamming windows at a rate of 10 ms considering the first $N = 100$ spectral lines of a 512-point FFT. Spectral flatness values of an audio signal containing overlap segments are demonstrated in Figure 8 and indicate a high variability of this feature. Even though both speaker overlap segments have, in

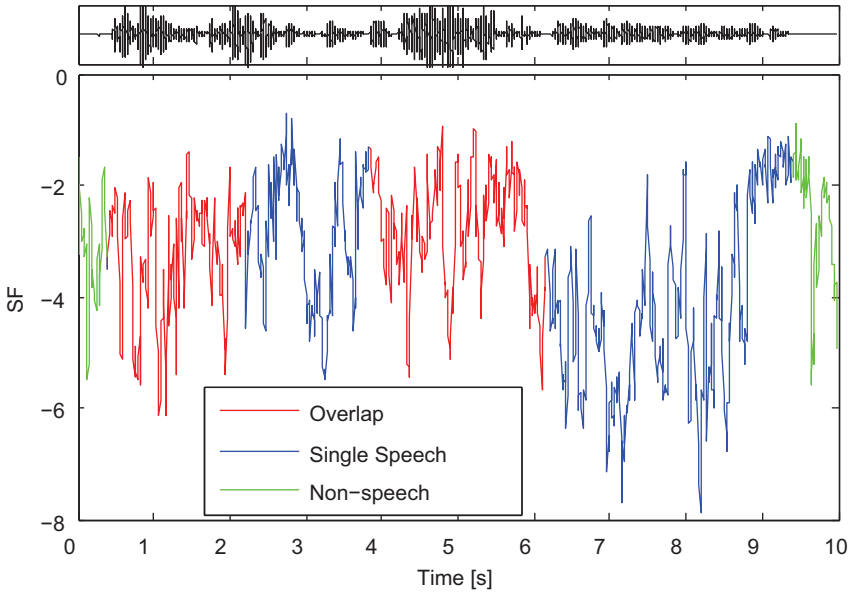


Figure 8: Spectral flatness of a sample audio signal (top) containing simultaneous speech. Sample taken from the recording IS1004d (AMI corpus) at approx. 00:16:32 h.

general, relatively high SF, the single-speaker segments in some frames show even higher flatness.

Pitch Prediction Feature

Pitch Prediction Feature (PPF) was developed to discriminate between one- and two-speaker speech [102]. A short summary of the computation process is as follows. In the first stage, an LPC representation is computed and an LP residual is obtained. The LP residual is furthermore smoothed with a Gaussian-shaped filter. After a threshold-based extraction of pitch peaks from the smoothed residual, the PPF measure is computed as the standard deviation of the distances between successive peaks. For single-speaker segments, a regular sequence of peaks will occur in the LP residual (corresponding to glottal closures), resulting in a low PPF value [97].

We extracted the PPF every 10 ms over 30 ms windows and an LPC of 12^{th}-order was used in the first computation stage. Values in unvoiced and silence regions were substituted with mean PPF from voiced regions. An example of PPF is illustrated in Figure 9.

Modulation Spectrogram

An alternative representation of speech signal with emphasis on temporal characteristics is the Modulation Spectrogram (MSG). Originally

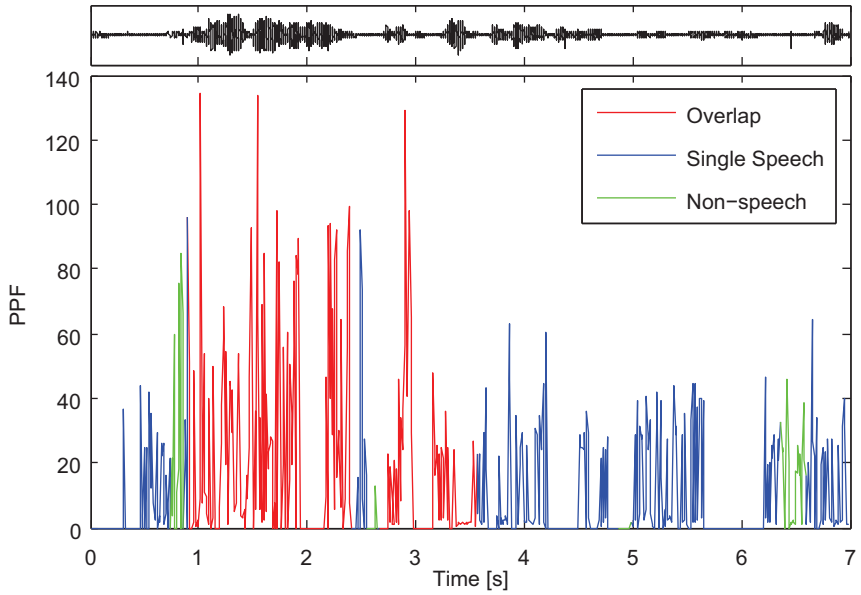


Figure 9: Pitch prediction feature of a sample audio signal (top) containing simultaneous speech. Sample taken from the recording IS1004d (AMI corpus) at approx. 00:21:58 h.

Modulation spectrogram emphasizes temporal evolution of signal spectrum

introduced in [120], Boakye et al. [113] suggested that this representation may act complementary to, for instance, MFCCs for the detection of overlapping speech.

An example of modulation spectrogram features is given in Figure 10. The parameters are computed as follows. After obtaining the FFT of a signal with 10 ms frame rate and 25 ms analysis window, the spectrogram is analyzed in 18 subbands according to Bark scale[1]. Square-root subband energies are computed for each frame. These are then filtered in time for each individual subband with two modulation filters: a low-pass 0–8 Hz filter, and a band-pass 8–16 Hz filter, to reflect temporal evolution of spectrogram related to the syllable rate in speech. The length of the filters is 0.21 s. Finally, the resulting 36 features are mean-variance normalized.

Voicedness Feature

Voicedness measure, which was used in combination with cepstral features for ASR in [121], was implemented based on Harmonic Product Spectrum (HPS) [122]. This method relies on the fact that the amplitude spectrum of voiced sounds show sharp peaks at integer multiples of the fundamental frequency. Harmonic product spectrum $P(n)$ is

1 Bark scale corresponds to the first 24 critical bands of hearing.

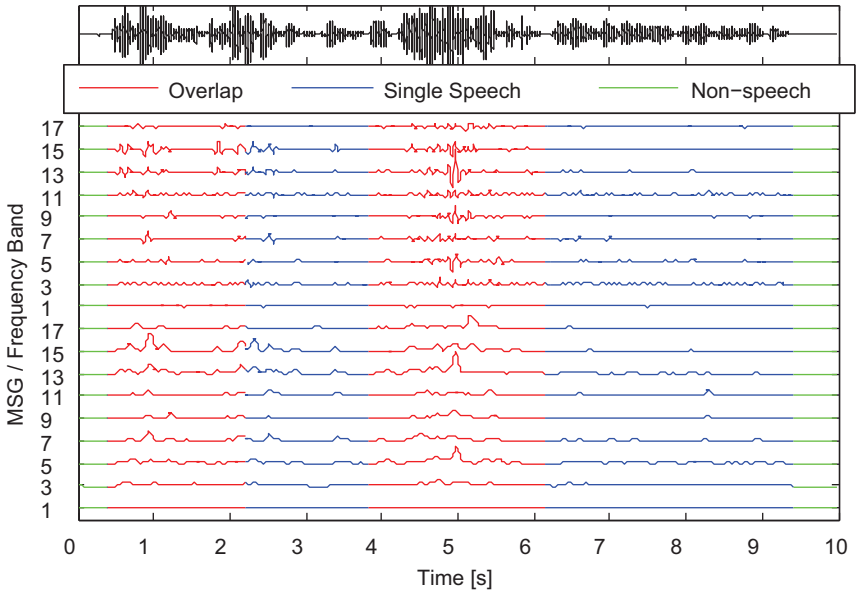


Figure 10: Modulation spectrogram features of a sample audio signal (top) containing simultaneous speech. Two modulation filters are applied to 18 subbands: 0–8 Hz filter (lower half) and 8–16 Hz filter (upper half). For better comprehensibility only every other subband is depicted. Sample taken from the recording IS1004d (AMI corpus) at approx. 00:16:32 h.

the product of R frequency-compressed replicas of the amplitude spectrum. Mathematically it can be expressed as

$$P(n) = \sqrt[R]{\prod_{r=1}^{R} |X(e^{in\Delta\omega r})|}, \tag{3.4}$$

where $\Delta\omega$ is the resolution of the FFT. The motivation is that for periodic signals the product spectrum should give high peaks at the pitch value and its near harmonics, and close-to-zero values otherwise. Unlike voiced speech, unvoiced frames do not have a clear peak structure and their HPS is typically flat.

Harmonic product spectrum peak-neighborhood ratio

The voicedness measure reflects how voiced a particular frame is. In [121], it is referred to as the height measure of the HPS peak, since it considers the peak amplitude:

$$v_h = \frac{P(n_{max})}{\sqrt[2W]{\prod_n P(n)}}, \tag{3.5}$$

where n_{max} is the position of the maximum amplitude, and n addresses the neighborhood of n_{max}, $n \in \langle n_{max} - W; n_{max} + W \rangle - n_{max}$. W is set as the half of the minimal distance between two harmonics, 40 Hz/$\Delta\omega$.

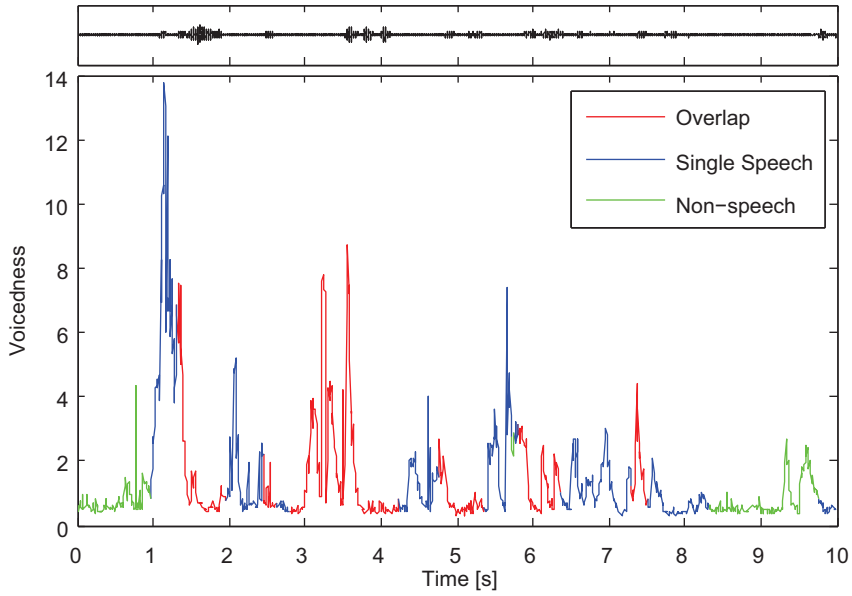


Figure 11: Voicedness feature derived from the harmonic product spectrum of a sample audio signal (top) containing simultaneous speech. Sample taken from the recording IS1004d (AMI corpus) at approx. 00:19:46 h.

We believe that HPS-derived voicedness measure may also be beneficial for the detection of speaker overlap, because the introduction of concurrent speaker harmonics will probably have an influence on the periodicity of voiced speech spectra of the first speaker. In such situation, the detected HPS peak will correspond to the pitch of the dominant speaker. The concept of this feature is very similar to the HER feature, mentioned in Section 2.3.2, which analyzes the energy distribution between harmonic and non-harmonic frequency bands.

Figure 11 shows an example of this feature on overlapping speech sample. In this case, the overlap segments exhibit in several instances relatively high voicedness values. The feature was computed with step size of 10 ms over 30 ms frames using a 2048-point FFT.

Zero-Crossing Rate

Zero-Crossing Rate (ZCR) is the rate at which the signal changes from positive to negative or back. It is commonly used in speech processing for various audio classification tasks including, for instance, voiced/unvoiced or speech/music discrimination.

This speech parameter is defined as follows,

$$M_{ZCR} = \sum_{i=1}^{N-1} 0.5\,|\operatorname{sign} x[i] - \operatorname{sign} x[i-1]|\,, \tag{3.6}$$

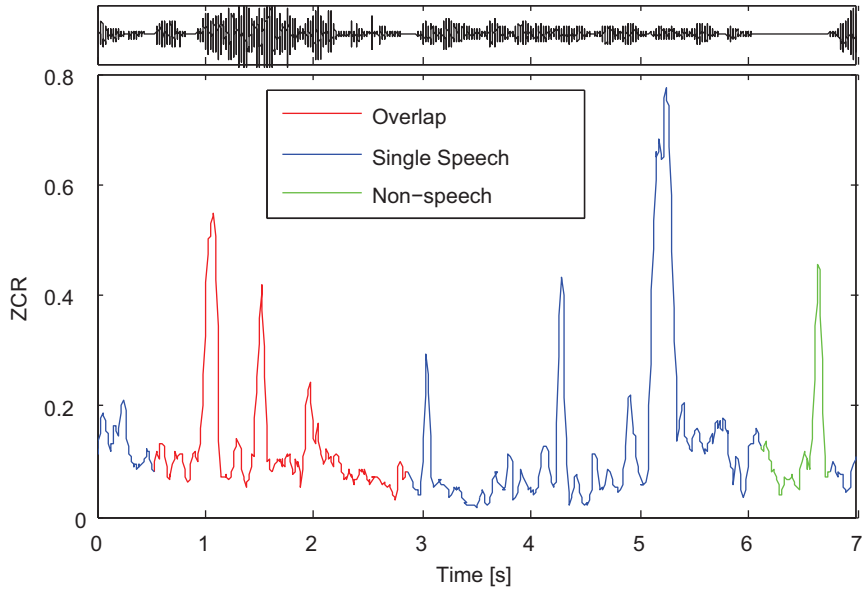


Figure 12: Zero-crossing rate of a sample audio signal (top) containing simultaneous speech. Sample taken from the recording IS1004d (AMI corpus) at approx. 00:16:35 h.

Temporal-based parameters

where x is an audio signal having N samples. In single-speaker voiced speech situation the values of ZCR are related to pitch periodicity and exhibit low values, whereas in unvoiced speech or background noise, the ZCR is high. The assumption regarding overlapping speech is that when two voiced signals are mixed, the resulting signal will then have due to different pitch periods increased periodicity, and thus higher ZCR. However, in the given speech sample in Figure 12, such assumed behavior cannot be verified. ZCR was computed over 50 ms frames every 10 ms.

Kurtosis

In statistics, kurtosis serves as a measure of how flat is the top of a symmetric distribution of a random variable. Distributions with longer tails and more acute peaks have positive kurtosis and are called *leptokurtic*. On the other hand, distributions with shorter tails and wider peaks use to have negative kurtosis and are called *platykurtic*. Given a random variable x, its kurtosis is defined as follows,

$$k_x = \frac{E(x-\mu)^4}{\sigma^4} - 3, \tag{3.7}$$

where μ and σ are the mean and standard deviation of x, respectively, and $E(\cdot)$ represents the expected value. The term -3 is introduced to make the kurtosis of Gaussian distribution equal to zero.

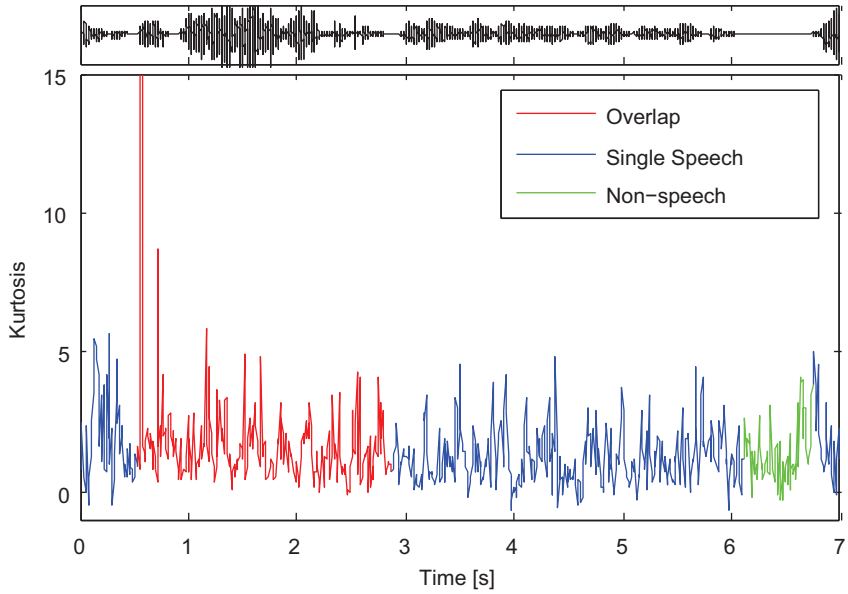


Figure 13: Kurtosis of a sample audio signal (top) containing simultaneous speech. Sample taken from the recording IS1004d (AMI corpus) at approx. 00:16:35 h.

The use of kurtosis in this context is inspired by Krishnamachari et al. [99] and later it was experimented by Wrigley et al. [97]. Since the sum of several random distributions has lower kurtosis than individual distributions, it was hypothesized that the kurtosis of overlapping speech is generally also lower than for isolated speech utterances.

In this work, kurtosis feature is extracted over a 20 ms Hamming window every 10 ms. Example values for an overlapping speech sample are illustrated in Figure 13, but unfortunately it is not clear from this illustration if overlap segments are less leptocurtic than single-speaker segments.

Root-Mean-Squared Energy

The basic idea behind the application of Root-Mean-Squared Energy (RMSE) for speaker overlap detection is that multiple concurrent speakers can produce higher energy than a single speaker. Furthermore, people also tend to start talking louder when competing simultaneously in an argument. The assumption is obviously very simple and does not take into account states of elevated emotions, such as laughter or anger.

RMSE for a framed audio signal x is defined as

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} x[i]^2}. \tag{3.8}$$

Table 1: F_{ratio} and KL2 divergence for a subset of candidate baseline features. Values are calculated according to AMI development data.

FEATURE	F_{RATIO}	KL2
Root-Mean-Squared Energy	0.1432	0.2378
Modulation Spectrogram (12^{th} prm.)	0.1356	0.3848
MFCC (7^{th} coef.)	0.0534	0.0831
Voicedness	0.0341	0.0978
Pitch Prediction Feature	0.0160	0.0466
LPC Residual Energy	0.0134	0.1220
Zero-Crossing Rate	0.0041	0.0179
Spectral Flatness	0.0023	0.0061
Kurtosis	0.0004	0.0210

For better comparability between different recording sessions, the signals were normalized to unit power before extracting the RMSE feature. The frame length used in this work is 20 ms.

Comparison of Baseline Feature Candidates

Feature discriminability can be assessed by means of Fisher's ratio or Kullback-Leibler divergence

There are several options to assess the potential discriminability of a particular feature. For the case of baseline overlap detection features we consider two metrics: Fisher's ratio (F_{ratio}) and the symmetric Kullback-Leibler divergence (KL2).

Fisher's ratio is a measure for linear discriminating power of some variable, given as

$$F_{ratio} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}, \tag{3.9}$$

with μ_1 and μ_2 being the means of class 1 and class 2, and σ_1^2 and σ_2^2 being the variances. The KL2 divergence, given in (2.9), measures the dissimilarity between two distributions of random variables. Though informative, these measures only give a preliminary idea of a parameter suitability for a given task and do not speak about the actual impact on system performance. Furthermore, some features can have good isolated discrimination properties but do not work so well in combination with others, since the features can be correlated.

Table 1 presents F_{ratio} and KL2 divergence for a subset of baseline features. For illustrational reasons we selected from MFCCs and MSG features only the 7^{th} and 12^{th} parameter, respectively. According to the AMI development data (refer to Section 5.1 for more details) the highest F_{ratio} exhibits the RMSE and the highest KL2 divergence the

MSG feature. The two discrimination metrics do not always follow the same trend, for instance LPCRE shows relatively high KL2, but only a moderate F_{ratio}.

Baseline feature histograms

The corresponding histograms for the subset of candidate baseline features are given in Figure 14. The distribution of SF pictured in Figure 14 (b) shows that the histograms of both classes are very much overlapping and explain the low scores in Table 1. A slight shift of the overlap distribution towards higher SF values is in agreement with the assumption that spectra of simultaneous speech are more flat. Figure 14 (f) demonstrates that overlapping speech also exhibits higher RMSE.

The majority of voicedness values in Figure 14 (e) are concentrated in the interval 0–2, bigger values clearly indicate voiced speech frames. It can be noted that relatively lower number of speaker overlap frames is unvoiced. This is not unexpected, since the mixture of independent voiced and unvoiced speech has rather a voiced speech appearance.

Surprisingly, the kurtosis distribution of simultaneous speech in Figure 14 (h) is shifted towards higher values in comparison with the single-speaker distribution. This observation is in contrast with the previously stated hypothesis that overlapping speech has, in general, lower kurtosis.

3.3 NOVEL SPATIAL-BASED FEATURES

Microphone arrays provide the ability to discriminate between sounds based on their source location. The application of features related to spatial location of speakers was proposed for speaker diarization in various works, such as [123]. Spatial sampling of the acoustic field can serve not only for meeting segmentation, but also makes it possible to detect more active speakers [40, 106]. Other publications in the context of speaker overlap detection relying on microphone arrays, and spatial information in one way or another, include [107, 108, 109].

Képesi et al. [109], for instance, presented a multi-source tracking algorithm based on the decomposition of two-channel cross-correlationinto a 2D Position-Pitch space. Concurrent speakers can be separated in such 2D representation when common periodicities and related DOA values are extracted by a specific sampling process applied to the cross-correlation function.

In this thesis, we elaborate on the idea of applying spatial information for the detection of simultaneous speech and propose a set of parameters derived from the cross-correlation between two distant microphones. Furthermore, three techniques for the fusion of information from different microphone pairs are investigated.

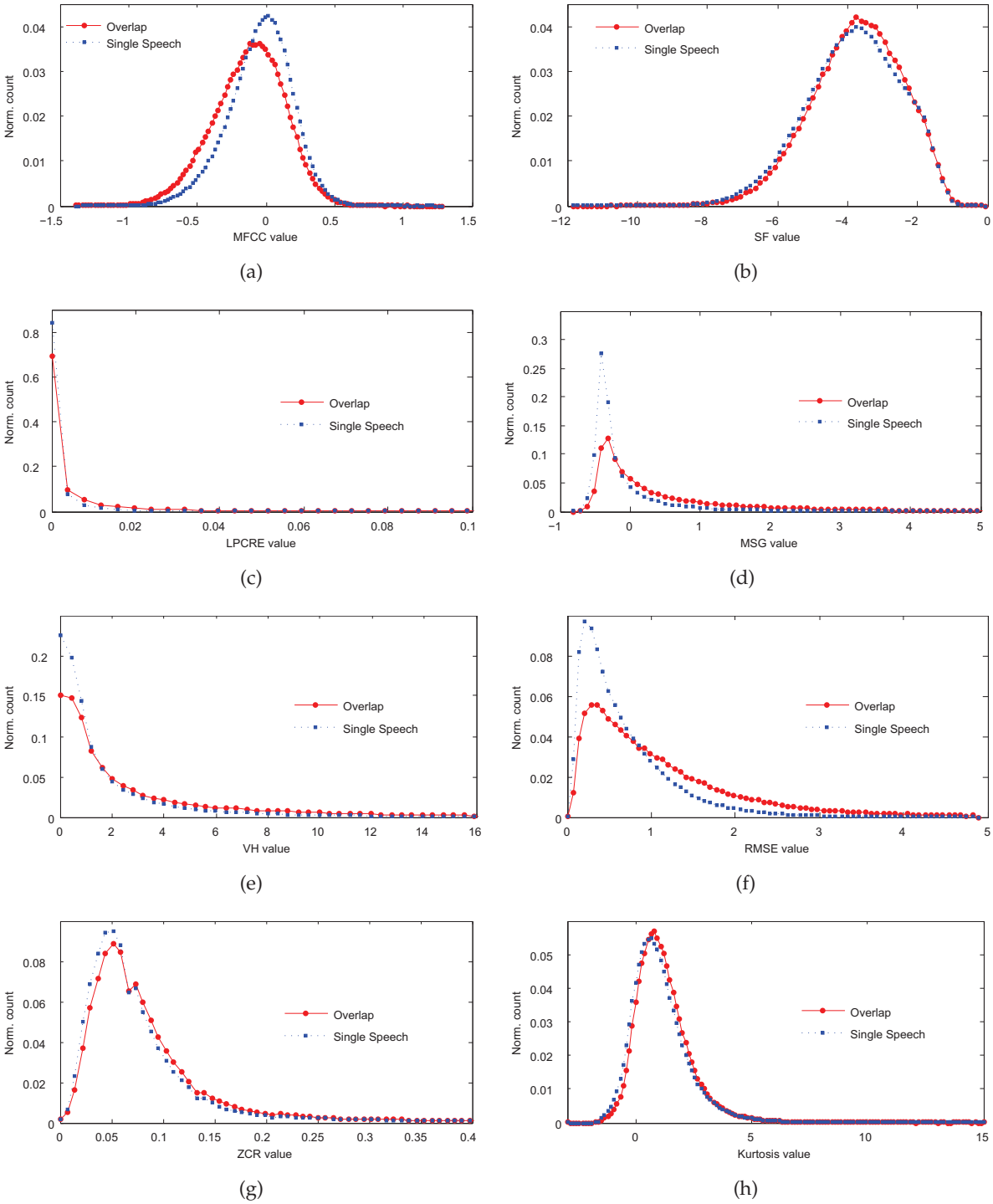


Figure 14: Histograms of baseline feature candidates: (a) 7^{th} MFCC, (b) SF, (c) LPCRE, (d) 12^{th} MSG parameter, (e) Voicedness, (f) RMSE, (g) ZCR, and (h) Kurtosis. Histograms are computed on AMI development data.

3.3.1 *Generalized Cross-Correlation*

Cross-correlation function

The cross-correlation function is well-known as a measure of the similarity between signals for any given time displacement and ideally its maximum lies in correspondence to the delay between the pair of signals [124]. A commonly used technique to estimate the time delay between two acoustic signals that performs robustly in reverberant environments is the Generalized Cross-Correlation with Phase Transform Weighting (GCC-PHAT) [125, 126]. Although it is a general purpose technique and not fully adapted to speech, it has turned out to be the most successful state-of-the-art approach to speaker localization and it has been employed by some researchers in the field of speaker diarization, including [127, 128]. For a pair of microphones m and n, the GCC-PHAT can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectrum , $G_{mn}(f)$, as follows,

$$R_{mn}(\tau) = \int_{-\infty}^{\infty} \frac{G_{mn}(f)}{|G_{mn}(f)|} e^{i2\pi f\tau} df, \tag{3.10}$$

and the Time Delay of Arrival (TDOA) is as follows,

$$\hat{\tau}_{mn} = \arg\max_{\tau} R_{mn}(\tau). \tag{3.11}$$

The GCC-PHAT function exhibits a prominent peak at the elapsed time corresponding to the dominant sound source in the room, minimizing the peaks of the non-dominant sources and reverberation at the same time. The value of the GCC-PHAT peak provides a measure of the coherence between signals independently of the microphone gains or the signal power, and varies with the distance between microphones, the distance between the acoustic source and the microphone pair, and with the environmental noise and reverberation conditions.

3.3.2 *Spatial Coherence, Dispersion, and delta TDOA*

In situations dealing with multiple, possibly moving, concurrent speakers, it was observed that the time delay estimates produced by the GCC-PHAT jump from one speaker to another at a very high rate as one source dominates due to the non-stationarity of the voice. The maximum value of the cross-correlation sequence is also lower than in the single speaker situation, since multiple speakers introduce random peaks, which attenuate the main peak.

Based on these observations we proposed several cross-correlation-based spatial features for every microphone pair that provide some degree of information on speaker overlaps [129, 130].

Coherence — cross-correlation peak value

An easily observable feature is the *coherence value*, defined in (3.12). This is the principal peak value of the GCC-PHAT, and in ideal conditions should be high for single-source situations, while the presence

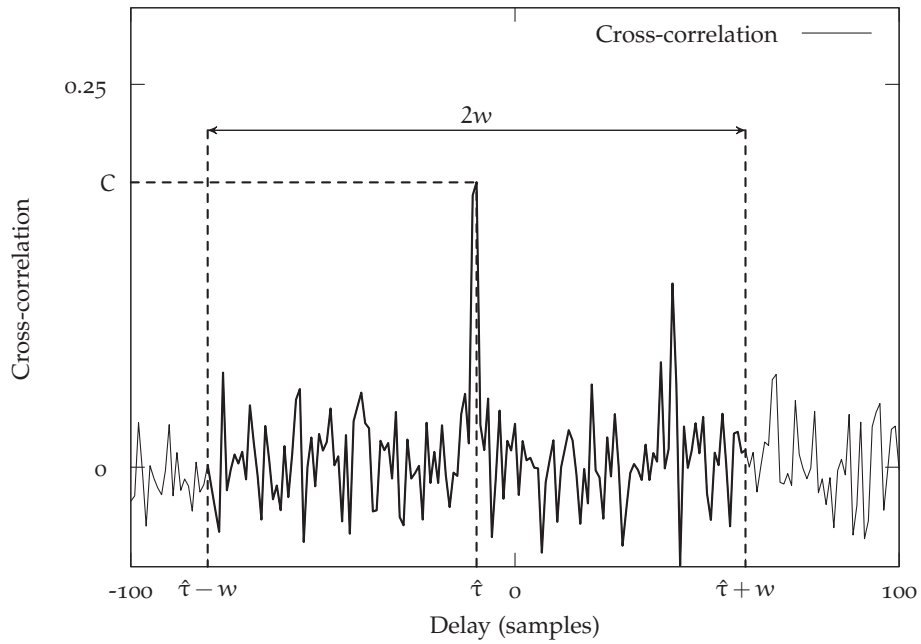


Figure 15: Example of the cross-correlation between a pair of microphones involving two concurrent speakers. The value of the main peak is the coherence feature C, and its time displacement $\hat{\tau}$ corresponds to the TDOA. The ratio between C^2 and the quadratic sum of the values in bold under the window is the dispersion feature D.

of noise, reverberation, and concurrent acoustic sources attenuate this value.

$$C_{mn} = \max(R_{mn}(\tau)) \tag{3.12}$$

Derived from the coherence value, we are also proposing to extract the coherence *dispersion ratio*, as follows,

$$D_{mn} = \frac{C_{mn}^2}{\sum_{t=-w_{mn}}^{w_{mn}} R_{mn}^2(t+\hat{\tau}_{mn})}. \tag{3.13}$$

Dispersion — cross-correlation peak-neighborhood ratio

This value is computed as the ratio between the square of the main peak value and the square quadratic sum of the cross-correlation values under a time delay window w_{mn}. The size of the window w_{mn} varies for different microphone pairs and it is set to the TDOA standard deviation of each pair. In this way, the dispersion ratio measures the relation between the energy of the main peak and the energy that is scattered in its neighborhood. Similar to the coherence feature (3.12), the dispersion ratio is close to 1 in the case of a single speaker and ideal conditions, while it has a lower value in reverberant conditions or concurrent acoustic sources situations.

Delta TDOA

Finally, the *delta of TDOA* obtained by (3.11) for every microphone pair also carries information on overlaps. The derivative of the TDOA

is high in situations where the speaker is moving, multiple non-concurrent speakers change turns at talk, or multiple speakers talk simultaneously.

An illustration of the cross-correlation between a pair of microphones and the proposed spatial features can be seen in Figure 15. The GCC-PHAT was estimated with step advances of 64 ms using a 1024-point FFT for each of the $\binom{m}{2}$ microphone pairs, with m being the number of available distant channels.

3.4 MICROPHONE DATA FUSION

High and variable dimensionality of spatial features

One of the main issues that arise is the high dimensionality of spatial feature vectors. For instance, a recording with 12 available microphone channels yields 66 pairs, and thus 198 features. Furthermore, the number of microphones can differ from site to site, making it difficult to train a general model. In the following, we discuss two transformation- and one neural-network-based approaches for the dimensionality reduction and normalization of spatial features.

3.4.1 *Principal Component Analysis Fusion*

First strategy is based on the application of the Principal Component Analysis (PCA), which is a useful statistical technique performing dimensionality reduction while preserving as much variability in the high-dimensional space as possible. It transforms the original feature space into a new coordinate system with the greatest variance lying on the first component. Otterson, for instance, used PCA for the reduction of spatial parameters for diarization in [131].

PCA, also known as Karhunen-Loève transformation

PCA is conceptually quite simple. Let $\mathbf{X} = \{\mathbf{x}_i\}$ be a data set formed by vectors $\mathbf{x}_i \in \mathbb{R}^n$, with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Next, the eigenvalues and eigenvectors are computed, and sorted according to decreasing eigenvalue. The eigenvalue equation is given by

$$\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}, \tag{3.14}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ lying on the diagonal and $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ is a $n \times n$ matrix consisting of eigenvectors, which are also called *principal components*. Eigenvectors are uncorrelated among each other. For more details on PCA, refer to one of [15, 11].

The transformed feature vectors are obtained as

$$\mathbf{y}_i = \mathbf{U}^T(\mathbf{x}_i - \boldsymbol{\mu}). \tag{3.15}$$

For dimensionality reduction only the first k eigenvectors with highest eigenvalues are picked, where $k < n$.

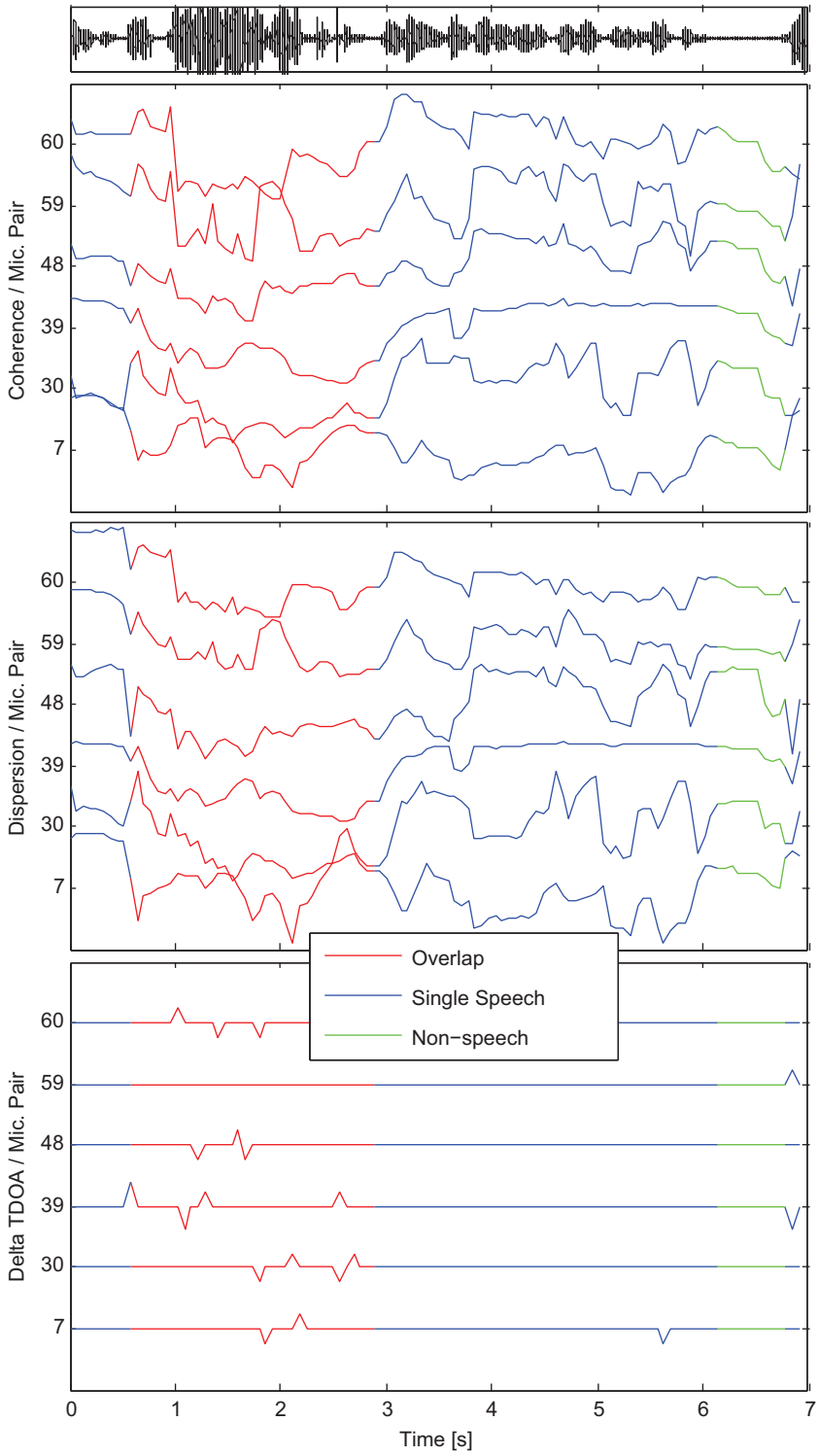


Figure 16: Spatial coherence, dispersion ratio, and delta TDOA for six randomly chosen microphone pairs of an audio signal (top) containing simultaneous speech. Sample taken from the recording IS1004d (AMI corpus) at approx. 00:16:35 h.

In this thesis a sequential PCA implementation [132], originally introduced in [133], is used. A transformation matrix is estimated for every discussed spatial feature kind and for each recording site, and then only the first principal component $\mathbf{u}_1$ is used [129, 130]. This practice is motivated by two reasons. First, the data across microphone pairs is correlated (see Figure 16) and usually the first eigenvalue is much more higher than the rest. The second motivation of using only one projection vector is to limit possible ambiguities when using data from various meeting rooms, i. e., which parameter in room A corresponds to which parameter in room B, etc. The final fused spatial feature vector $\mathbf{y}_{spat}$ for a given frame and a particular site can be expressed as follows,

$$\mathbf{y}_{spat} = \left[\mathbf{u}_{C\,1}^{T}\,\mathbf{x}_C,\ \ \mathbf{u}_{D\,1}^{T}\,\mathbf{x}_D,\ \ \mathbf{u}_{dT\,1}^{T}\,\mathbf{x}_{dT}\right], \tag{3.16}$$

where $\mathbf{u}_{C\,1}$, $\mathbf{u}_{D\,1}$, and $\mathbf{u}_{dT\,1}$ are first principal components for spatial coherence, dispersion ratio, and delta TDOA, respectively. The corresponding microphone-pair parameter vectors $\mathbf{x}_C$, $\mathbf{x}_D$, and $\mathbf{x}_{dT}$ are already assumed to have subtracted means.

3.4.2 *Linear Discriminant Analysis Fusion*

Linear Discriminant Analysis (LDA), in contrast to PCA, explicitly attempts to model the difference between the classes of data. However, when the discriminatory information is not in the mean but rather in the variance of the data, LDA is not a suitable option.

LDA is related to R. Fisher's linear discriminant (1936)

Let us consider a data set of N independent feature vectors $\{\mathbf{x}_i\}$ where each of the vectors $\mathbf{x}_i \in \mathbb{R}^n$ belongs to only one class $j \in \{1, \ldots, J\}$. The objective of LDA is to find a linear projection from n-dimensional space onto $(J-1)$ dimensions, $f : \mathbb{R}^n \to \mathbb{R}^{(J-1)}$, $\mathbf{y} = f(\mathbf{x})$.

Each class j is characterized by its own mean $\boldsymbol{\mu}_j$, covariance $\boldsymbol{\Sigma}_j$, and sample count N_j satisfying $\sum_{j=1}^{J} N_j = N$. The class information is represented by two scatter matrices $\mathbf{S}_W$ and $\mathbf{S}_B$ called *within-class* scatter and *between-class* scatter, respectively:

$$\mathbf{S}_W = \frac{1}{N}\sum_{j=1}^{J} N_j \boldsymbol{\Sigma}_j, \qquad \mathbf{S}_B = \frac{1}{N}\sum_{j=1}^{J} N_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^{T} - \boldsymbol{\mu}\boldsymbol{\mu}^{T}. \tag{3.17}$$

The linear projection should maximize the ratio of between-class and within-class scatter. In the general case, however, determinants of the scatter matrices are used. The criterion function can be written as

$$J(\theta) = \frac{\left|\theta\,\mathbf{S}_B\,\theta^{T}\right|}{\left|\theta\,\mathbf{S}_W\,\theta^{T}\right|}, \tag{3.18}$$

and $\hat{\theta} = \arg\max J(\theta)$ is the estimated transformation matrix. The LDA-projected feature vectors are computed as $\mathbf{y} = \theta\,\mathbf{x}$.

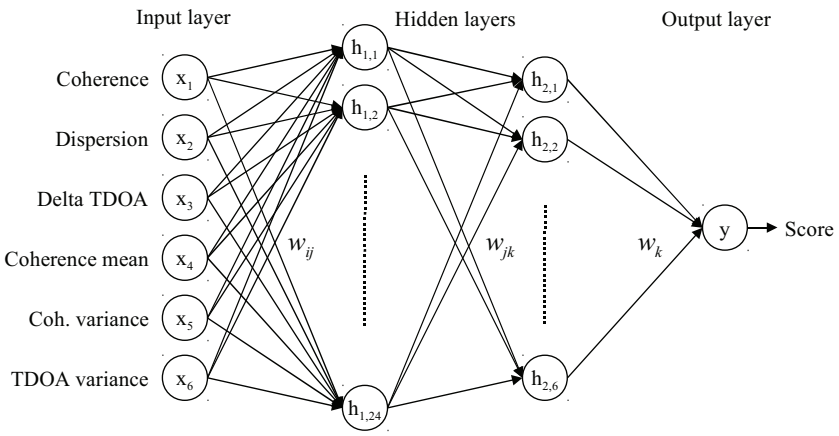


Figure 17: Topology of the MLP used for fusion of spatial data from several microphone pairs.

Similarly to the PCA-fusion strategy, we estimated a LDA-projection matrix for each spatial feature kind: θ_C for coherence, θ_D for dispersion, and θ_{dT} for delta TDOA. We performed the LDA considering only the overlapping and single-speaker speech, since the discrimination between these two classes is the focus of our attention. Consequently, the $1 \times n$ projection matrices, n being here the number of microphone pairs, project the data for a given frame to a one-dimensional parameter. The joint three-dimensional spatial feature vector can be expressed as in (3.16), only replacing the principal components $\mathbf{u}_{C\,1}$, $\mathbf{u}_{D\,1}$, and $\mathbf{u}_{dT\,1}$ with LDA projections θ_C, θ_D, and θ_{dT}, respectively.

3.4.3 *Artificial Neural Network Fusion*

Another issue is that the proposed spatial features are, in general, not comparable across different microphone pairs, since they are intrinsically tied to physical characteristics of the pair like the inter-microphone distance. To normalize the spatial features and reduce their dimensionality we consider a Multi-layer Perceptron (MLP) neural network [129]. The MLP is composed by four layers with six input neurons, its topology is given in Figure 17. The input corresponds to three spatial features and three normalization values (*mean of coherence*, *variance of coherence*, *variance of TDOA*) for every microphone pair. The two hidden layers have 24 and 6 neurons, respectively. The output is a binary score classifying between overlap and non-overlap, which is commensurable across microphone pairs. For a given frame the average score across all microphone pairs is taken.

Multi-layer perceptron classification

The process of obtaining the output score can be mathematically expressed as

$$y = \sigma\left(\sum_{k=1}^{6} w_k \cdot \sigma\left(\sum_{j=1}^{24} w_{jk} \cdot \sigma\left(\sum_{i=1}^{6} w_{ij} x_i\right)\right)\right), \tag{3.19}$$

where w denotes the weight of a connection between two neurons and $\sigma(\cdot)$ is the sigmoid function (also called *logsig*) defined by the formula

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{3.20}$$

Starting with the first hidden layer, the input to each neuron is taken and the output is found by first calculating the weighted sum of inputs $\{x_i\}$. Then, the sigmoid function is applied to it, and it is passed forward to the next layer until the output layer is updated. The standard way to train a multi-layer perceptron is using a supervised learning method called *error back-propagation* [11].

Comparison of PCA-, LDA-, and ANN-based fusion

The histograms of spatial parameters after the application of PCA-, LDA-, and MLP-based microphone data fusion are given in Figure 18. The absolute values of the transformed features are not important, since the original parameter values were projected to different space. However, the histograms show that the PCA-transformed coherence and dispersion ratio in Figures 18 (a) and (c) are spanned over larger intervals compared to LDA-transformed coherence and dispersion in Figures 18 (b) and (d). LDA parameters, on the other hand, seem to have less overlapping distributions. These observations are consistent with the characteristics of the two statistical techniques, PCA being focused on signal representation and high variance, and LDA on enhancing the class-discriminatory information. In case of the MLP scores in Figure 18 (g), the values range from -1 to 1 where the lower value corresponds to single-speaker speech and the higher value to overlapping speech.

PCA emphasizes parameter variance while LDA the class discriminability

In terms of the KL2 divergence calculated on AMI single-site development data (refer to Section 5.1 for more details on data sets), the PCA coherence, dispersion, and delta TDOA features yield 0.0513, 0.1078, and 0.0435, respectively. For the LDA features in the same order, the values are 0.2515, 0.2360, and 0.0479. The KL2 divergence of MLP score distribution is 0.3387.

3.5 PROSODY-BASED FEATURES

A few studies were published which researched the relationship between prosodic cues and the interaction of conversation participants.

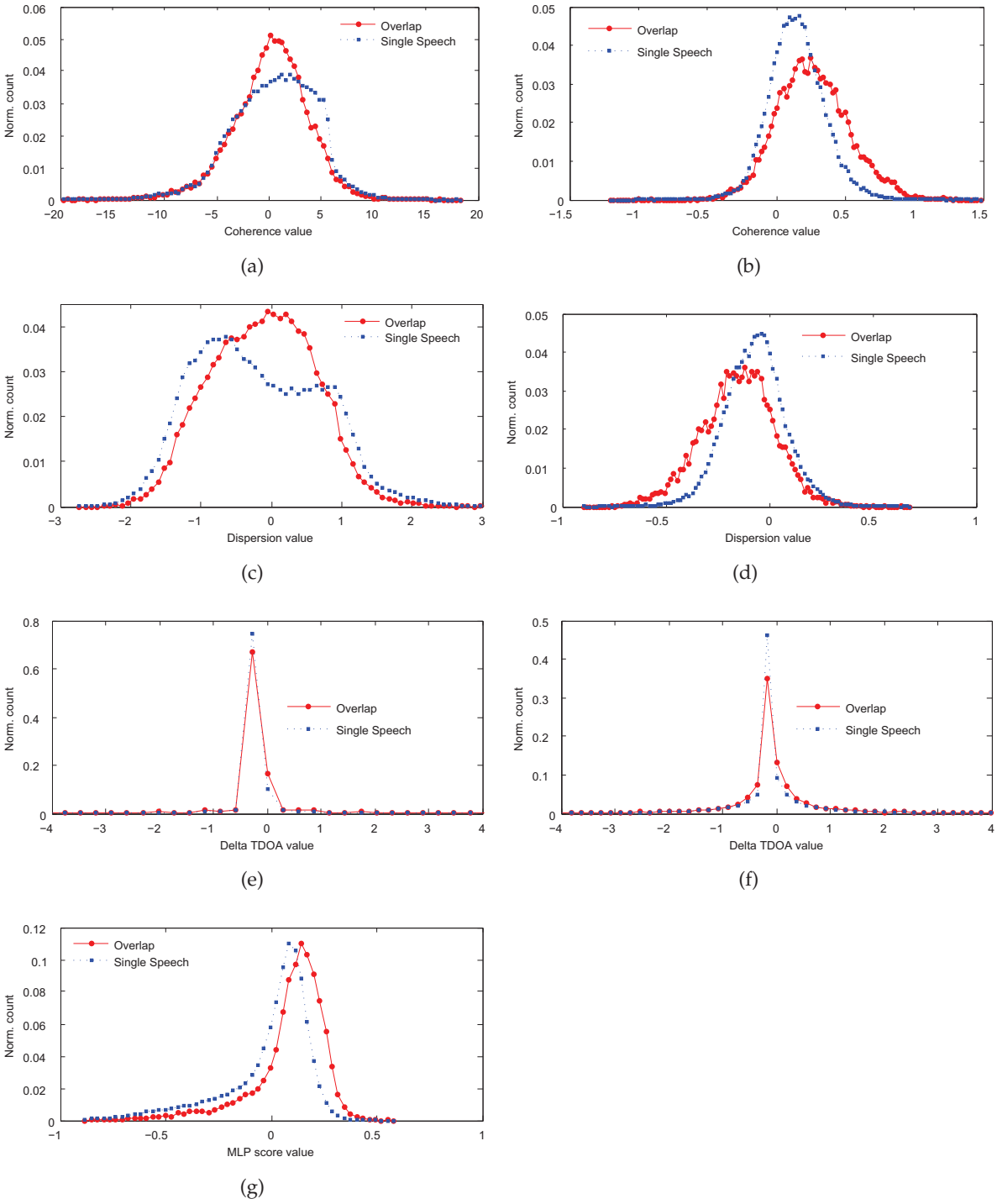


Figure 18: Histograms of spatial feature candidates: (a) PCA- and (b) LDA-transformed coherence, (c) PCA- and (d) LDA-transformed dispersion ratio, (e) PCA- and (f) LDA-transformed delta TDOA, and (g) spatial MLP score. Histograms are computed on AMI single-site development data.

The work by Ward and Tsukahara [134], for instance, suggests that stretches of low pitch can trigger backchannel feedback from listener (*yeah, uh-huh, right*). In another publication, Shriberg et al. [135] showed that speakers raise their voices when starting their utterance during somebody else's talk, compared to starting in silence.

This section concerns with the application of prosodic features and their statistical characteristics, which are obtained over relatively long time spans. Candidate features are investigated for their overlap discrimination properties, and a two-stage feature selection process is outlined.

3.5.1 *Candidate Features and Long-Term Statistics*

Prosody, in general, is characterized by rhythm, intonation, stress, and juncture in speech. These attributes, however, cannot be measured directly, only their acoustic or perceptual correlates can be extracted from speech signal. For the detection of overlapping speech a number of prosody-based features are considered [136, 137]. They can be assigned to one of the following categories:

- fundamental frequency,
- acoustic intensity,
- formant frequencies.

Fundamental frequency, acoustic intensity, and formants

Fundamental frequency (F0) is the rate of vibration of the vocal folds during voiced speech measured in Hz. The term F0 is often, though incorrectly, used interchangeably with pitch, which is the perceptual correlate of F0. Intonation, for instance, is represented by changing suprasegmental—representation above the level of a phoneme—patterns of F0. Fundamental frequency is influenced by age and gender. For male voices it typically ranges from 100 to 150 Hz while for females it is 170–220 Hz.

The values of acoustic intensity, usually expressed in dB, indicate the energy of the speech signal. Changes in intensity, or loudness, are relevant for marking stress and can reflect emotions of speakers.

Formants denote concentrations of acoustic energy around particular frequencies at approximately 1000 Hz intervals. These frequencies correspond to the resonances of the vocal tract tube, however, they only occur in voiced speech segments. The first two formants are the most important for determining the phonetic content. The higher formants are assumed to convey mainly the speaker-specific information. Four formant frequencies, i. e., F1, F2, F3, and F4 were extracted from the speech signal.

In addition to the actual values of the prosodic parameters (F0, formants, intensity) for any given time point, suprasegmental statistical

characteristics are also computed. The long-term statistics include the mean, median, minimum, maximum, standard deviation, and the difference between the minimum and maximum of the extracted prosodic parameters. They are computed over 500 ms windows with a 10 ms step for synchronization reasons with the baseline features. Missing values, such as F0 estimates for unvoiced speech, or parameters in non-speech regions, are substituted with default values. Moreover, non-speech regions are not considered when computing statistical parameters.

Long-term statistics

A similar set of candidate features was investigated for speaker-discriminative properties in [36]. Prosodic features were extracted with the help of Praat[2].

3.5.2 *Feature Selection*

Using the full set of candidate features might be suboptimal for the overlap detection system. There are two options for reducing the dimensionality of feature vectors, the first one being the feature extraction/transformation approach with methods covered in Section 3.4, such as the PCA or LDA. The other approach is to select a subset of existing features without a transformation, referred to as *feature selection*. Feature selection might be preferable in situation when features are expensive to obtain or the measurement units of features want to be maintained. Furthermore, less features means reduced complexity and run-time.

Feature selection requires two things, a search strategy to select candidate subsets, and an objective function to evaluate these candidates. Search strategies can be grouped in one of the following categories: exponential, sequential, and random algorithms. An example of the exponential algorithm is the simple exhaustive search which involves 2^N possible combinations, N being the number of feature candidates. Sequential forward selection is a representative of sequential algorithms. It starts from an empty set and sequentially adds candidate features that result in the highest objective function when combined with the features that have already been selected.

Search strategy and objective function — two components of feature selection

Objective functions are divided into two groups: filters and wrappers. Filter objective function evaluates feature subsets by their information content, such as interclass distance, for instance. Wrapper objective function, on the other hand, is actually a patter classifier which evaluates according to the recognition rate on some test data.

In this work, the used feature selection process consists of two stages [136, 137]. In the first, a minumum Redundancy Maximum Relevance (mRMR) algorithm [138] was applied on held-out development

mRMR-based feature selection

2 Praat: doing phonetics by computer [Computer program]. Version 5.2.04, retrieved from http://www.praat.org/

Table 2: Candidate prosodic features sorted according to the mRMR criterion, fundamental frequency (f0), intensity (int), and formants (f1-4).

ORD.	NAME	SCORE	ORD.	NAME	SCORE
1.	f0_max	0.026	22.	f1	-0.037
2.	f4_max	0.000	23.	f2_max	-0.041
3.	f4	0.000	24.	int_std	-0.040
4.	f0_min	0.000	25.	f3_med	-0.041
5.	int	-0.002	26.	f3_diff	-0.044
6.	f2_min	-0.007	27.	f0_mean	-0.044
7.	f4_min	-0.009	28.	f4_diff	-0.046
8.	f1_min	-0.011	29.	f3	-0.045
9.	f2_med	-0.015	30.	f4_mean	-0.047
10.	f3_max	-0.015	31.	f0_diff	-0.047
11.	int_diff	-0.013	32.	f3_std	-0.049
12.	f3_min	-0.018	33.	int_max	-0.051
13.	f0	-0.019	34.	f1_med	-0.051
14.	f2	-0.026	35.	f2_diff	-0.051
15.	f2_std	-0.024	36.	f1_std	-0.059
16.	f0_med	-0.025	37.	f2_mean	-0.060
17.	f4_med	-0.030	38.	int_med	-0.067
18.	f1_max	-0.029	39.	f1_mean	-0.069
19.	f0_std	-0.034	40.	f1_diff	-0.069
20.	int_min	-0.038	41.	f3_mean	-0.077
21.	f4_std	-0.039	42.	int_mean	-0.086

data to individually score the candidate features against the target class (overlapping speech vs. single-speaker speech), and sort them according to their minimum redundancy and maximal relevance. The mRMR criterion is commonly used for first-order incremental feature selection and it is an equivalent form of the maximal statistical dependency criterion based on mutual information. Table 2 gives the sorted 42 candidate features. The highest scores yield the F0 maximum, the F4 maximum, the actual F4 estimate every time step, and the F0 minimum.

The second feature selection stage involves conventional hill climbing wrapper approach, i. e., iteratively adding candidate features to the feature subset, creating a model, and evaluating the system on the development data. The obtained experimental results will be presented in Section 6.3.

3.6 MODELS AND DECODING NETWORK

Three acoustic classes: non-speech, single speech, and overlap

Overlap detection system considers three acoustic classes representing non-speech, single-speaker speech, and overlapping speech. For each class a continuous density HMM is defined. In order to obtain a more accurate modeling of transitions between classes, the HMMs have three states. Three states empirically showed to be a fair compromise between imposed minimum duration constraint and simplicity. As most other continuous density HMM systems, the output distributions are modeled by Gaussian mixture densities (GMMs) using diagonal covariance.

The output distribution of a particular state Θ_t represented by a GMM was defined in (2.2). When modeling different information sources (e. g., spectral, spatial, or prosodic features), a further generalization is made. Each observation vector $\mathbf{y}_t \in \mathbb{R}^n$ at time t can be split into a number of S independent data streams $\mathbf{y}_{st}$. The formula for computing $b_{\Theta_t}(\mathbf{y}_t)$ is then

$$b_{\Theta_t}(\mathbf{y}_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} w_{\Theta_t m s} \, N(\mathbf{y}_{st}, \boldsymbol{\mu}_{\Theta_t m s}, \boldsymbol{\Sigma}_{\Theta_t m s}) \right]^{\gamma_s}, \tag{3.21}$$

where M_s is the number of components in stream s, with corresponding component weights $w_{\Theta_t m s}$. The exponent γ_s is a stream weight which is used to give a particular stream more emphasis [13]. The Gaussian density $N(\cdot)$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is defined as follows,

$$N(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \, |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}-\boldsymbol{\mu})}. \tag{3.22}$$

Since the amount of training data among different classes is not balanced (refer to Chapter 5), the baseline system uses in each state 256 Gaussian components for single-speaker speech and a smaller number, 64 components, for overlapping speech and non-speech. Unlike the baseline feature GMMs, the spatial and prosodic likelihood distributions use only 32 Gaussian components regardless the class.

In general, there is less spatial feature vectors compared to baseline MFCCs, for instance, because the frame rate in the first case is 64 ms whereas it is only 10 ms in the latter case. In order to synchronize the frames, the spatial features are repeated in time accordingly. The three HMM states will not capture much temporal evolution of spatial features and it has little sense to use completely different GMMs. For this reason, the spatial Gaussian mixtures share their means and variances across the three states of a particular HMM. The mixture weights in different states are not shared, though.

Given the pooled training data, the iterative training process for the estimation of joint overlap detection models can be described with following steps:

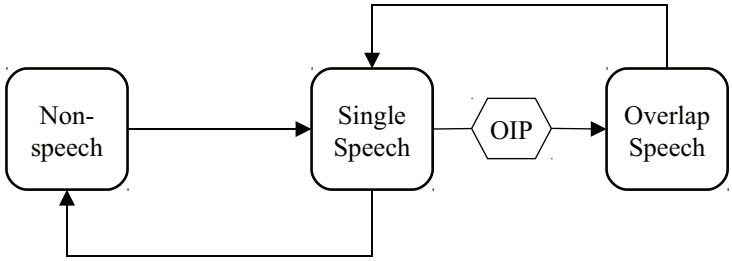


Figure 19: Work network topology in decoding process of the overlap detection system. OIP refers to Overlap Insertion Penalty.

Iterative training algorithm with Gaussian splitting

1. HMMs with only a single Gaussian per state are initialized.
2. An initial Baum-Welch re-estimation is performed.
3. The number of mixture components is doubled by using a Gaussian-splitting technique.
4. (Mean and variance parameter tying in case of spatial Gaussian mixtures.)
5. A single Baum-Welch re-estimation of the GMM parameters.
6. Go to the step 3 if the final number of GMM components is not reached yet, or finish.

Viterbi decoding with imposed overlap insertion penalty

Detection hypothesis is obtained by Viterbi (maximum-likelihood) decoding and applying a word network whose topology is depicted in Figure 19. The transition probabilities between different HMMs are not trained. They are set manually. In order to inhibit the number of false overlap segments, and thus increase the precision, the transition from single speech to overlap speech can be penalized with an Overlap Insertion Penalty (OIP) and certain transitions are completely forbidden. The OIP parameter could be perceived as a compensation for an undertrained model. After obtaining an overlapping speech hypothesis, the information about overlap segments is used as an input in the speaker diarization system, as was shown in Figure 6. The model training and decoding is performed using the HTK[3] framework.

3.7 EVALUATION METHOD

There are two types of errors in Overlap Detection (OD), missed and false overlapping speech amounting to total durations of $T_{miss}^{(ov)}$ and $T_{false}^{(ov)}$, respectively. Missed overlaps correspond in a classical binary

3 HTK: Hidden Markov Toolkit [Computer program]. Version 3.4, retrieved from http://htk.eng.cam.ac.uk/

detection scheme to False Negative errors and false overlaps correspond to False Positive errors. Let denote the total amount of reference overlapping speech as $T_{ref}^{(ov)}$ and the amount detected by the overlap detection system as $T_{sys}^{(ov)}$.

The first evaluation metric reflecting the amount of True Positives is *Recall*. It is defined as the ratio between the true detected and the reference overlap time:

Recall, Precision, and Overlap detection error

$$R = \frac{T_{sys}^{(ov)} - T_{false}^{(ov)}}{T_{ref}^{(ov)}}, \tag{3.23}$$

where the amount of correctly detected overlapping speech in the nominator can also be expressed as $T_{sys}^{(ov)} - T_{false}^{(ov)} = T_{ref}^{(ov)} - T_{miss}^{(ov)}$.

The second metric is *Precision*, which is the ratio between true and all detected overlap time:

$$P = \frac{T_{sys}^{(ov)} - T_{false}^{(ov)}}{T_{sys}^{(ov)}}. \tag{3.24}$$

For instance, a system with a freely set operation point can yield very high recall, but if the system introduces a lot of false overlapping speech segments, the precision will be low. In some publications on speaker overlap detection [112, 113] another metric called F_{score} is also used. It is defined as the harmonic mean of precision and recall, but here it is not considered.

Evaluation metric which measures all the error related to overlapping speech is calculated as the sum of missed and false overlap time divided by the reference overlap time. In this thesis it is referred to as *Overlap detection error*, or simply Error. It can be expressed as:

$$E = \frac{T_{miss}^{(ov)} + T_{false}^{(ov)}}{T_{ref}^{(ov)}}. \tag{3.25}$$

The ratio between false positive time and reference time is referred to as False Alarm rate, $FA = T_{false}^{(ov)} / T_{ref}^{(ov)}$.

Note that the evaluation metrics are very strongly influenced by the overlap insertion penalty, since this penalizing parameter controls the number of overlap segments the system will hypothesize. A common way of demonstrating the performance of a binary detection system is by means of a Receiver Operating Characteristic (ROC) curve. This ROC curve plots the false positive rate (FA) against the false negative rate (or its complement, recall R) for a number of sensitivity thresholds. In case of overlap detection system, the operating point is controlled by the OIP.

ROC curve and OD error area

When comparing two systems by means of ROC curves, it is unfortunately not always clear which system performs better. One system

can have lower detection error for high penalization, but the contrary is true for low penalization. In order to solve this issue, we suggest to calculate the area under these curves, and use this measure as a decision factor. The amount of area reflects the overlap detection error of a particular system. For a fair comparison, every curve is extended with the same fictional start point $[R = 100\%, FA = 100\%]$ and end point $[R = 0\%, FA = 0\%]$.

4

HANDLING OVERLAPPING SPEECH IN SPEAKER DIARIZATION

The two questions raised by Otterson and Ostendorf in [5], i. e., if the diarization score can be improved by assigning more speaker labels and if discarding speech containing multiple speakers from training data in the diarization process will result in purer speaker models, are outlining the topic of this chapter. Before, however, a description of the UPC speaker diarization system considered as baseline in this work is given. The techniques that handle overlapping speech, namely overlap exclusion and labeling, and their integration into the diarization system are discussed afterwards.

4.1 UPC BASELINE SPEAKER DIARIZATION SYSTEM

4.1.1 *Diarization System Architecture*

The UPC speaker diarization system follows the commonly used agglomerative clustering approach. Firstly, speech is broken into short uniform segments, and then the successive clustering stage groups acoustically similar segments and assigns them to speaker clusters. Figure 20 depicts an overall scheme of the diarization system submitted to previous RT '07 and RT '09s evaluations [127]. The main stages of the diarization can be condensed into the following points:

Main stages of the diarization algorithm

- Feature extraction and removal of non-speech frames. At this stage, a clustering initialization is also performed based on an homogeneous partitioning of the data (Figure 20 block A).
- Complexity selection of the models based on the amount of data per cluster and the cluster complexity ratio (R_{CC}) which fixes the amount of speech per Gaussian. HMM/GMM training and cluster realignment by Viterbi decoding based on maximum likelihood (Figure 20 block B).
- Agglomerative clustering based on the Bayesian Information Criterion (BIC) metric among clusters. The stopping criterion, also based on the BIC, drives the ending point of the algorithm (Figure 20 block C).

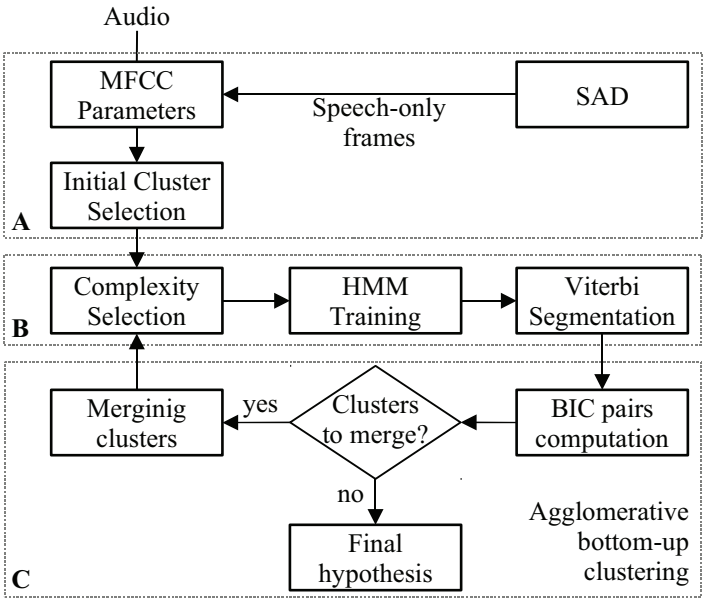


Figure 20: Speaker diarization system architecture.

4.1.2 *Integrated Segmentation and Clustering Algorithm*

More in detail, the audio parametrization for the integrated segmentation and clustering consists of the extraction of 20 MFCCs from 30 ms windows with 10 ms shifting. Aiming to avoid non-speaker information such as background or channel noise, non-speech frames are discarded from further processing based on either reference speech/non-speech annotation (default) or a SAD hypothesis.

Clusters are initialized by uniform segmentation

At the beginning of the clustering algorithm, an uniform initialization is performed so that the system starts with a homogeneous splitting of the whole data among the initial number of clusters (Figure 20 block A). The number of initial clusters is determined automatically depending on the meeting length with minimum and maximum value constraints. In this work the total amount of clusters was constrained to a minimum and a maximum of 20 and 55 clusters, respectively, in order to avoid overclustering and to reduce the computational cost of the iterative approach.

Once the initial segmentation is performed, each cluster is modeled by one mixture of Gaussians, fitting the probability distribution of the features by the classical EM algorithm (Figure 20 block B). The automatic selection of the number of clusters (K_{init}) is defined as

$$K_{init} = \frac{N}{G_{init}\, R_{CC}}. \tag{4.1}$$

This expression takes into account the total number of speech frames in the meeting (N), the number of Gaussians initially assigned to each speaker cluster (G_{init}), and the cluster complexity ratio (R_{CC}). The

R_{CC} is a constant value across all meetings that defines the number of frames per Gaussian. It was fixed to 7 s of speech per Gaussian whereas the initial number of Gaussians per model (G_{init}) was set to 5.

Bottom-up clustering with several HMM alignment and re-estimation iterations

It follows an iterative bottom-up strategy driven by a loop of BIC estimations and HMM alignments (Figure 20 block C). In this step the segments which belong to the same speaker are combined into a new model at each iteration. A time constraint as in [20] is also imposed on the duration of the speaker segments by setting the transition probability among each cluster. In that sense, Viterbi decoding decisions are taken based on the accumulation of the emission probabilities in a 3 s window.

We used a modified BIC-based metric [20] to determine the most likely pair of clusters to merge. The segmentation obtained at the output of block B (see Figure 20) defines a new set of speaker clusters which will be retrained. Most of the systems based on agglomerative clustering perform just one merge at each BIC iteration, where the cluster pair with the highest BIC values is merged. This system, however, applies a threshold that depends on the standard deviation of the BIC values obtained across cluster pairs. It was decided to merge all cluster pairs (i, j) which are fulfilling

Modified BIC metric to speed-up cluster merging

$$BIC_{ij} > BIC_{\mu} + \frac{3}{2} BIC_{\sigma}, \tag{4.2}$$

where BIC_{ij} is the BIC value between the clusters i and j, BIC_{μ} is the mean of BIC_{ij} for $i \neq j$, and BIC_{σ} is the standard deviation for the same set. For this reason, it is possible for the system to merge more than one pair of clusters per iteration, speeding up the agglomerative clustering.

Automatic model complexity estimation

At each iteration n, the number M_i^n of Gaussians to model the cluster i is updated by

$$M_i^n = \left\lfloor \left(\frac{N_i^n}{R_{CC}} \right) + \frac{1}{2} \right\rfloor, \tag{4.3}$$

where N_i^n is the number of frames belonging to the cluster i. Whenever two segments are merged, a new segment model is also trained pooling all the features from the merged segments and fixing the model complexity according to the R_{CC} value. Such automatic selection of the modeling complexity has demonstrated a successful performance while avoiding the use of the penalty term in the classical BIC metric [139]. This procedure is iterated until all the BIC values of the remaining cluster pairs are negative, which means that no suitable candidates for merging are found anymore. Finally, at the last iteration and once the stopping criterion is met, each remaining state represents a different speaker. A detailed description of the system can be found in [127].

4.1.3 *Multi-Microphone Approach*

Weighted delay-and-sum beamforming

The baseline speaker diarization system can be improved by multi-channel approach based on conventional techniques. Firstly, the Wiener filtering is applied using the noise reduction implementation from the QIO front-end [140] on each microphone signal. Next, we apply the weighted delay-and-sum technique [141] to perform the signal enhancement. In order to synchronize two microphone signals and enhance the signal-to-noise ratio of the signal mixture, TDOAs are estimated. In addition, TDOAs can serve as a second stream of information when combined with the classical MFCC parameters in the diarization algorithm.

TDOA feature stream in combination with MFCCs

TDOAs are computed by means of the Generalized Cross-Correlation with Phase Transform Weighting (GCC-PHAT) method which was already defined in (3.10). The TDOAs for two microphones are computed similarly to (3.11), using a window of 500 ms at a rate of 250 ms applied on the Wiener-filtered channels. The TDOA information is combined with the MFCC stream along the diarization process in the Viterbi path as well as in the BIC estimation. The joint log-likelihood is estimated as a weighted linear combination of the log-likelihoods of each stream. Each stream is considered to be statistically independent from each other, as in [123].

4.1.4 *Speech Activity Detection*

SVM-based SAD system, however, the default are reference speech/non-speech annotations

In some experiments in this work a SAD system developed at the UPC, which has shown a good performance in last RT SAD evaluations [142], is applied. The algorithm is based on a proximal SVM (PSVM) [143] and on a fast training technique which allows the training of huge amounts of data.

The SVM-based SAD system was trained with the RT'05, RT'06, RT'07 conference data, the CHIL'07 meeting data, and the Speecon far-field microphone data. It yielded to more than 25 hours of training material.

Nevertheless, the default option regarding SAD is to use reference speech/non-speech annotations. The reason is that this thesis mainly focuses on studying the impact of overlapping speech on speaker segmentation and clustering. The addition of another tunable system (SAD) to the processing chain only complicates this task and can possibly introduce more confusion. The use of a real SAD system should rather complete the picture for the reader about the performance of speaker diarization.

4.1.5 *Diarization Scoring*

Performance of speaker diarization systems is evaluated by means of DER, a time-weighted metric defined by NIST[1] [118]. The audio file is divided into contiguous segments demarcated by all reference and system speaker change points so that the set of compared speakers in one segment does not change. Then, the DER metric is defined as follows,

$$\mathrm{DER} = \frac{\sum_{\forall s} \mathrm{dur}(s) \cdot (\max(\mathrm{N}_{ref}(s), \mathrm{N}_{sys}(s)) - \mathrm{N}_{correct}(s))}{\sum_{\forall s} \mathrm{dur}(s) \cdot \mathrm{N}_{ref}(s)}, \quad (4.4)$$

where $\mathrm{dur}(s)$ is the duration of a particular segment s, $\mathrm{N}_{ref}(s)$ is the number of reference speakers speaking in segment s, $\mathrm{N}_{sys}(s)$ is the number of system speakers in segment s, and $\mathrm{N}_{correct}(s)$ is the number of matching reference and system speakers who are speaking in segment s. DER represents the ratio of incorrectly attributed speech time to the total amount of speech time. It can be decomposed into missed speaker time error, false alarm error, and speaker error (speech assigned to the wrong speaker). Since there is no a priori relation between the system and reference speaker clusters, an optimum one-to-one mapping of reference speaker IDs to system output speaker IDs is determined separately for each audio file.

Diarization error consists of missed speech, false alarms, and speaker confusion

As will be shown in Section 5.1 the median overlap duration is rather short. To make sure that overlap segments are considered in scoring, normally no forgiveness collar was applied around segment boundaries.

4.2 OVERLAP HANDLING TECHNIQUES

Techniques which handle overlapping speech in speaker diarization comprise the exclusion of overlap frames from model training and the assignment of second speaker labels for overlap segments, also referred to as overlap labeling. The aim of the first is to achieve purer cluster models and thus a more precise segmentation. The latter strives to recover missed speaker time which contributes to the DER.

Figure 21 shows the relationship of these two techniques to function blocks of the UPC diarization system. They work independently from each other, or better said sequentially, since the exclusion works throughout the diarization process whereas the labeling is performed at the end of the iteration process. Understandably, overlap exclusion affects the outcome of overlap labeling by means of the trained cluster models. These two techniques do not necessarily have to share the

Independence of overlap exclusion and labeling allows for their separate optimization

1 NIST scoring tool available at: http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl

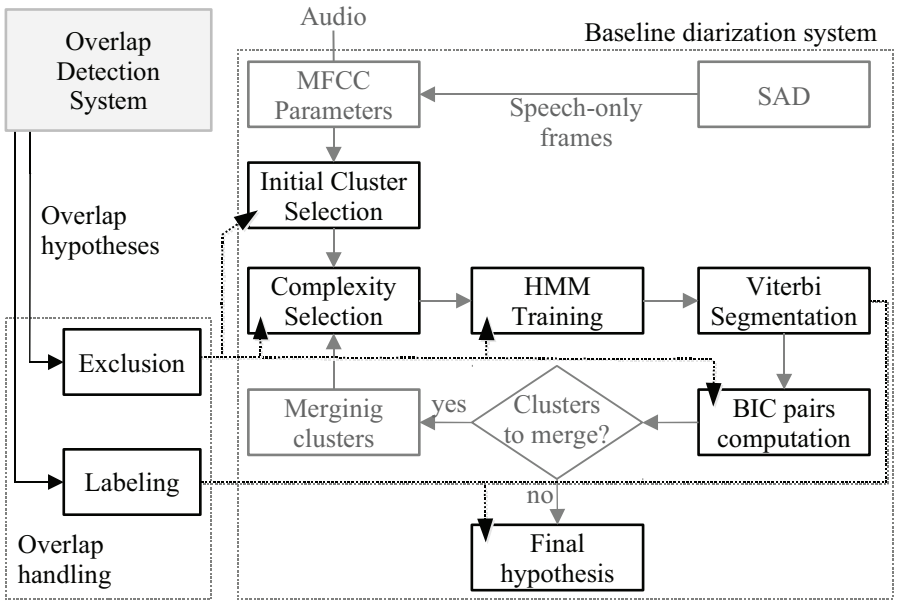


Figure 21: Overlap handling in speaker diarization system. Affected modules by the exclusion and the labeling of overlapping speech.

same overlap hypothesis but can possibly use two different hypotheses, i. e., one for each technique. In other words, overlap labeling and exclusion can be optimized independently regarding the used overlap detection hypotheses. This method was firstly suggested in [129] and then also applied in the following works [130, 136].

4.2.1 *Overlap Exclusion*

Single-speaker models could be corrupted due to multiple-speaker speech in training data

Cluster models should in the end correspond only to single speakers. The original assumption behind overlap exclusion is that overlapping speech frames can lead to corruption of these cluster models, since they implicitly introduce speech from more people. Furthermore, a large amount of overlapping speech can possibly result in over-clustering, i. e., stopping the clustering process with more final speaker clusters than the correct number of speakers.

The overlapping speech hypotheses are, however, not perfect and there exists a risk that too much clean data could be taken away. Consequently, the speaker models trained on less data will not be trained as well as normally.

Exclusion of overlapping speech does not mean that these frames are completely thrown away. They are not considered in some steps of the diarization algorithm, but are maintained, for instance, in Viterbi decoding. The functional modules affected by discarding overlap frames, visible in Figure 21, are automatic cluster selection, complexity selection, HMM training, and BIC pairs computation.

The UPC speaker diarization system uses a uniform segmentation in the cluster initialization among the automatically computed number of clusters. In this stage of the diarization process there are several variations how to implement overlap exclusion. Firstly, the Automatic Number of Clusters (ANC) can be computed with the original formula defined in (4.1), or the formula can be modified so that overlapping frames are not considered,

Possible variations of overlap exclusion implementation

$$K_{init} = \frac{N - N^{(ov)}}{G_{init}\, R_{CC}}, \tag{4.5}$$

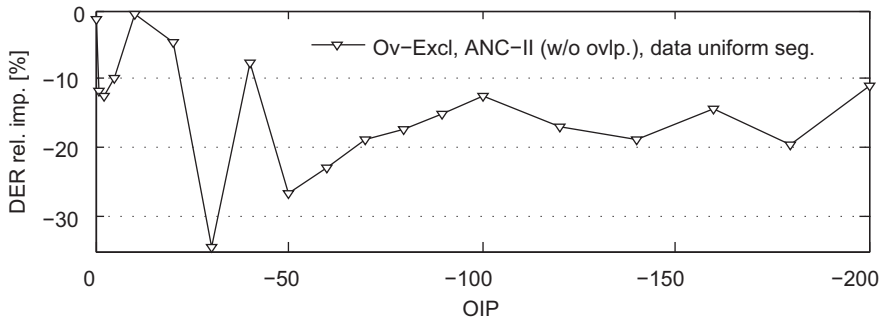
where $N^{(ov)}$ is the total number of detected overlapping speech frames. In the following, the original formula (4.1) is referred to as ANC-I, and the modified version (4.5) as ANC-II. Secondly, in case of using ANC-I, overlapping frames can be discarded before or after uniformly dividing the data among clusters. Discarding the overlapping frames before makes the initial clusters data-uniform in the sense that there will be an equal amount of frames in each cluster, whereas discarding these frames after the splitting will distribute the cluster equally in time, i. e., time-uniformly. The initial cluster boundaries in the latter case match the start and end times of clusters when no overlap exclusion is applied. The ANC-II approach implicitly assumes discarding overlapping speech before the initial segmentation.

It may be polemic which of these options is the most correct. In order to select one, the performance of the diarization system in terms of relative DER improvement for the three variations of overlap exclusion implementation in the initialization stage was obtained and is demonstrated in Figure 22. Experiments were performed on AMI development data using overlapping speech hypotheses acquired for a number of OIP working points. The results show that there is no clear optimal implementation from among using ANC-II (Figure 22 (a)), or using ANC-I with data-uniform clusters (Figure 22 (b)) or time-uniform clusters (Figure 22 (c)). In the end, the ANC-II variation was selected as the final implementation for overlap exclusion since it is probably the most logical alternative.

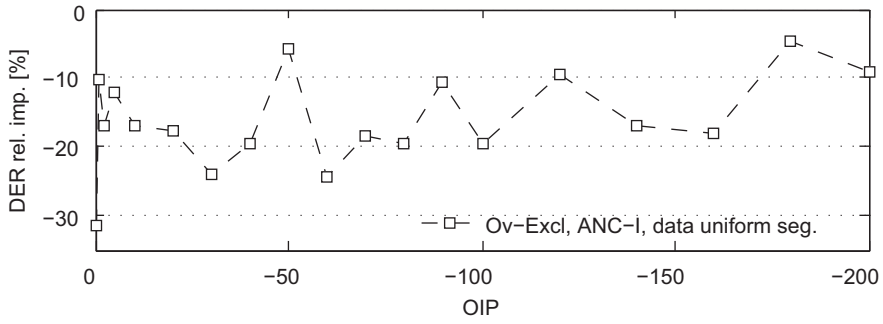
For complexity selection, the employment of overlap exclusion modifies the way the appropriate number of Gaussians is determined from (4.3) to

$$M_i^n = \left\lfloor \left(\frac{N_i^n - N_i^{(ov)\,n}}{R_{CC}} \right) + \frac{1}{2} \right\rfloor . \tag{4.6}$$

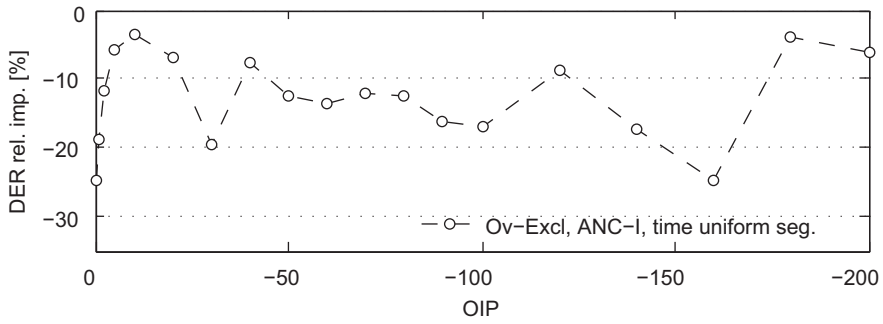
Here, $N_i^{(ov)\,n}$ is the number of overlapping speech frames belonging to the cluster i at iteration n. In case of HMM training and BIC pair computation, respective formulas are modified similarly and overlapping frames are not considered in these steps.



(a)



(b)



(c)

Figure 22: Relative improvement over baseline DER 28.3% by excluding simultaneous speech from cluster-model training as a function of OIP applied in overlap detection. Overlapping speech frames are discarded (a) (b) before, or (c) after the initial uniform segmentation. The Automatic Number of Clusters (ANC) is computed (b) (c) with, and (a) without considering overlap frames. Experiments are performed on AMI single-site development data.

4.2.2 *Overlap Labeling*

Given the start and end times of overlapping speech segments, the goal of overlap labeling is to determine the overlapping speaker(s) on top of the original choice of a conventional diarization system. A number of strategies were already proposed in the literature for the assignment of second speaker labels. They range from simple schemes such as the most talkative speaker [119] or nearest neighbor speaker [5, 117] to technique like the one used in [112, 113], which relies on posterior speaker probabilities.

The technique for second speaker-label assignment proposed in this thesis is integrated into Viterbi decoding. Altough a third and even more speaker labels could theoretically also be assigned, here only two speakers are considered for speaker overlap segments. As a matter of fact, two concurrent speakers represent the vast majority of overlapping speech situations (see Chapter 5).

Overlap labeling technique integrated in Viterbi algorithm

Viterbi algorithm can be regarded as a dynamic programming algorithm applied to the HMM, which uses a computationally efficient approximation to estimate the optimal (maximum likelihood, in practice) sequence of states given in (2.3). Instead of summing up probabilities from different paths coming to the same destination state, Viterbi algorithm selects and memorizes just the best path.

For a given HMM with N states—each state represents one cluster—let $\delta_t(j)$ be the probability of the most likely state sequence at time t, which have generated the observation vectors $\mathbf{y}_1, \ldots, \mathbf{y}_t$ and finished in state j. Furthermore, let $\phi_t(j)$ be a back-pointer variable which points to the optimal predecessor of the current state j. For $t = 1$ the variables are initialized as

$$\begin{aligned} \delta_1(j) &= \pi_j b_j(\mathbf{y}_1) \\ \phi_t(j) &= 0, \end{aligned} \tag{4.7}$$

where $1 \leqslant j \leqslant N$, and π_j denotes the initial probability of state j. The induction step of the recursive algorithm is as follows,

$$\begin{aligned} \delta_t(j) &= \max_i \left(\delta_{t-1}(i) a_{ij}\right) b_j(\mathbf{y}_t) \\ \phi_t(j) &= \arg\max_i \left(\delta_{t-1}(i) a_{ij}\right), \end{aligned} \tag{4.8}$$

where $1 \leqslant j \leqslant N$ and $2 \leqslant t \leqslant T$. The best state sequence $\hat{S} = \{\hat{s}_1, \ldots, \hat{s}_T\}$ is obtained by first identifying the last state ($t = T$), $\hat{s}_T = \arg\max_i \delta_T(i)$, and then *backtracking* the other states of the sequence by observing that

$$\hat{s}_t = \phi_{t+1}(\hat{s}_{t+1}), \qquad t = T-1, T-2, \ldots, 1. \tag{4.9}$$

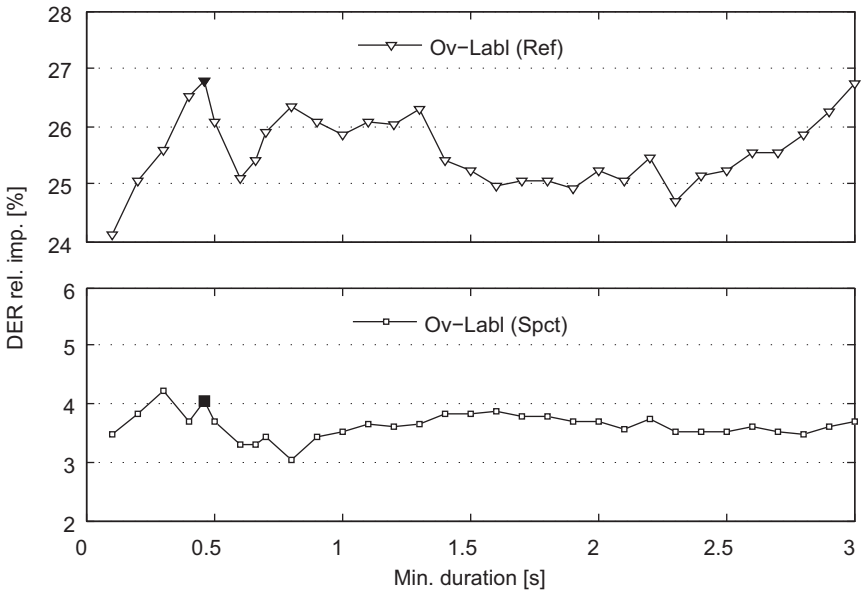


Figure 23: Relative improvement over baseline DER 28.3% by labeling simultaneous speech as a function of the minimum duration decoding parameter when determining the second speaker label. Overlapping speech segments are extracted from annotation reference (top), or real OD system hypothesis. Experiments are performed on AMI single-site development data.

The implementation of overlap labeling requires the introduction of another variable which tracks the second most likely preceding cluster state. Being in time t and current state j, it is defined as

$$\phi_t^{(ov)}(j) = \underset{i;\, i \neq \phi_t(j)}{\arg\max} \left(\delta_{t-1}(i) a_{ij} \right). \tag{4.10}$$

It is initialized in the same manner as $\phi_t(j)$ in (4.7). For overlapping speech segments the alternative state at time t, $\hat{s}_t^{(ov)}$, is determined based on the most likely decoding path with identified states $\hat{s}_1, \ldots, \hat{s}_T$ as

$$\hat{s}_t^{(ov)} = \phi_{t+1}^{(ov)}(\hat{s}_{t+1}). \tag{4.11}$$

Hence, given that speaker overlap was detected at time t, $\hat{s}_t$ represents the first speaker and $\hat{s}_t^{(ov)}$ represents the second, overlapping, speaker.

Minimum duration constraint imposed on decoded clusters

In Section 4.1.2 it was stated that a minimum duration constraint of 3 s is imposed in the decoding in order to prevent too short speaker segments. This value, however, may eventually be sub-optimal in case of the second-speaker segments. A series of development experiments was launched to estimate the most appropriate minimum length of continuous overlapping-speaker chunk. Figure 23 demonstrates the relative DER improvement of diarization system by overlap labeling

with the minimum duration set to a range of values. The minimum duration constraint for the first (default) speaker was left at 3 s, though. When using overlapping speech segments from a real overlap detection system the performance improvements are roughly 3–4%. For minimum durations higher than 1 s the observed improvements seem relatively insensitive to increasing values. In case of assigning second speaker labels to reference overlapping speech, the observed relative improvements are much higher compared to previous case (approx. 25–27%). The reason is that in the former case only around 21% of overlapping speech is actually labeled and there is also a certain amount of false overlaps. In both cases it is possible to observe a relatively high DER improvement at 0.46 s, which is the value that was finally selected for overlap labeling. It is worth mentioning that this value actually corresponds with the median duration of overlapping speech segments (see Chapter 5).

When decoding the first and the overlapping speaker with different minimum duration constraints—two decodings are performed in fact—sometimes the same speaker cluster is picked. In such situations the overlapping speaker is changed to the most likely cluster of the second decoding, which is different from the already selected first speaker from the first decoding.

High-Precision Overlap Detection Requirement

As was already implied, the possible improvement of speaker diarization by overlap labeling is negatively affected by the amount of false overlapping speech. The overlap hypothesis which should be used for labeling needs to be sufficiently precise, since all falsely detected overlaps will directly contribute to the diarization error, but only a perfect selection of speaker labels will recover the missed overlapping speaker time. This requirement was previously also explained in [111] and [110]. The critical precision for overlap hypothesis is 50%, in such case the DER after overlap labeling will at best be the same.

Critical precision of detected overlapping speech to be suitable for labeling is 50%

A model example of this mechanism is illustrated in Figure 24. The missed speech error in a system with overlap labeling is decreased by the amount of detected overlapping speech. The false alarm error, on the other hand, will grow exactly by the amount of falsely detected overlapping speech, because the labeling technique will introduce to all such segments a false speaker. The actual net profit of overlap labeling depends on the difference between overlap recall on one side, and the amount of false overlaps and the increase of speaker error on the other. The speaker error part of DER will almost certainly be increased, since more speaker labels are assigned, in general.

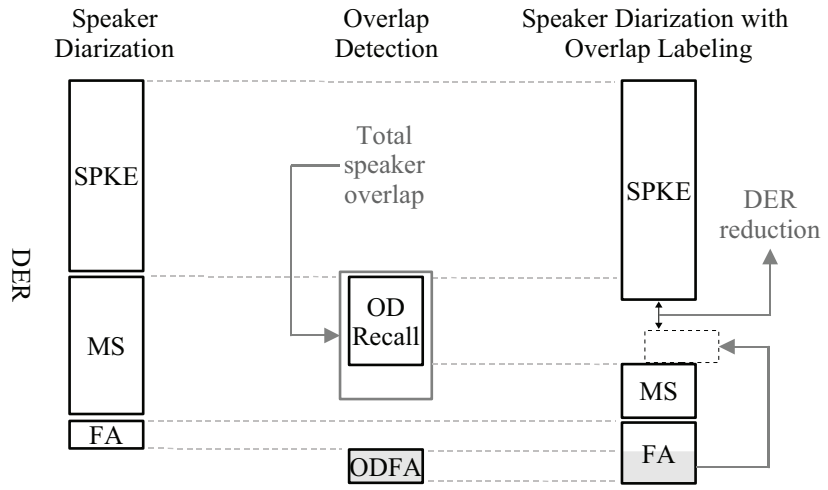


Figure 24: Impact of second label assignment on speaker diarization performance in terms of missed speech error (MS), false alarm error (FA), and speaker error (SPKE). Falsely detected overlapping speech is denoted as OD FA.

5 DATABASES

This chapter gives a description of the corpora used for overlapping speech detection as well as for speaker diarization experiments. The distribution of recordings into various sets is detailed, and issues regarding overlap annotations are discussed. The main experimental data comes from the AMI Meeting corpus. Additional experiments were conducted on the NIST RT recordings.

5.1 AMI MEETING CORPUS

The Augmented Multi-party Interaction (AMI) Meeting corpus [144, 145] consists of 100 hours of audio in 171 meeting recordings which use a range of signals synchronized to a common timeline. These include close-talking and far-field microphones. In addition, this database provides individual and room-view camera videos. The meetings were recorded in English using three different recording rooms with different acoustic properties. They were located at *Idiap*, *Edinburgh*, and *TNO* site. The participants are mostly non-native English speakers, and there are normally four speakers in one meeting. The audio signals are sampled at 16 kHz with 16 bit precision.

Two experimental scenarios involving single and multiple recording sites

This thesis concerns with the use of far-field microphone channels. For experiments on AMI data, two experimental scenarios were defined. The first, *single-site scenario*, included recordings only from the Idiap site, and the other, *mutli-site scenario*, included meeting recordings also from the Edinburgh and TNO sites. Full recordings were divided for both scenarios into training, development, and evaluation sets.

The total duration of audio in single-site sets (maintaining the given order) is 9.7 h, 1.6 h, and 4.8 h, which corresponds to 22, 3, and 11 recordings[1], respectively. The recording distribution is given in Table 3.

In the case of multi-site scenario, the training, development, and evaluation data amounts to 10.8 h, 4.4 h, and 5.9 h corresponding to 22, 9, and 10 recordings[2], respectively. The distribution of meetings is given in Table 4.

1 Originally, the development set also included the recording IS1007d, and the evaluation set the recording IS1003b, but these recordings were later discarded due to the unavailability of multi-channel data. The sets were meant to maintain the distribution used in [110].

2 The evaluation set originally included the recording IS1003b, but due to the unavailability of multi-channel data it was discarded and substituted with the recording IS1008c. As in single-site scenario, the distribution of recordings was meant to be similar to [110].

Table 3: Experimental sets for the AMI single-site scenario.

SITE	TRAINING		DEVELOPMENT	EVALUATION
Idiap	IS1000b	IS1005b	IS1000d	IS1000a
	IS1000c	IS1005c	IS1002d	IS1001a
	IS1001d	IS1006a	IS1004d	IS1001b
	IS1002b	IS1006c		IS1001c
	IS1002c	IS1007a		IS1003d
	IS1003a	IS1007b		IS1006b
	IS1003c	IS1007c		IS1006d
	IS1004a	IS1009a		IS1008a
	IS1004b	IS1009b		IS1008b
	IS1004c	IS1009c		IS1008c
	IS1005a	IS1009d		IS1008d

Table 4: Experimental sets for the AMI multi-site scenario.

SITE	TRAINING		DEVELOPMENT	EVALUATION
Edinburgh	EN2002d	ES2008d	EN2004a	EN2003a
	ES2003b	ES2011a	ES2013c	EN2009b
	ES2005b	ES2012b		ES2008a
	ES2006a	ES2014b		ES2015d
	ES2007a	ES2016c		
Idiap	IB4005	IS1004a	IS1001c	IN1008
	IN1001	IS1006a	IS1001d	IN1012
	IN1009	IS1007a	IS1005a	IS1002c
	IS1001a		IS1007b	IS1008b
			IS1007c	IS1008c
TNO	TS3003c	TS3010a	TS3006a	TS3009c
	TS3006b	TS3010b	TS3012b	
	TS3008b			

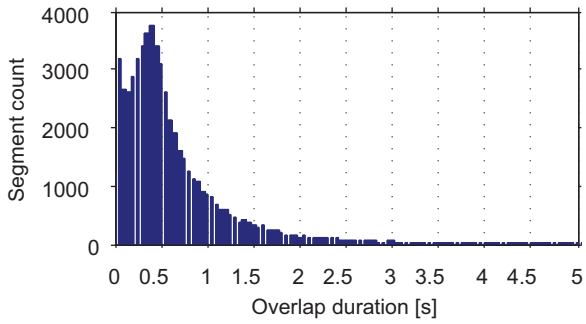


Figure 25: Overlapping speech duration distribution in the AMI corpus.

Overlapping speech in AMI data

The total amount of data in single-site scenario with respect to the three acoustic classes considered in overlap detection system (see Section 3.6) is 4.0 h of non-speech and 10.4 h of single-speaker speech. Overlapping speech yields 1.8 h which is 14.4% out of all speech.

The same categorization of data in multi-site scenario is as follows. The duration of non-speech is 5.3 h, the total amount of single-speaker speech is 13.5 h, and finally, the 2.4 h of overlapping speech constitues 15.1% of all speech.

Different properties of the recordings in the AMI sets, such as the number of speakers or the amount of overlapping speech by two, three, and four speakers, are given in Tables 5 and 6. The amounts of overlapping speech in the training, development, and evaluation set are 12.6%, 16.5%, and 17.4% in single-site scenario and 13.5%, 16.2%, and 17.0% in multi-site scenario, respectively.

Note that the term overlapping speech is used to refer to speech signal whereas speaker time refers to speech in the sense of speech utterances from various speakers. Speaker time is calculated for each speaker independently. For instance, three speakers speaking simultaneously for 3 s produce 3 s of overlapping speech, but it equals 9 s of speaker time. In this sense, overlapping speaker time (OV-SPKT) denotes the time that will be missed by a conventional speaker diarization system which assigns one speaker label per segment, assuming oracle speech/non-speech detection. Moreover, it sets the upper diarization improvement limit by overlap labeling if a perfect overlap detection system would be used. This would have to specify also the number of involved speakers. The upper performance bound of our system, which considers only two-speaker overlap, matches the ratio of overlapping speech (OV).

The duration of continuous segments of simultaneous speech varies, however, the lengths are rather short. The median value in AMI data is 0.46 s, whereas the mean is 0.66 s. The distribution of overlap segment durations is depicted in Figure 25.

Reference annotations, impact of mistakes on overlap detection is negligible

The force-aligned annotations used, for instance, for training and scoring were obtained by SRI's DECIPHER recognizer[3] [146]. Reference annotations are, naturally, not perfect. In order to evaluate the extent of this issue, overlap annotations for the 3 AMI single-site development recordings were manually corrected. In some situations, however, it is very difficult to determine what can and what cannot be considered overlapping speech. For instance, when several consecutive attempts by an interrupting speaker to grab floor are accompanied by non-verbal breathy sounds.

Let us denote the reference annotations as R and their manually corrected version as C. The amount of overlapping speech in common, $R \cap C$, is 635.6 s. Overlap annotated in R, but then discarded during the correction since it was observed to be false, $R \cap C'$, accounts for 42.9 s which is 6.3% of the original amount of overlapping speech. Overlap discovered to be missing in the reference annotations, $C \cap R'$, has a total duration of 9.01 s, being a 1.3% addition to the original overlap. To assess the impact of the annotation differences on the evaluation metrics, the output overlap hypotheses of 50 experimental setups were pairwise scored with R and C ground-truth annotations. The mean difference in the obtained OD errors was $0.57\% \pm 0.14\%$, which we consider acceptable for further use of the reference annotations (R) in our experiments.

Laughter in the AMI corpus

Another open question in the context of simultaneous speech is the presence of laughter, because it is reasonable to assume that people are often laughing together. The occurence of laughter can trigger spontaneous concurrent utterances by speakers willing to share their immediate thoughts, and vice versa (occurence of overlap resulting in laughter). In the corpus annotations, laughter is treated as an acoustic event independent from speech. This means that laughter can be annotated concurrently with speech segments. In the single- and multi-site data there is 1935.0 s and 1686.5 s of annotated laugter, respectively. The correlation with overlapping speech is obvious when comparing segments of both types. Laughter matches overlapping speech in 1427.5 s (73.8% of laughter) for single-site data, and in 1186.8 s (70.4% of laughter) for multi-site data.

3 The annotation were kindly provided by K. Boakye (ICSI, Berkeley).

Table 5: Statistics of recordings in AMI single-site experimental sets. Number of speakers (#Spks); duration of speech, two- (OV-2), three- (OV-3), and four-speaker overlap (OV-4) in (s); overlapping speech as a portion of all speech (OV), and overlapping speaker time as a portion of all speaker time (OV-SPKT) in (%).

MEETING	#SPKS	SPEECH	OV-2	OV-3	OV-4	OV	OV-SPKT
IS1000b	4	1636.7	160.2	11.9	2.6	10.7%	10.5%
IS1000c	4	1559.3	141.7	14.5	1.0	10.1%	10.0%
IS1001d	4	516.7	72.3	9.3	2.2	16.2%	15.9%
IS1002b	4	1872.7	128.2	8.7	0.3	7.3%	7.2%
IS1002c	4	1614.7	151.3	28.8	3.6	11.4%	12.2%
IS1003a	4	457.3	78.1	12.0	0.8	19.9%	18.6%
IS1003c	4	1355.0	170.6	18.1	0.9	14.2%	13.7%
IS1004a	4	479.1	26.7	0.2	0.0	5.6%	5.4%
IS1004b	4	1675.2	144.0	9.8	2.8	9.4%	9.3%
IS1004c	4	1704.1	213.8	26.0	3.8	14.3%	14.0%
IS1005a	4	539.0	46.4	15.2	4.0	12.2%	14.2%
IS1005b	4	1608.1	133.0	17.3	3.9	9.6%	10.0%
IS1005c	4	1455.6	131.1	27.1	2.7	11.1%	11.7%
IS1006a	4	607.5	108.5	51.5	19.6	29.6%	31.0%
IS1006c	4	1454.1	205.7	57.9	19.7	19.5%	21.2%
IS1007a	4	598.9	83.3	20.7	2.3	17.7%	18.0%
IS1007b	4	940.4	147.9	40.6	7.9	20.9%	21.3%
IS1007c	4	1549.8	194.6	25.8	8.2	14.8%	15.1%
IS1009a	4	575.2	79.3	15.8	0.4	16.6%	16.3%
IS1009b	4	1669.5	152.3	28.2	3.3	11.0%	11.7%
IS1009c	4	1409.0	79.5	15.6	1.4	6.8%	7.5%
IS1009d	4	1448.2	136.2	30.4	6.0	11.9%	13.0%
ami-ss_train		**26725.9**	**2784.4**	**485.3**	**97.3**	**12.6%**	**13.2%**
IS1000d	4	1874.0	237.8	23.1	1.0	14.0%	13.3%
IS1002d	4	911.3	111.2	40.0	4.6	17.1%	18.7%
IS1004d	4	1330.0	216.5	37.9	5.8	19.6%	19.0%
ami-ss_dev		**4115.3**	**565.5**	**101.1**	**11.4**	**16.5%**	**16.4%**
IS1000a	4	868.0	98.0	20.1	0.0	13.6%	13.7%
IS1001a	4	590.6	79.6	12.1	1.6	15.8%	15.5%
IS1001b	4	1508.0	129.3	12.9	0.0	9.4%	9.3%
IS1001c	4	1010.5	88.4	9.0	0.6	9.7%	9.7%
IS1003d	4	1602.5	386.7	135.1	26.5	34.2%	31.9%
IS1006b	4	1600.3	183.8	61.4	12.6	16.1%	17.9%
IS1006d	4	1454.3	338.4	172.5	56.7	39.0%	37.6%
IS1008a	4	700.7	33.0	0.9	0.0	4.8%	4.7%

Continued on next page

Table 5—continued from previous page

MEETING	#SPKS	SPEECH	OV-2	OV-3	OV-4	OV	OV-SPKT
IS1008b	4	1275.5	74.7	9.5	0.7	6.7%	7.0%
IS1008c	4	1184.7	138.1	20.8	2.4	13.6%	13.8%
IS1008d	4	1130.3	122.3	22.1	1.8	12.9%	13.2%
ami-ss_eval		**12925.4**	**1672.2**	**476.4**	**102.8**	**17.4%**	**18.7%**
ami-ss		**43766.6**	**5022.2**	**1062.8**	**211.4**	**14.4%**	**15.2%**

Table 6: Statistics of recordings in AMI multi-site experimental sets. Number of speakers (#Spks); duration of speech, two- (OV-2), three- (OV-3), and four-speaker overlap (OV-4) in (s); overlapping speech as a portion of all speech (OV), and overlapping speaker time as a portion of all speaker time (OV-SPKT) in (%).

MEETING	#SPKS	SPEECH	OV-2	OV-3	OV-4	OV	OV-SPKT
EN2002d	4	1764.9	333.1	133.1	45.4	29.0%	30.2%
ES2003b	4	1566.8	59.7	2.2	0.0	3.9%	3.9%
ES2005b	4	1795.1	205.3	26.5	0.8	13.0%	12.7%
ES2006a	4	901.0	81.9	16.0	4.2	11.3%	12.3%
ES2007a	4	717.9	72.7	9.9	1.1	11.7%	11.8%
ES2008d	4	1959.4	219.1	38.7	5.5	13.5%	13.9%
ES2011a	4	707.7	92.5	23.6	4.7	17.1%	17.9%
ES2012b	4	1524.8	104.9	14.2	3.7	8.1%	8.7%
ES2014b	4	1588.3	117.0	10.0	1.8	8.1%	8.2%
ES2016c	4	1462.3	83.7	21.6	5.2	7.6%	8.9%
IB4005	4	1683.9	179.6	27.9	8.7	12.8%	13.5%
IN1001	3	2708.1	456.1	51.1	0.0	18.8%	17.1%
IN1009	4	892.7	108.3	14.0	0.8	13.8%	13.6%
IS1001a	4	590.6	79.6	12.1	1.6	15.8%	15.5%
IS1004a	4	479.1	26.7	0.2	0.0	5.6%	5.4%
IS1006a	4	607.5	108.5	51.5	19.6	29.6%	31.0%
IS1007a	4	598.9	83.3	20.7	2.3	17.7%	18.0%
TS3003c	4	1479.5	79.5	8.1	1.8	6.0%	6.4%
TS3006b	4	1891.7	346.8	55.5	7.5	21.7%	20.3%
TS3008b	4	1806.6	209.5	31.2	2.7	13.5%	13.4%
TS3010a	4	379.3	32.4	3.6	0.4	9.6%	9.7%
TS3010b	4	1128.4	36.1	1.8	0.0	3.4%	3.4%
ami-ms_train		**28234.4**	**3116.2**	**573.4**	**117.6**	**13.5%**	**14.1%**
EN2004a	4	2829.3	426.4	126.5	39.5	20.9%	22.1%

Continued on next page

Table 6—continued from previous page

MEETING	#SPKS	SPEECH	OV-2	OV-3	OV-4	OV	OV-SPKT
ES2013c	4	1727.6	156.5	28.4	0.3	10.7%	11.0%
IS1001c	4	1010.5	88.4	9.0	0.6	9.7%	9.7%
IS1001d	4	516.7	72.3	9.3	2.2	16.2%	15.9%
IS1005a	4	539.0	46.4	15.2	4.0	12.2%	14.2%
IS1007b	4	940.4	147.9	40.6	7.9	20.9%	21.3%
IS1007c	4	1549.8	194.6	25.8	8.2	14.8%	15.1%
TS3006a	4	845.2	133.5	25.2	3.7	19.2%	19.0%
TS3012b	4	1918.2	287.7	26.9	1.6	16.5%	15.3%
ami-ms_dev		**11876.7**	**1553.6**	**306.8**	**67.9**	**16.2%**	**16.7%**
EN2003a	3	1583.7	132.7	9.9	0.0	9.0%	8.8%
EN2009b	3	1894.1	311.1	52.6	0.0	19.2%	18.2%
ES2008a	4	690.0	34.6	3.2	0.0	5.5%	5.6%
ES2015d	4	1485.5	310.9	84.9	34.0	29.0%	28.6%
IN1008	4	2682.3	219.4	19.1	3.4	9.0%	9.1%
IN1012	4	2649.9	611.7	107.6	13.9	27.7%	24.7%
IS1002c	4	1614.7	151.3	28.8	3.6	11.4%	12.2%
IS1008c	4	1184.7	138.1	20.8	2.4	13.6%	13.8%
IS1008b	4	1275.5	74.7	9.5	0.7	6.7%	7.0%
TS3009c	4	1884.3	385.5	98.8	14.9	26.5%	25.1%
ami-ms_eval		**16944.6**	**2370.0**	**435.1**	**72.8**	**17.0%**	**17.1%**
ami-ms		**57055.7**	**7039.7**	**1315.3**	**258.3**	**15.1%**	**15.6%**

5.2 NIST RT DATA

NIST Rich Transcription evaluation series

The alternative database to the AMI data for the experiments presented in this thesis consists of the Rich Transcription (RT) conference meeting recordings. This data was released for the RT evaluation series organized by NIST since 2002 [118]. In general, the RT evaluation encompasses more domains, such as telephone speech and broadcast news, but in the last years the focus has been directed exclusively at the meeting environment.

The meetings are held in English and recorded at various sites. The number of participants ranges from 4 to 11. Each speaker is equiped with a personal microphone, and there are several table top microphones located between the participants. For speaker diarization, for instance, NIST defines two evaluation conditions: Single Distant Microphone (SDM) and Multiple Distant Microphones (MDM). In the first, only one of the microphones located at the table is used, whereas

in the second condition, several microphones are used. The audio files are sampled at 16 kHz and samples have 16 bit depth. The reference transcriptions are derived from force-aligned annotations which are released by NIST after a particular RT evaluation campaign finishes.

Overlapping speech in NIST RT data

The duration of overlappig speech in all NIST RT sets is roughly 1.0 h, which corresponds to 10.5% of total speech duration. For particular sets the amounts are as follows, RT '05 has 8.6%, RT '06 11.1%, RT '07 10.3%, and RT '09 15.4% of overlapping speech.

We used the RT '05, RT '06, and RT '07 data for training of the overlap detection system, and the RT '09 corpus for testing. The total duration of audio is 7.4 h in so-defined joint training set and 3.0 h in RT '09 evaluation set. Table 7 gives detailed statistical properties of the 10, 9, 8, and 7 meetings of the RT '05, RT '06, RT '07, and RT '09 data, respectively.

Laughter in NIST RT data

The amount of annotated laughter in NIST RT '05, '06, '07, and '09 together is only 129.3 s and matches overlapping speech (totaling 3427.9 s) only in 22.7 s. Contrary to the situation on AMI data, laughter can be considered insignificant in this case.

Table 7: Statistics of NIST RT conference meeting recordings. Number of speakers (#Spks); duration of speech, two- (OV-2), three- (OV-3), and four-speaker overlap (OV-4) in (s); overlapping speech as a portion of all speech (OV), and overlapping speaker time as a portion of all speaker time (OV-SPKT) in (%).

MEETING	#SPKS	SPEECH	OV-2	OV-3	OV-4	OV	OV-SPKT
AMI_20041210-1052	4	557.9	14.5	0.1	0.0	2.6%	2.6%
AMI_20050204-1206	4	524.8	34.0	3.3	0.1	7.1%	7.3%
CMU_20050228-1615	4	586.5	79.9	13.8	1.1	16.2%	15.9%
CMU_20050301-1415	4	544.0	48.2	3.1	0.0	9.4%	9.1%
ICSI_20010531-1030	5	548.0	31.4	0.7	0.0	5.9%	5.6%
ICSI_20011113-1100	9	568.0	73.1	9.0	1.2	14.7%	14.4%
NIST_20050412-1303	6	494.5	70.7	6.1	0.2	15.6%	14.5%
NIST_20050427-0939	4	542.7	37.2	1.0	0.0	7.0%	6.7%
VT_20050304-1300	5	563.6	4.8	0.0	0.0	0.8%	0.8%
VT_20050318-1430	5	450.1	26.1	3.4	0.0	6.6%	6.8%
RT '05 Conf.		**5380.1**	**419.9**	**40.6**	**2.6**	**8.6%**	**8.7%**
CMU_20050912-0900	4	884.5	152.7	12.4	0.5	18.7%	16.8%
CMU_20050914-0900	4	837.9	122.1	8.8	0.3	15.6%	14.4%
EDI_20050216-1051	4	767.8	58.2	4.1	1.0	8.2%	8.3%
EDI_20050218-0900	4	809.6	68.9	9.8	1.1	9.9%	10.2%
NIST_20051024-0930	9	886.3	83.7	6.6	1.1	10.3%	10.2%
NIST_20051102-1323	8	839.6	63.1	5.2	1.1	8.3%	8.4%
TNO_20041103-1130	4	794.9	61.0	4.3	0.0	8.2%	8.1%

Continued on next page

Table 7—continued from previous page

MEETING	#SPKS	SPEECH	OV-2	OV-3	OV-4	OV	OV-SPKT
VT_20050623-1400	5	799.6	101.9	15.8	1.3	14.9%	14.7%
VT_20051027-1400	4	659.5	24.3	1.4	0.0	3.9%	3.9%
RT '06 Conf.		**7279.8**	**736.0**	**68.2**	**6.4**	**11.1%**	**10.9%**
CMU_20061115-1030	4	1100.5	170.0	8.9	0.1	16.3%	14.6%
CMU_20061115-1530	4	1030.6	93.0	3.6	0.0	9.4%	8.9%
EDI_20061113-1500	4	1094.9	170.5	27.9	0.9	18.2%	17.3%
EDI_20061114-1500	4	964.7	55.0	1.4	0.0	5.8%	5.6%
NIST_20051104-1515	4	1054.9	104.5	3.2	0.2	10.2%	9.6%
NIST_20060216-1347	6	1053.5	64.3	5.9	0.5	6.7%	6.9%
VT_20050408-1500	5	1023.8	20.0	0.6	0.0	2.0%	2.0%
VT_20050425-1000	4	1031.3	121.5	9.3	0.0	12.7%	12.0%
RT '07 Conf.		**8354.2**	**798.9**	**60.6**	**1.7**	**10.3%**	**10.0%**
EDI_20071128-1000	4	1355.4	108.2	6.1	0.0	8.4%	8.2%
EDI_20071128-1500	4	1266.9	178.0	11.1	0.2	14.9%	13.7%
IDI_20090128-1600	4	1615.7	163.1	13.4	0.9	11.0%	10.7%
IDI_20090129-1000	4	1366.9	124.8	8.9	0.3	9.8%	9.5%
NIST_20080201-1405	5	1088.7	302.6	76.2	6.8	35.4%	30.6%
NIST_20080227-1501	6	1021.4	183.9	30.1	2.6	21.2%	19.9%
NIST_20080307-0955	11	1121.1	119.4	19.0	2.5	12.6%	12.9%
RT '09 Conf.		**8836.1**	**1179.9**	**164.7**	**13.4**	**15.4%**	**15.0%**

6 OVERLAP DETECTION EXPERIMENTAL RESULTS

The system for speaker overlap detection was introduced in Chapter 3, where also various features which may act contributory to this objective were discussed. In addition, Chapter 5 presented the audio databases which are used for conducting experiments, leaving only the results yet to be shown. This chapter completes the topic on the detection of overlapping speech by demonstrating the performance of different systems.

In the first part, a subset of spectral and temporal features is selected in order to define a baseline overlap detection system. After the definition of the baseline, experiments with the application of the novel spatial features are discussed. First, the proposed microphone-data fusion strategies are evaluated, and then different combinations of the three spatial parameters are analyzed. Afterwards, the focus is dedicated to the application of prosodic features, starting with the selection of an optimal number of parameters. Finally, some remarks are given about the relationship between detected overlapping speech and laughter which is present in the recordings.

Remember that the detector has a tunable parameter called Overlap Insertion Penalty (OIP) which influences the amount of overlapping speech the system will hypothesize. The dilemma which OIP value to chose for comparing different systems, or according to which performance metric to optimize this parameter, was solved as follows. Detection results are typically presented for four OIP values empirically selected based on development data experiments, accounting for hypotheses with high recall (OIP 0, no penalization), F-ratio (OIP −10), low detection error (OIP −50), and acceptable high precision (OIP −100). The extreme option of using a lot of OIPs with high resolution is normally unnecessarily time- and computationally demanding.

Overlap detection performance is typically measured at four OIPs

6.1 DEFINITION OF THE BASELINE OVERLAP DETECTION SYSTEM

The purpose of a baseline system is to establish a reference to compare the performance improvement, or in some cases also decline, by the newly proposed techniques. There are a lot of parameters which can be tuned in order to achieve the best system possible, e. g., number of HMM states, number of GMM components, extraction of features, etc. In practice, some of these parameters are not completely tuning-independent and a full grid search is neither computationally feasible nor actually necessary for the scope of this work. Therefore, some

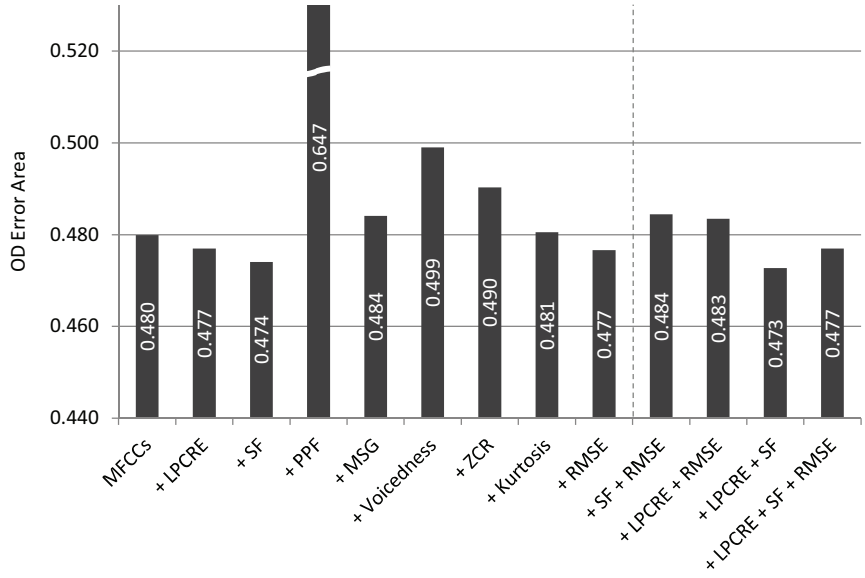


Figure 26: Selection of baseline system features for overlap detection, OD error area for detection performance on AMI single-site development data. All feature setups, except modulation spectrogram parameters, include their first-order derivatives (deltas).

of the system parameters are fixed empirically based on previous experiences.

From the set of spectral and temporal features discussed in Section 3.2 the aim is to select a subset which will compose the baseline overlap detection system. The selection strategy in this case is rather naive. Each of the candidate features is combined individually with 12 MFCCs, specific models are trained, and then they are tested on AMI development data. In addition, all feature vectors are mean-variance normalized according to statistics obtained on training data, and normally first-order derivatives are added.

Candidate features are compared according to OD error under their ROC curves

In order to evaluate which feature actually contributes to overlap detection, we suggest to calculate the OD error area under their appertaining ROC curves. This concept was previously explained in Section 3.7. Figure 26 gives these OD error areas for several combinations of MFCCs with other candidate features. In the first part of the table, it can be seen that only adding LPCRE, SF, and RMSE reduced the error area compared to MFCCs only. However, the truth is that the majority of values are not very different.

It is interesting to note that this results are only vaguely correlated with the F_{ratio} and KL2 divergences from Table 1. The spectral flatness parameter, for instance, obtained low preliminary F_{ratio} and KL2 scores, but in combination with mel cepstrum it shows to be performing well in the actual overlap detection. On the contrary, parameters such as voicedness or PPF, which had average preliminary scores, either do not

convince in the real use or even completely fail (PPF). This suggests that one cannot rely too much on measures such as F_{ratio} and KL2 divergence for assessing discriminability properties of parameters as far as real detection performance is concerned.

In the right part of Figure 26 there are OD error areas for further combinations of LPCRE, SF, and RMSE. Sometimes two parameters are not well-compatible and their joint performance is bellow their individual ones. The lowest value is achieved by the combination of MFCCs with LPCRE and SF. Consequently, the overlap detection system using this feature combination is considered the best in this context, and in the following it will be referred to as the *baseline* or the *spectral* system. Note, for the sake of clarity, that the length of feature vectors in the baseline system including the deltas is 28. The performance of the baseline system will be discussed in the next sections.

Baseline overlap detection system uses 12 MFCCs, LPC residual energy, and spectral flatness

As was explained in Section 3.6, the number of GMM components in overlapping speech model is 64. Nevertheless, during the development of the system, experiments were also performed with other numbers of Gaussian components (e. g., 32, 128). In our experience, the more Gaussian components are used in overlapping speech model, the higher overlap recall has the detection system, but with a lower precision. In this regard, 64 components seems like a reasonable choice.

From the point of view of speech/non-speech discrimination, the overlap detection system has a tendency to generate more missed speech errors than false alarms. The effect on detected overlapping speech is in fact almost negligible regarding overlap recall, almost all the missed speech affects only the single speech class. Regarding false overlap error, such SAD operation is responsible for an increase of less than 1%.

Overlapping speech in the context of SAD

6.2 APPLICATION OF SPATIAL INFORMATION

In this section our attention turns towards spatial features which were proposed in Section 3.3, namely spatial coherence, dispersion ratio, and delta TDOA computed for every microphone pair. Furthermore, Section 3.4 suggested different strategies in order to deal with the high, and possibly variable, dimensionality of the spatial feature space. In the following, their application will be evaluated in single- and multi-site scenarios and compared to the baseline system relying on spectral features only.

When combining spectral and spatial features, the two feature streams are considered to be statistically independent. The joint emission probability is obtained by weighing the streams with weights

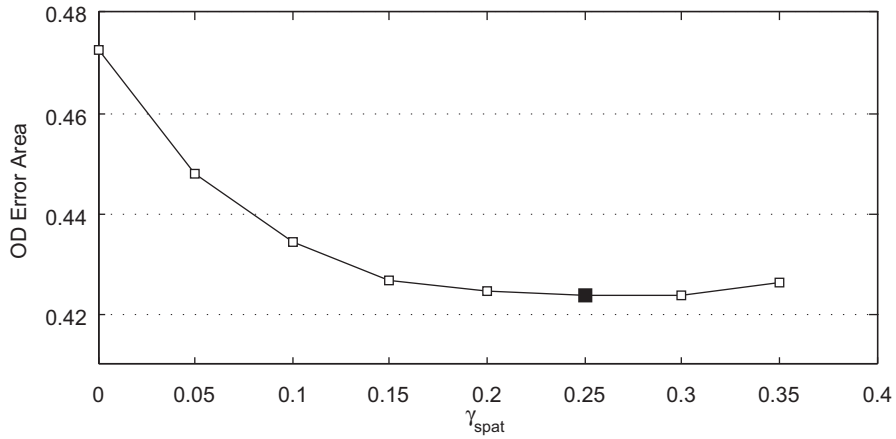


Figure 27: Overlap detection error area for different values of spatial stream weight γ_{spat}. Weight 0 means that spatial features were not used. Experiments conducted on AMI single-site development data.

γ_{spct} and γ_{spat}, respectively, where $\gamma_{spct} + \gamma_{spat} = 1$. For any given frame, the log-likelihood is computed as:

$$\begin{aligned} &L(y_{spct}, y_{spat}|\lambda_{spct}, \lambda_{spat}) \\ &= \gamma_{spct}\, L(y_{spct}|\lambda_{spct}) + \gamma_{spat}\, L(y_{spat}|\lambda_{spat}), \end{aligned} \tag{6.1}$$

where λ_{spct}, y_{spct} is the spectral model and data, and λ_{spat}, y_{spat} is the spatial model and data.

Spatial feature stream weight optimization

A series of experiments was conducted on AMI development data in order to determine the optimal weight values. The obtained results in terms of OD error area under ROC curves are given in Figure 27. Based on this graph the values $\gamma_{spct} = 0.75$ and $\gamma_{spat} = 0.25$ were chosen for further experiments. Although the optimization was done for PCA-transformed spatial parameters, these weights are also applied in case of the other fusion strategies.

6.2.1 *Comparison of Fusion Strategies*

Figure 28 demonstrates the overlap detection performance of the baseline spectral system (*Spct*) and of three systems also employing spatial information on AMI evaluation data. The performance is given in terms of recall, precision, and overlap detection error. The spatial systems are as follows:

- *Spct+Spat PCA* — System combining spectral and PCA-transformed spatial coherence, dispersion ratio, and delta TDOA.

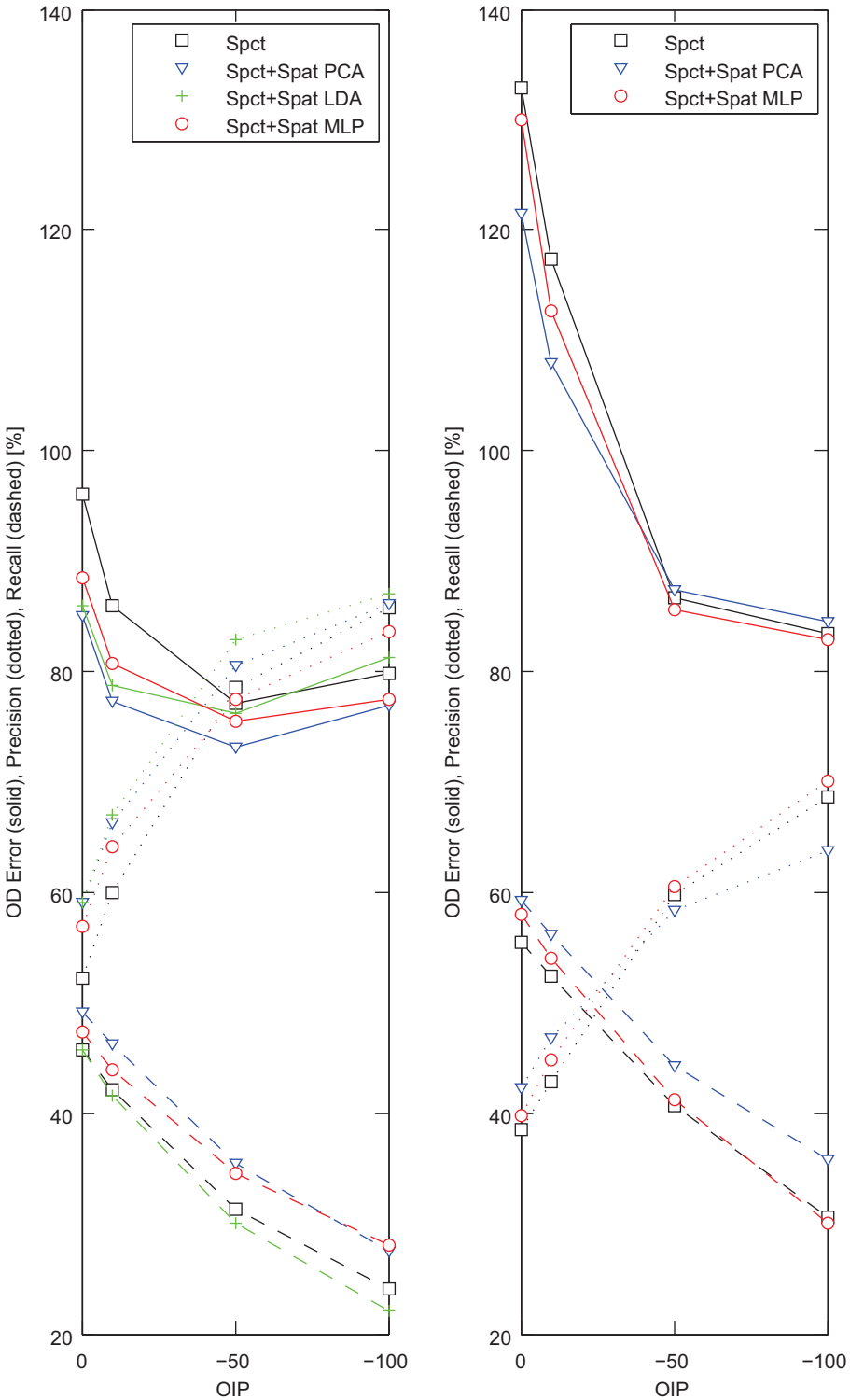


Figure 28: Overlap detection performance for AMI (a) single- and (b) multi-site evaluation data using spectral features alone (*Spct*) or in combination with PCA- (*Spct+Spat PCA*), LDA-transformed spatial features (*Spct+Spat LDA*), or spatial MLP score (*Spct+Spat MLP*). Detection error, precision, and recall are delineated with solid, dotted, and dashed line, respectively.

- *Spct+Spat LDA* — Three spatial parameters for all microphone pairs are projected using LDA and combined with baseline spectral features (single-site scenario only[1]).
- *Spct+Spat MLP* — This system combines the spectral features with a classification score of an MLP based on spatial features.

Comparison of microphone-data fusion strategies

The results on single-site recordings (Figure 28 (a)) show that in the lower penalty region the spatial systems outperform the spectral in all three evaluation metrics, just with the exception of *Spct+Spat LDA* recall which is similar to the one of the baseline system. Even though the differences among detection errors and precisions for higher penalization are becoming smaller, the *Spct+Spat PCA* system continues with a performance better than the baseline. It also achieves the lowest error of 73%, which corresponds to a precision of 80% and a recall of 35% at OIP −50. The system using MLP score maintains good detection error for high OIPs, but its precision drops bellow the one of the baseline system. The *Spct+Spat LDA*, on the contrary, falls at the end with the detection error behind the *Spct*, but it exhibits the highest precision in all experiments.

Multi-site scenario results presented in Figure 28 (b) show an overall degradation of the detection performance compared to the single-site data. Nevertheless, the performance patterns in the low penalization region are similar. Here, the spatial PCA system seems to be the best at low OIPs in all metrics. With increasing penalization, however, the detection errors get almost alike, and even though *Spct+Spat PCA* maintains the highest recall, in terms of precision it is overrun by both *Spct* and *Spct+Spat MLP* system. The latter also achieves the best result with 83% detection error, 70% precision and 30% recall at OIP −100. The results of the baseline system are very similar in this case. *Spct+Spat PCA*, on the other hand, yields at this point an error of 85%, precision of only 64%, and recall of 36%. Obviously, the less precise multi-site models need a higher amount of overlap penalization to arrive to the lowest detection errors. A possible explanation for the relatively lower precision of the spatial PCA system at the lowest error operating point (in comparison with the other two system) could be that PCA is a too simple technique to compensate for the variability of the multi-site scenario. These results were also presented in [129]. All numerical values are given in Table 8 at the end of this chapter.

6.2.2 *Comparison of Spatial Parameter Combinations*

Since the PCA fusion strategy has shown good potential, especially on single-site data, we decided to investigate more on the individual

1 LDA fusion strategy was added as the last one at the end of the work, for that reason results are available only for single-site data.

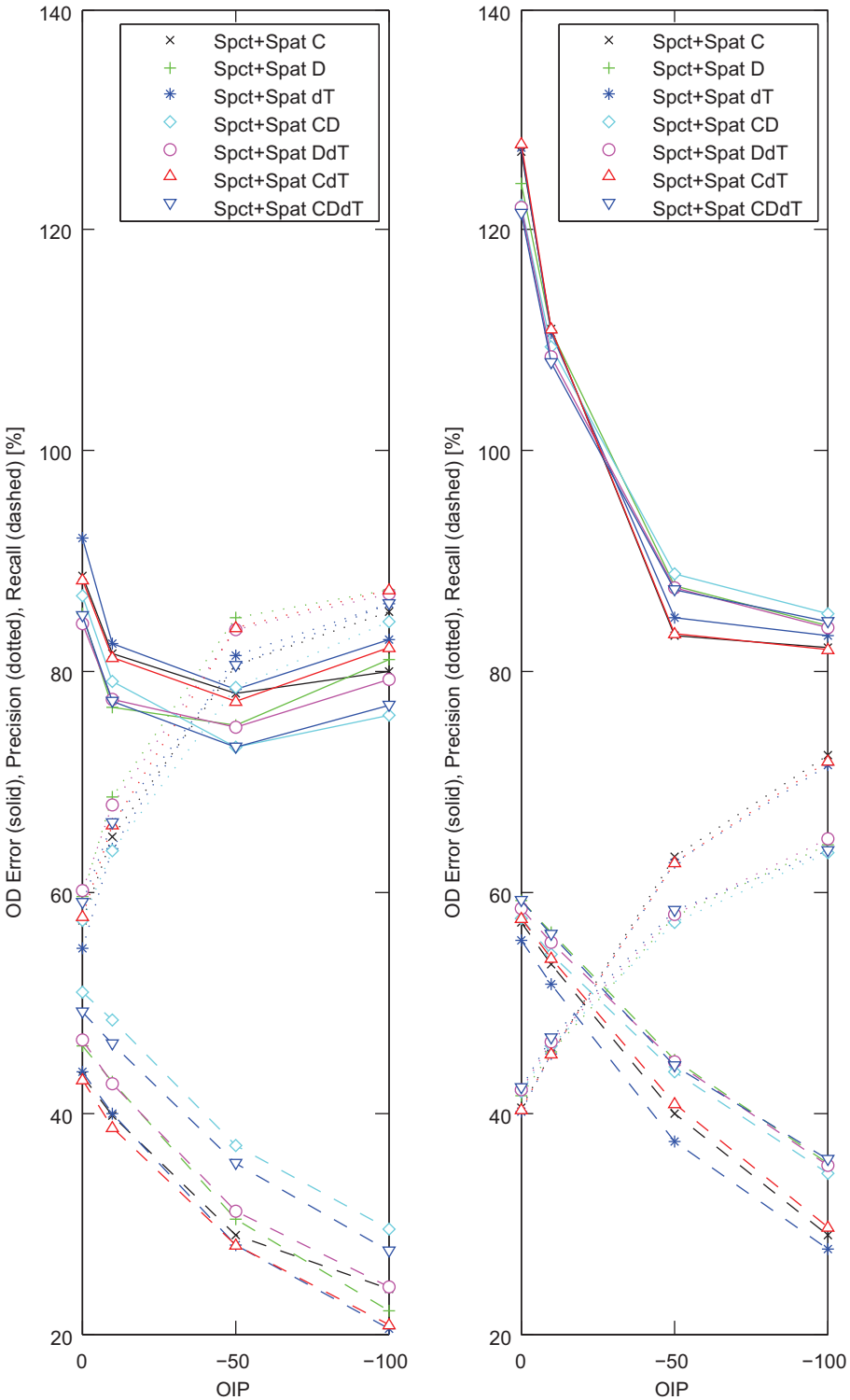


Figure 29: Overlap detection performance for AMI (a) single- and (b) multi-site evaluation data using different combinations of spectral features (*Spct*) and PCA-transformed spatial coherence (*Spat C*), dispersion (*Spat D*), and delta TDOA (*Spat dT*). Detection error, precision, and recall are delineated with solid, dotted, and dashed line, respectively.

importance of the three spatial parameters in the given system. We try different combinations of spatial coherence (*Spat C*), dispersion (*Spat D*), and delta TDOA (*Spat dT*) in order to have a better understanding of the contribution of each of these parameters. Some of the following results were published in [130].

Different combinations of spatial parameters

Besides the already presented setup involving all three PCA-fused parameters (*Spct+Spat PCA* ≡ *Spct+Spat CDdT*), there are six other possible feature setups (*Spct+Spat C*, *Spct+Spat D*, etc.). The overlap detection performance of all seven setups on single-site recordings is given in Figure 29 (a). We can observe that in terms of detection error the *Spct+Spat CDdT* feature setup is performing well together with *Spct+Spat D* and *Spct+Spat DdT* at low OIPs, and *Spct+Spat CD* at high OIPs. The lowest error 73% was obtained by both *Spct+Spat CDdT* and *Spct+Spat CD* at OIP of −50. In case of *Spct+Spat CD* it corresponds to a precision and recall of 79% and 37%, respectively (*Spct+Spat CDdT* results were given before). Note that what all these setups have in common is the spatial dispersion ratio parameter. *Spct+Spat CD* and *Spct+Spat CDdT* are yielding the highest recall values but the former setup at the cost of the lowest precision. Here, the combinations *Spct+Spat D* and *Spct+Spat DdT*, and later also *Spct+ Spat CdT* are the better ones as far as precision is concerned. For instance, the precision of *Spct+Spat D* at OIP −50 is 85% and at OIP −100 it increases to 87%. Interesting is the relatively worse performance of setups with either *Spat C* or *Spat dT* parameter alone and, maybe except the mentioned high penalty precision, also with both of them together.

The situation in multi-site scenario illustrated in Figure 29 (b) is different to single-site data, particularly regarding the detection error. The difference is actually twofold. Not only are the absolute numbers significantly worse as was commented before, but the performance pattern of the feature setups changes with increasing penalization. The relative error positioning of the setups at OIP 0 is to a certain extent similar to the single-site data, i. e., the setups including *Spat D* are slightly better than the others. The values of recall and precision are scattered in a smaller interval. However, with penalization −10 the error detection values are coming closer to each other, and at OIP −50 we can observe a clear switch of the error performance between the *Spat D* setups and the rest. This event is even more visible by looking at the precision as well as recall lines. Despite the fact that for the highest penalization the detection errors start to converge again, setups *Spct+Spat C*, *Spct+Spat dT*, and *Spct+Spat CdT* maintain lower error and higher precision. In fact, the gap in precision becomes even wider. Even though it is not directly depicted, their precision is actually higher than the one of the baseline spectral system, and the detection error is lower or equal. The lowest error of 82% was achieved by both *Spct+Spat C* and *Spct+Spat CdT* setup (OIP −100) with the same

Distinct behavior of spatial dispersion parameter in singe- and multi-site conditions

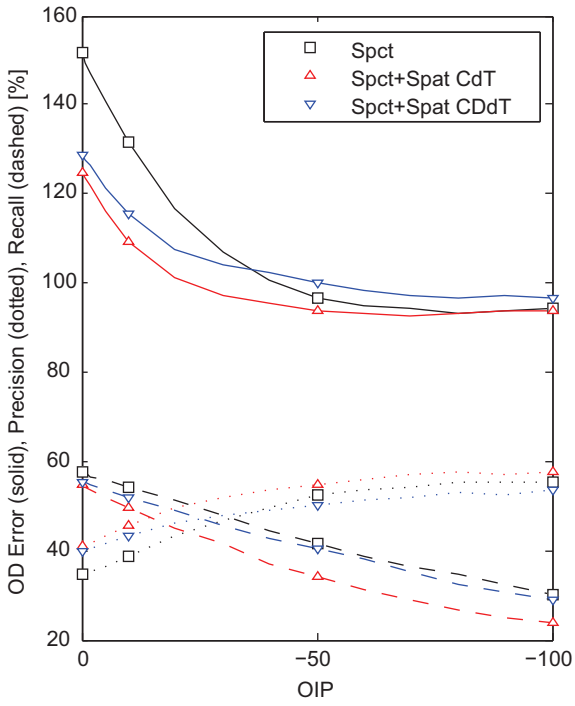


Figure 30: Overlap detection performance for NIST RT '09 data. Feature setups are as follows, spectral (*Spct*), spectral with spatial coherence and delta TDOA (*Spct+Spat CdT*), and spectral with all three spatial parameters (*Spct+Spat CDdT*). Detection error is delineated with solid line, precision with dotted line and recall with dashed line. The four predefined OIP values are marked, but the performance was tested with more penalties.

precision of 72% and a recall of 29% and 30%, respectively. It seems that when spatial dispersion is used, the systems are more prone to hypothesize a higher amount of overlapping speech. Since the multi-site scenario is more challenging and the models obviously less precise, this behavior can turn to be eventually problematic.

Another possible reason for the worse performance of feature setups involving the *Spat D* parameter, besides the simplicity of PCA mentioned earlier, is the fact that this parameter may be closely dependent on the spatial distribution of microphones in a room. Such dependency would most probably result in a lower robustness in multiple room scenarios.

Overlapping speech detection with spatial features on NIST RT '09 data

The difficulty to detect simultaneous speech on data originating from various rooms is even more visible in Figure 30, where the overlap detection results on NIST RT '09 are given. This data comprises recordings from three sites. The recordings from previous RT evaluations used for training were collected from six different sites. We decided to build overlap detection models for three feature setups which showed to be the most interesting on AMI data in Figure 29, i. e.,

Spct, *Spct+Spat CdT*, and *Spct+Spat CDdT*. This time the performance was tested for more penalizations as usual.

The *Spct+Spat CdT* system maintains the highest precision and lowest detection error, but the absolute numbers are worse than the AMI multi-site results (see Figure 29 (b)). As far as the detection error is concerned, the *Spct* system reduces the performance gap to *Spct+Spat CdT* with increased penalization. Both systems yield the same lowest detection error of 93%, *Spct+Spat CdT* at penalization −70 achieving 57% precision and 29% recall, and *Spct* at OIP −80 with a lower precision of 55% but a higher recall of 35%. These numbers are not particularly good. Nevertheless, to the knowledge of the author much better results for the detection of overlapping speech on NIST RT meeting recordings have not been published [107, 117][2].

To summarize, the combination of spectral and spatial parameters improves the detection of overlapping speech compared to baseline system, more significantly for the low penalization values. When spatial coherence, dispersion, and delta TDOA estimates are fused by means of a PCA, it was observed that the application of dispersion ratio is very beneficial for single room use, but in case of multiple recording rooms it can result in a lower precision of the detected simultaneous speech.

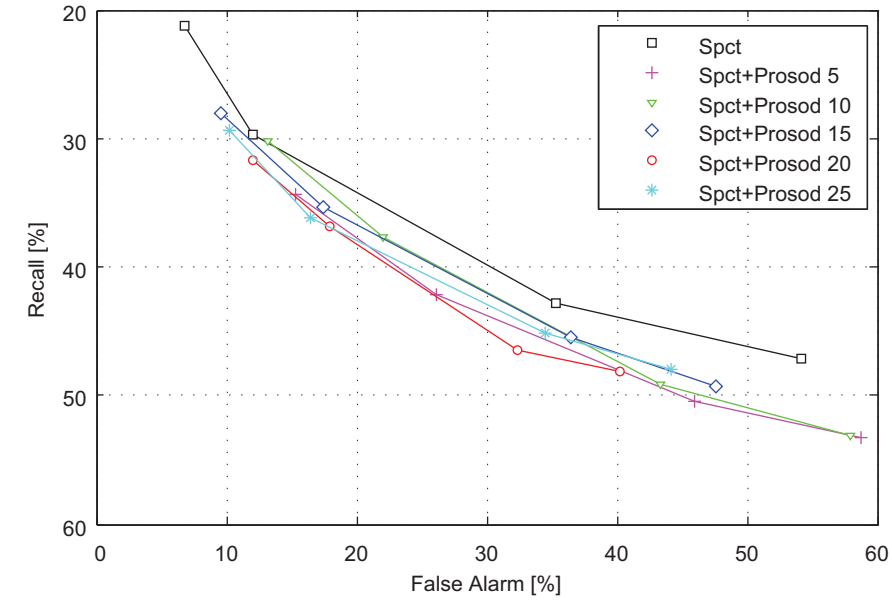
6.3 APPLICATION OF PROSODIC INFORMATION

In Section 3.5 a set of candidate prosodic features was introduced and a two-stage feature selection mechanism was outlined. After applying the mRMR algorithm, the candidate features were scored and sorted accordingly (see Table 2). In this section the results of the second feature-selection stage are presented, together with the performance of the overlap detection system using the selected optimal number of prosodic features.

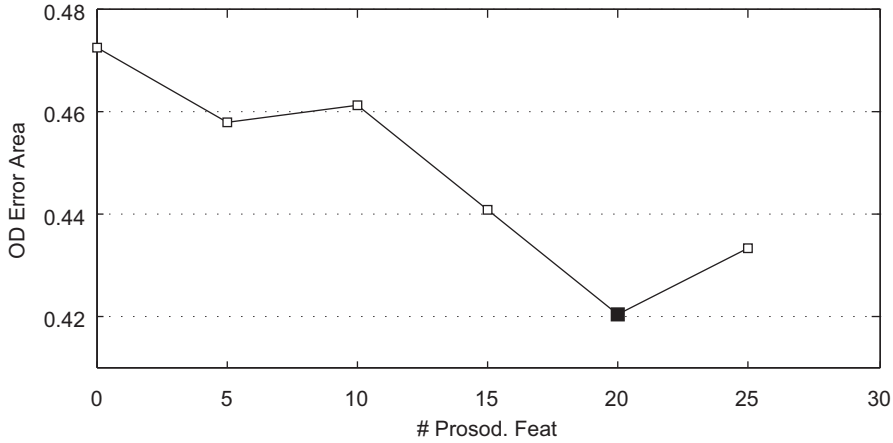
Selection of the optimal number of prosodic features

Following the hill-climbing wrapper strategy, the baseline spectral features were combined with the first 5, 10, 15, etc. candidate prosodic features from Table 2. New models were trained and then tested on AMI development data. The performance in terms of ROC curve is given in Figure 31 (a). It can be seen that the systems with prosodic features achieve lower error, especially for low penalization values when compared to the spectral-only system. However, it is not easy to decide from this plot what number of prosodic features is the optimal value. Similarly to the selection of the baseline features, the area under the ROC curves is used as a decision factor. The resulting OD error area values for the considered numbers of prosodic features are given

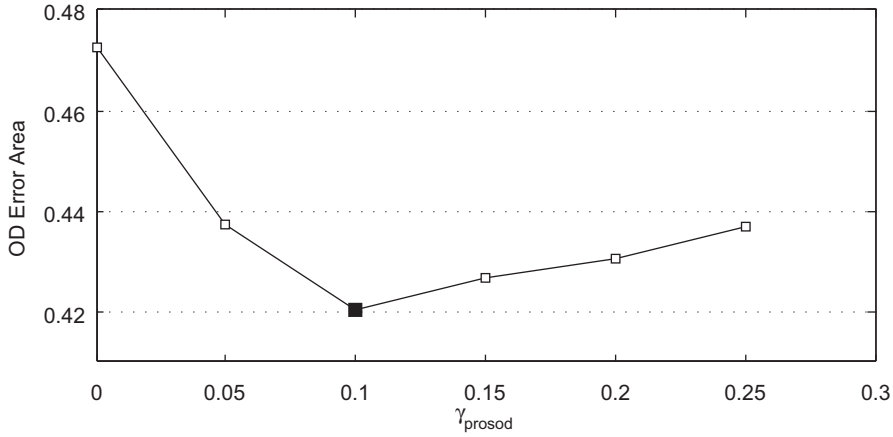
2 Although Huijbregts et al. [117] does not specifically give results for overlap detection, an approximate notion can be deduced from the differences between misses and between false alarms of the diarization baseline and the diarization handling overlaps.



(a)



(b)



(c)

Figure 31: Overlap detection performance for (a) (b) different numbers of selected prosodic features and (c) for different values of prosodic stream weight γ_{prosod} when 20 features are selected. Performance given in terms of (a) ROC curves and (b) (c) OD error area. Experiments conducted on AMI single-site development data.

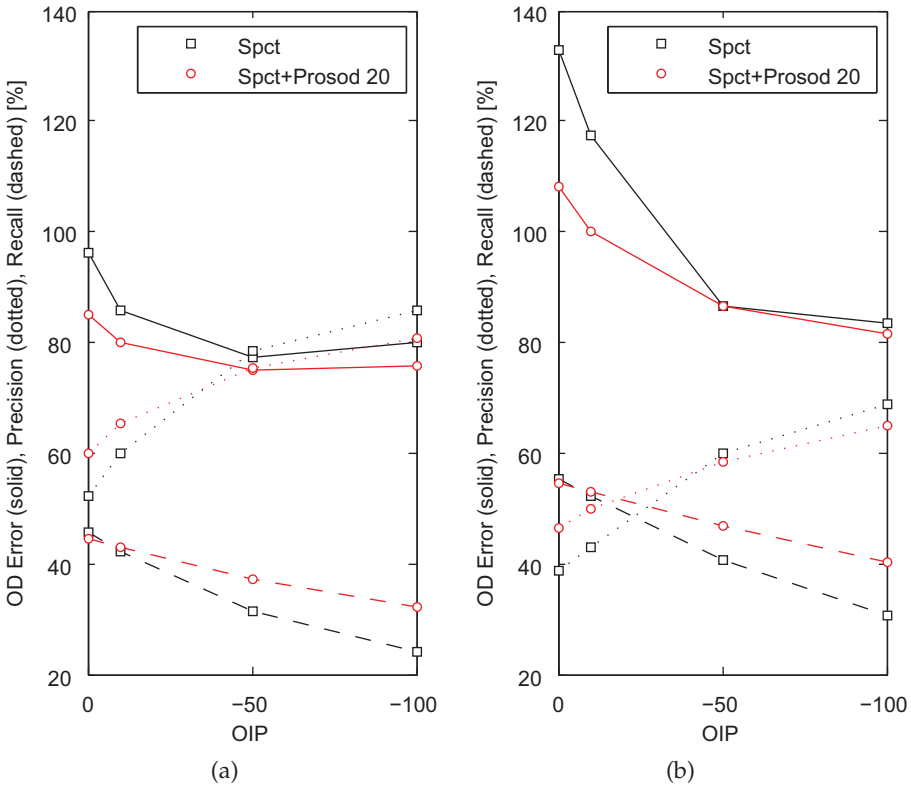


Figure 32: Overlap detection performance for AMI (a) single- and (b) multi-site evaluation data using spectral features only (*Spct*), and the combination of spectral and 20 prosodic features (*Spct+Prosod 20*) in terms of detection error (solid line), precision (dotted line) and recall (dashed line).

in Figure 31 (b). Based on this graph the first 20 features from the candidate set are picked as the optimal number [136].

Optimization of prosodic feature stream weight

The strategy for fusion of the spectral with the prosodic information is basically the same as the one applied with spatial features. The two feature streams are considered statistically independent and similarly to (6.1) the output HMM probability is obtained by weighting particular streams with weights γ_{spct} and γ_{prosod}, while $\gamma_{\mathrm{spct}} + \gamma_{\mathrm{prosod}} = 1$. The final weights are tuned on AMI development data in the same way as the number of prosodic features. The OD error areas for a range of examined γ_{prosod} values is depicted in Figure 31 (c), where a local minimum appears at $\gamma_{\mathrm{prosod}} = 0.1$. Hence, this value was eventually selected.

Comparison of the system using prosodic features with the baseline

The comparison of the baseline spectral and combined spectro-prosodic system on AMI single- and multi-site evaluation data is presented in Figure 32 (a) and (b), respectively. Detection performance is given in terms of recall, precision, and detection error. In single-site scenario the combined-feature system (*Spct+Prosod 20*) outperforms

the spectral (*Spct*) in terms of error for all OIPs, with the lowest value of 75% at OIP −50. On the other hand, the situation is not so unequivocal with precision. The precision does not rise so steeply with increasing OIP in the new system. At the highest OIP of −100, the precision of the *Spct+Prosod 20* system is 81% while the baseline system yields 86%. According to our experience such behavior can be probably related to the higher amount of model parameters which need to be trained in the combined system. The rest of the numerical results can be found in Table 8.

In case of multi-site overlap detection the lowest detection error of 81% is achieved by *Spct+Prosod 20* at OIP −100 with a precision of 65% and a recall of 40%. At the same penalization point the baseline system obtains 83%, 69%, and 31% of detection error, precision, and recall, respectively. Considering the first two metrics, these results indicate a worse detection performance compared to single-site condition. This observation, again, emphasizes the higher difficulty of multi-site scenario, especially with regard to the use of a single general model of overlapping speech. However, the relative behavior of the two systems is similar to the single-site case. The *Spct+Prosod 20* system outperforms the *Spct* in the low OIP region, but with increasing penalization the detection errors are basically converging and the precision of the baseline system surpasses the prosodic one.

6.4 REMARKS ON LAUGHTER

Another topic that is remaining to discuss is the behavior of our overlap detection system in relation to laughter. To repeat a bit, it was observed that the annotated laughter and overlapping speech often coincide (in AMI corpus $>$ 70% of laughter time). Laughter and speaker overlap can be correlated because people can laugh when they accidentally jump into each other's speech. Furthermore, when something funny is discussed people are prone to add their remarks instantly, without waiting for the others to finish.

It was suspected that part of the false overlaps detected by the system may have been due to the detected laughter segments. The basis for this suspicion was that when laughter is included in the training data of the overlap model, this model will be susceptible to detect laughter even when it is not coinciding with simultaneous speech—laughter and normal speech can be considered acoustically different to some extent.

Does the inclusion of laughter in the training data of the overlap model make it detect also laughter which is not occurring together with simultaneous speech?

In order to clear this doubt, oracle laughter segments were either subtracted from detected speaker overlap segments or they were joined together. Both alternatives are compared to the original detected overlapping speech and the result in terms of ROC curves is given in Figure 33. The subtraction of laughter almost did not decrease the

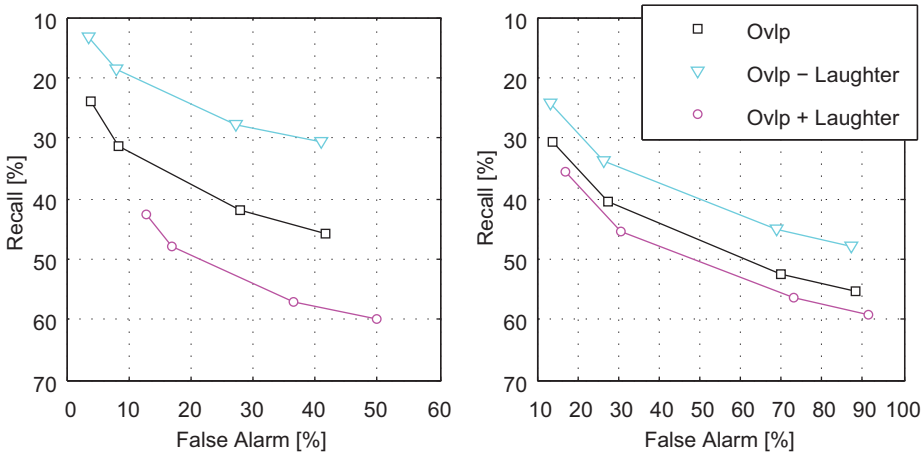


Figure 33: Overlap detection performance of the baseline spectral system when reference laughter segments are subtracted from or added to detected overlapping speech, results for (a) single- and (b) multi-site data.

FA error compared to the original results. It only decreased recall, which means that only laughter which occurred together with speaker overlap was discarded. On the other hand, the union of overlap and laughter segments increased the amount of detected overlapping speech (both true and false), to a greater extent for single-site data than for multi-site data. These observations imply that our former suspicion could not be verified and the system is not prone to detect laughter. In fact, for single-site data the contrary is true, overlapping speech associated with laughter is rather missed.

Table 8: Overlapping speech detection on AMI single- and multi-site data. Results given in terms of recall, precision, and OD error in (%) for four OIPs.

		SINGLE-SITE			MULTI-SITE		
OIP	OVERLAP DET.	RCL	PRC	ERR	RCL	PRC	ERR
0	Spct	45.7	52.2	96.1	55.4	38.6	132.8
	Spct+Spat (PCA) C	43.3	57.5	88.7	57.3	40.4	127.1
	Spct+Spat (PCA) D	46.1	59.6	85.2	59.3	41.5	124.3
	Spct+Spat (PCA) dT	43.8	55.0	92.1	55.6	40.1	127.5
	Spct+Spat (PCA) CD	50.9	57.4	86.9	57.6	41.9	122.2
	Spct+Spat (PCA) CdT	43.0	57.9	88.3	57.7	40.3	127.8
	Spct+Spat (PCA) DdT	46.6	60.1	84.3	58.6	42.1	122.1
	Spct+Spat (PCA) CDdT	49.2	59.0	85.0	59.3	42.3	121.5
	Spct+Spat LDA	45.8	59.1	85.9			
	Spct+Spat MLP	47.3	57.0	88.4	58.0	39.8	129.9
	Spct+Prosod 20	44.4	60.0	85.2	54.6	46.5	108.1
-10	Spct	42.1	60.1	85.9	52.4	42.9	117.3
	Spct+Spat (PCA) C	39.7	65.0	81.7	53.5	45.3	111.1
	Spct+Spat (PCA) D	42.9	68.6	76.7	56.3	45.6	110.8
	Spct+Spat (PCA) dT	40.0	63.9	82.6	51.6	45.3	110.6
	Spct+Spat (PCA) CD	48.4	63.8	79.1	54.4	46.0	109.4
	Spct+Spat (PCA) CdT	38.7	66.0	81.2	54.1	45.4	111.0
	Spct+Spat (PCA) DdT	42.7	68.0	77.4	55.4	46.4	108.6
	Spct+Spat (PCA) CDdT	46.3	66.2	77.3	56.1	46.7	107.9
	Spct+Spat LDA	41.6	67.1	78.8			
	Spct+Spat MLP	43.9	64.1	80.7	54.1	44.8	112.6
	Spct+Prosod 20	42.8	65.2	80.0	52.8	50.0	99.9
-50	Spct	31.3	78.6	77.2	40.7	59.9	86.6
	Spct+Spat (PCA) C	29.0	80.6	78.0	39.9	63.2	83.3
	Spct+Spat (PCA) D	30.3	84.9	75.1	44.9	57.9	87.8
	Spct+Spat (PCA) dT	28.0	81.4	78.4	37.4	62.8	84.8
	Spct+Spat (PCA) CD	37.0	78.6	73.1	43.7	57.3	88.8
	Spct+Spat (PCA) CdT	28.1	83.9	77.3	40.8	62.7	83.5
	Spct+Spat (PCA) DdT	31.1	83.8	74.9	44.6	58.1	87.6
	Spct+Spat (PCA) CDdT	35.4	80.5	73.2	44.2	58.3	87.4

Continued on next page

Table 8—continued from previous page

OIP	OVERLAP DET.	RCL	PRC	ERR	RCL	PRC	ERR
	Spct+Spat LDA	30.0	82.9	76.2			
	Spct+Spat MLP	34.5	77.5	75.5	41.2	60.6	85.6
	Spct+Prosod 20	37.3	75.5	74.8	46.7	58.5	86.4
-100	Spct	24.1	85.8	79.9	30.5	68.7	83.4
	Spct+Spat (PCA) C	24.1	85.5	80.0	28.9	72.4	82.1
	Spct+Spat (PCA) D	22.1	87.4	81.1	35.5	64.3	84.2
	Spct+Spat (PCA) dT	20.4	86.1	82.9	27.7	71.6	83.3
	Spct+Spat (PCA) CD	29.4	84.5	76.0	34.6	63.6	85.2
	Spct+Spat (PCA) CdT	20.9	87.4	82.1	29.6	71.8	82.0
	Spct+Spat (PCA) DdT	24.3	87.1	79.3	35.3	64.8	83.9
	Spct+Spat (PCA) CDdT	27.5	86.2	76.9	35.7	63.8	84.6
	Spct+Spat LDA	22.1	87.0	81.2			
	Spct+Spat MLP	28.1	83.6	77.4	30.0	70.1	82.8
	Spct+Prosod 20	32.1	80.7	75.6	40.2	65.0	81.4

7 SPEAKER DIARIZATION EXPERIMENTAL RESULTS

Previous chapter was dedicated to the evaluation of overlap detection systems. Here in this chapter, the segments of simultaneous speech which were detected by these systems are employed in order to reduce the error of speaker diarization. The resulting improvements are compared among each other.

Firstly, we try to establish a relationship between the diarization improvement and the operation of overlap detection system, which is controlled with a penalization parameter. Relative DER reduction is plotted as a function of the OIP used for the detection of simultaneous speech segments.

Then, we discuss overlap handling experiments on evaluation data with the baseline diarization system. These include the application of overlap exclusion, labeling, and both techniques together. In the last section, a subset of the previous experiments is repeated, but in this case the baseline diarization system operating on single distant microphone is switched for a diarization system which makes use of multiple distant microphones. Finally, an analysis on the diarization performance on individual meeting recordings is given.

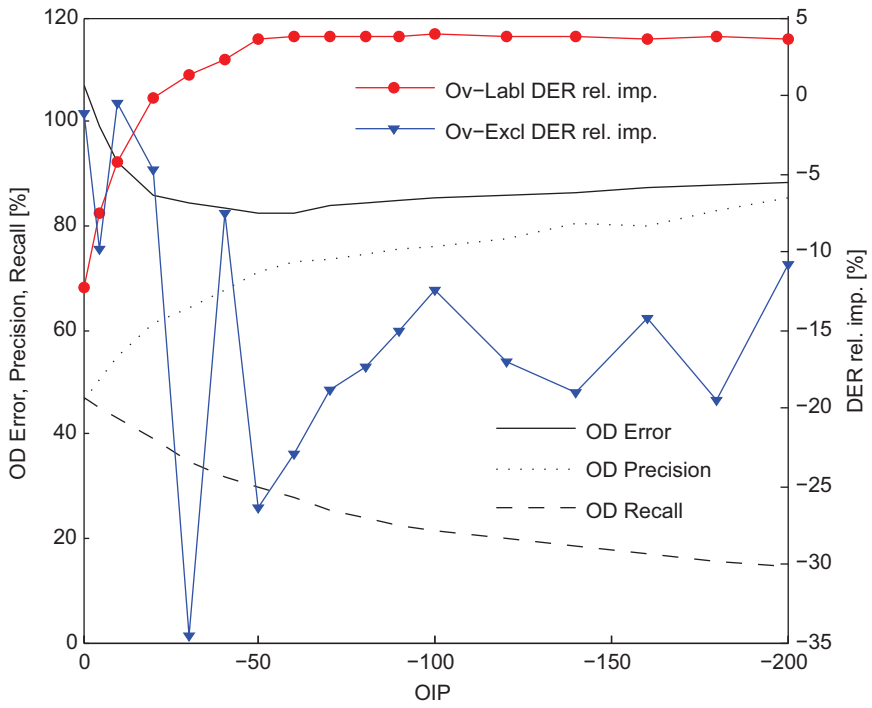
It is worth mentioning that in order to evaluate just the effect of overlapping speech on speaker diarization, detected overlaps are normally masked with reference speech/non-speech segments before given to diarization system. However, in experiments involving a real SAD system (Section 7.3.1), proper SAD hypotheses are used instead of the reference annotations.

7.1 OVERLAP DETECTION VS. DIARIZATION IMPROVEMENT RELATIONSHIP

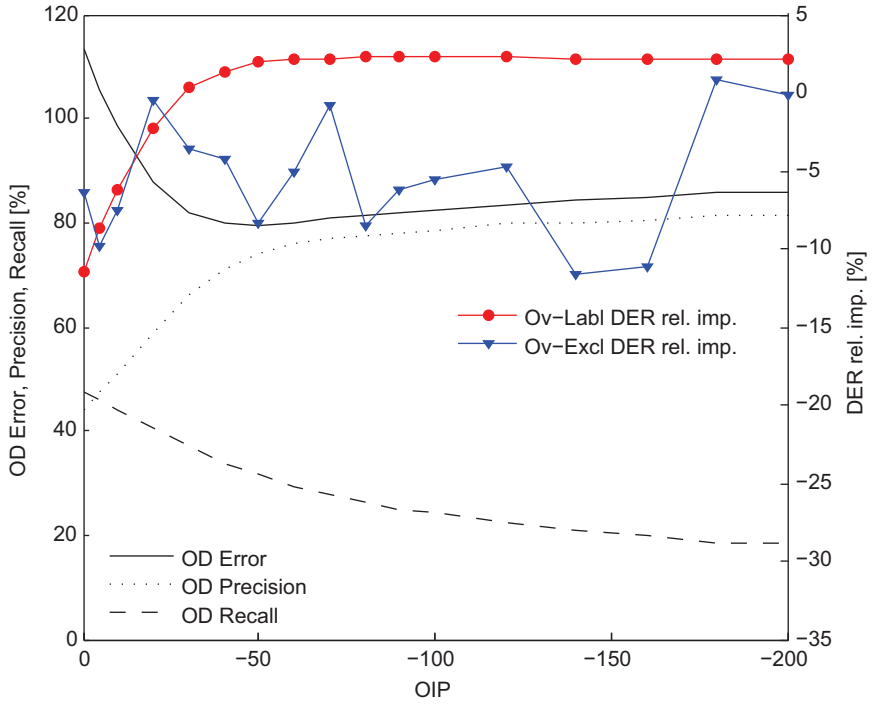
Complement of the overlap detection error marks the maximal possible gain by labeling

The complement of the overlap detection error tells us how much the diarization can possibly gain with labeling when using a particular overlap hypothesis. All of the overlap false positives will be propagated to the DER, but only a perfect labeling would transform all true positives into a reduction of missed speaker time. Sufficiently high precision is also important for obtaining good results. The relationship between DER improvements and overlap detection properties was discussed in [147, 130] and here it is depicted in Figure 34.

Overlap hypotheses, which were produced for development recordings for several OIP values with the spectral overlap detection system, were subsequently applied in the diarization system for the assign-



(a)



(b)

Figure 34: Overlap detection performance of spectral system and corresponding relative DER improvements by overlap exclusion (*Ov-Excl*) and assignment of second speaker labels to detected overlap segments (*Ov-Labl*) for AMI (a) single and (b) multi-site development data.

ment of second speaker labels. We can observe in both scenarios that the DER improvement curve has its maximum in the region of the lowest detection errors, and maintains high values with increasing OIP towards higher overlap precisions. The relative improvements are computed against baseline DER, which is 28.3% for single- and 39.5% for multi-site development data.

Similar experiments were also performed with overlap exclusion, but the system behavior in this case is not easily predictable. The relationship between DER improvements and the overlap detection performance metrics in Figures 34 (a) and (b) is not clear.

Different OIPs are applied for producing overlapping speech segments for exclusion, and for labeling

In practice, it is useful to have one overlap hypothesis for overlap exclusion and another for overlap labeling. Even though doing exclusion has influence on labeling output, we can say that these two techniques work independently and may have different requirements on overlap detection from the perspective of DER improvement. When each technique has its own hypothesis, more room is left for the optimization of the performance.

The OIP value for overlap labeling experiments on the evaluation data is fixed based on the results on the development data. For each overlap detection system a high-penalty hypothesis at OIP −100 is selected. Since it was not possible to clearly identify a successful working point for overlap exclusion, we decided to use the overlap hypotheses without penalization (OIP 0) for this technique.

7.2 EVALUATION OF OVERLAP HANDLING TECHNIQUES

7.2.1 *Application of Overlap Exclusion*

Baseline DER and relative improvements by applying overlap exclusion in experiments conducted on AMI evaluation data are given in Table 9. The relative DER improvements are presented according to overlap detection systems discussed in Chapter 6, which were used for finding segments of simultaneous speech. The results for overlap exclusion show that the most successful overlap detection setups in single-site condition are *Spct+Spat CDdT* with an improvement of 5.2% and *Spct+Spat D* with 5.1%. Their common characteristic is that both setups yield high recall and precision, and low detection error from among the zero-penalty hypotheses in Figure 29 (a). On the other hand, the exclusion of speaker overlap that originates from *Spct+Spat DdT* setup, having comparable overlap detection performance, results in a much lower error reduction (1.8%). This observation suggests that the overlap detection systems though having similar numerical performance they are not detecting exactly the same overlap segments. Besides, it also indicates that the exclusion of different segments of

overlapping speech does not have the same effect on the diarization system.

With the alternative microphone-pair fusion approaches, which included either the deployment of a pre-trained MLP (*Spct+Spat MLP*) or an LDA projection (*Spct+Spat LDA*), the achieved relative improvements are bellow the one with the spectral overlap detection system (*Spct*). The exclusion of speaker overlap detected also with the help of prosodic features is performing slightly better.

Exclusion in multi-site scenario resulted in higher relative improvements than in single-site scenario

The comparison of single- and multi-site data results shows one difference. The improvements are significantly higher for the multi-site as for the single-site recordings. Considering the PCA-transformed spatial features, high improvements by exclusion are obtained by following setups: *Spct+Spat C*, *Spct+Spat D*, *Spct+Spat dT*, and particularly by *Spct+Spat CdT* setup that yields 13.9%. Unexpected is the lower improvement with hypothesis originating from the combination of spatial coherence and dispersion (*Spct+Spat CD*), which probably affected also the *Spct+Spat CDdT* setup.

Noteworthy is the DER improvement of 12.1% relative with the *Spct+Spat MLP* overlap detection system, which is the second-best observed result. The overall high improvements are also confirmed by the combined prosodic system *Spct+Prosod 20*. The exclusion of overlaps in this case reduced the baseline DER by 9.2% relative.

From the point of view of detected number of speakers, the effect of overlap exclusion on clustering is that normally the algorithm finishes with a lower number of final clusters. In this way both the number of true and false detected speakers are decreased. In single-site scenario we can speak on average about 3–7 less true speakers and 5–10 less false speakers. In multi-site condition, where better improvements were observed, exclusion typically results in 0–1 less detected true speakers and 20–24 less false speakers. According to the reference annotations there are 44 speakers in single-site scenario and 38 speakers in multi-site scenario (see Chapter 5).

Overlap exclusion may also be perceived as a frame purification mechanism. It is a bit surprising that sometimes the improvements with overlap segments detected by a real system are higher than by using oracle overlapping speech. For instance, on multi-site data the relative DER reduction with reference overlaps is only 2.5%. On single-site data it is 6.2%.

7.2.2 *Application of Overlap Labeling*

Overlap labeling on AMI single-site evaluation data exhibits comparable improvements over the baseline DER of 38.3% for all setups. These reach from 4.3% relative for *Spct+Spat dT* up to 5.5% relative for *Spct+Spat CD* and *Spct+Prosod 20*. Results are given in Table 9.

Table 9: Speaker diarization with exclusion and labeling of simultaneous speech detected by different systems on AMI evaluation data, baseline DER and relative improvements over the baseline (in %).

		SINGLE-SITE			MULTI-SITE		
	OVERLAP DET.	Excl.	Labl.	Ex. + Lb.	Excl.	Labl.	Ex. + Lb.
Baseline			38.3			37.3	
+ Ovlp. Excl., Labl., or Both	Spct	+3.9	+4.7	+6.9	+4.9	+1.3	+6.7
	Spct+Spat (PCA) C	-0.3	+4.8	+4.0	+10.4	+2.0	+13.1
	Spct+Spat (PCA) D	+5.1	+4.5	+9.6	+11.6	+0.3	+13.5
	Spct+Spat (PCA) dT	+1.7	+4.3	+4.4	+10.0	+1.8	+12.5
	Spct+Spat (PCA) CD	+0.3	+5.5	+3.4	+5.4	-0.1	+5.6
	Spct+Spat (PCA) CdT	+3.0	+4.4	+5.7	+13.9	+2.0	+17.0
	Spct+Spat (PCA) DdT	+1.8	+4.9	+6.7	+9.7	+0.3	+10.2
	Spct+Spat (PCA) CDdT	+5.2	+5.4	+11.2	+6.9	+0.2	+8.0
	Spct+Spat MLP	+3.1	+5.1	+5.7	+12.1	+1.4	+13.9
	Spct+Spat LDA	+3.0	+4.6	+5.8			
	Spct+Prosod 20	+4.3	+5.5	+7.2	+9.2	+0.7	+11.1

Although the differences between *Spct* overlap detection and one of the better performing combined systems, spatial *Spct+Spat CDdT* or prosodic *Spct+Prosod_20*, for instance, is not dramatic, they still indicate a slight increase of improvement by the addition of spatial or prosodic features. These setups, which are more successful than the spectral system, have better overlap detection properties at the selected penalization point (OIP −100), detection error and recall in particular (see Figures 28 (a) and 32 (a)).

Improvements by labeling are much more interrelated with OD performance compared to exclusion

The multi-site improvements in comparison to single-site scenario are lower, they range from a small degradation of −0.1% in case of *Spct+Spat CD* to an improvement of 2.0% over the baseline DER of 37.3%. The latter result is achieved by both *Spct+Spat C* and *Spct+Spat CdT* setups. Worse labeling results were expected, since the detection error and precision properties of multi-site overlap detection do not attain those from single-site detection (see Figures 28, 29, and 32).

The factor that has the most influence on the results here is probably overlap detection precision. This becomes well visible when recalling Figure 29 (b) and taking note of the split in precision between setups with and without spatial dispersion. While the labeling of overlaps that correspond to setups not including this parameter performs better than *Spct*, setups with spatial dispersion achieve only insignificant improvements.

7.2.3 *Joint Application of Exclusion and Labeling*

In the previous sections we observed that the individual application of the two overlap handling techniques produced some improvement over the baseline diarization. Our expectation when applying them both together in one experiment is to observe some kind of synergic effect, and consequently obtain even higher improvements. As it turns out, it is true in the majority of cases, but not always.

Table 9 shows that all setups are yielding improvements over the baseline diarization. A visual representation of some of them is illustrated in Figure 35. The best relative improvement of 11.2% in single-site condition is achieved with *Spct+Spat CDdT* detection system. This corresponds with the results of overlap detection in Figure 28 (a) where spatial PCA was the overall best performing setup. Good result with exclusion alone most probably stimulated also the relatively good improvement with *Spct+Spat D* overlap segments (9.6%). Somehow surprising is that spatial MLP could not turn the improvements by exclusion and labeling separately also into a higher combined performance.

When overlaps are detected by the combined prosodic system, their discarding from training process together with the assignment of second speaker labels reduce the single-site baseline DER by 7.2%

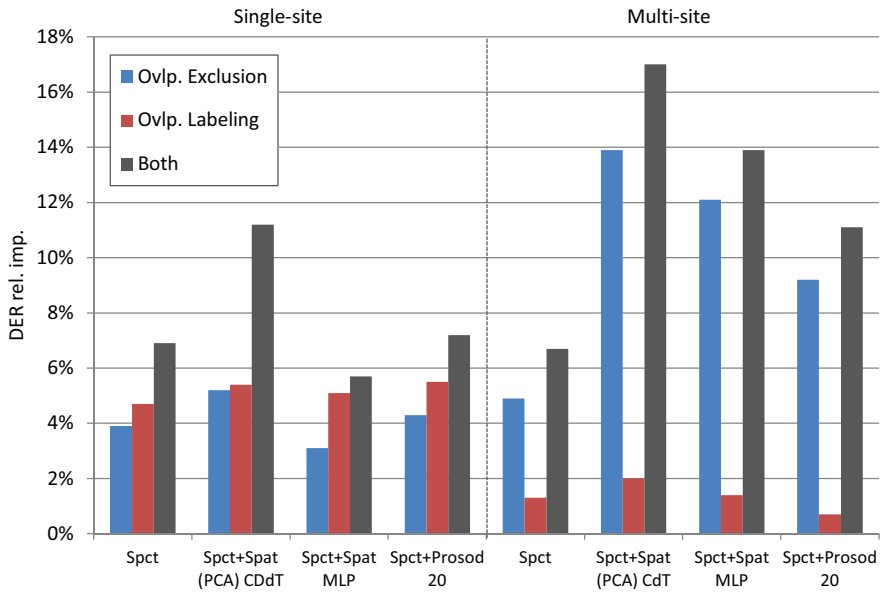


Figure 35: Improvement of baseline speaker diarization by exclusion and labeling of simultaneous speech detected by different systems on AMI single- and multi-site evaluation data.

relative. However, this result is not much higher compared to the *Spct* case. On the other hand, on multi-site data the relative improvement by using *Spct+Prosod 20* is 11.1%, which is much higher than the spectral overlap detection setup.

Very good results with the application of both overlap handling techniques are mostly due to exclusion

Similarly to the application of exclusion alone discussed in Section 7.2.1, the relative DER improvements observed on multi-site data are, in general, higher in comparison to single-site data. Driven by the very good improvement with exclusion, the use of overlap hypotheses from *Spct+Spat CdT* setup obtains the overall best result of 17.0% DER reduction. The same observation applies for a couple of other spatial PCA setups. The spatial MLP setup confirms with 13.9% relative improvement its good performance on multi-site data from before. A set of similar diarization experiments was presented in [129], but with the difference that the overlapping speech model, used for detecting simultaneous speech segments, had 32 Gaussians components in its GMMs (here 64).

Unfortunately, there is no standard procedure how to estimate confidence intervals in speaker diarization experiments that would be defined, for instance, by NIST for the Rich Transcription competition. In this work we followed the same procedure that was used in [148]. For a $100(1-\alpha)\%$ confidence interval the margin of error is $m_e = z_{\alpha/2}\sqrt{\mathrm{DER}(100-\mathrm{DER})/N_f}$, where $z_{\alpha/2}$ is the upper $\alpha/2$ critical value of the normal distribution, and N_f is in this case the number of frames. This formula assumes that a frame-level decision of a speaker

Confidence interval in diarization experiments estimated to ±0.1%

diarization system is a Bernoulli trial with DER percentage of success. The margin of error corresponding to a 95% confidence in the obtained DER results varies in the range 0.063–0.078%, so after rounding to one decimal, the confidence interval around DER values is ±0.1%. The relative improvements were computed according to the center values.

7.3 OVERLAP HANDLING WITHIN EXTENDED SPEAKER DIARIZATION

An interesting question is, if it will be still possible to achieve improvement by overlap handling when the baseline diarization system is improved with state-of-the-art techniques like beamforming and the use of an additional TDOA feature stream. In such case, less room would be left for improvement by other techniques.

7.3.1 *Overlap Labeling and Superior Clustering*

Table 10 gives the DERs of the new, improved, baseline system for single- and multi-site data. In order to demonstrate the effects of overlap labeling and the use of a real SAD system, the DER is also decomposed into three components: missed speaker time error (MS), false alarm error (FA), and speaker error (SPKE).

The application of beamforming and TDOA features did improve the baseline system despite little optimization. The change in performance was from 38.3% to 35.7% DER and from 37.3% to 32.5% DER for single- and multi-site scenario, respectively.

We repeated overlap labeling experiments for several overlap detection setups, namely *Spct*, *Spct+Spat MLP*, *Spct+Spat LDA* (single-site), *Spct+Prosod 20*, and one most promising spatial PCA setup for single- (*Spct+Spat CDdT*) and multi-site data (*Spct+Spat CdT*) each.

Assignment of second speaker labels benefits from improved clustering

It can be seen that the labeling algorithm takes advantage of the improved clustering, since in all cases the relative DER improvements in Table 10 increased compared to previous results in Table 9. This observation is consistent, since the better clustering process also implies a higher effectiveness of the attribution of second speaker labels.

In single-site scenario the best improvement of 6.5% relative, from 35.7% to 33.4% DER, is achieved by labeling of overlaps from the *Spct+Prosod_20* detection system. As far as the spatial setups are concerned, *Spct+Spat CDdT* and *Spct+Spat MLP* achieve comparable results of 6.2% and 5.9% relative improvement, respectively.

The best achieved result in multi-site scenario is the DER reduction from 32.5% to 31.4% when second-speaker labels are assigned to overlap segments from the spatial setup *Spct+Spat CdT* (3.4%). The results with either the combined spatial MLP or the pure spectral overlap detection are not much different though. When using a real

SAD system the improvements are lower, but they basically follow the same pattern as with oracle speech/non-speech segmentation.

The addition of prosodic features for overlap detection and their labeling improved the diarization result a little. However, the improvements were lower compared to the *Spct* system. Even though the overlap hypothesis of the combined prosodic system corresponding to OIP −100 exhibits high recall (40%), the precision of 65% (see Figure 32 (b)) is obviously not sufficient and too many false speech segments are introduced to the diarization hypothesis. A possible explanation may be the fact that the selection and integration of prosodic features was basically tuned for single-site condition, which probably was not optimal for multi-site scenario.

It is worth reminding that the DER scores are computed with no forgiveness collar. Scoring with a collar of 0.25 s, which is common for instance in NIST RT evaluations, reduces the DER values. These values are given in Table 10 in parentheses next to the no-collar DERs. The DER improvements remain consistent, but the application of a forgiveness collar in some cases mitigates the gain in segmentation precision introduced by using some (spatial) overlap setups, because the segmentation changes are rather short.

Applying a scoring forgiveness collar lessens the DER differences, but they basically remain consistent

Comparison of Labeling Techniques

Our labeling technique for assigning second speaker labels to overlapping speech segments is integrated into the Viterbi decoding in the diarization system. Table 11 shows its comparison to two simple labeling schemes in terms of relative DER improvement over the diarization with beamforming and TDOAs. The improvements are also illustrated in Figure 36. The first of these techniques a posteriori attributes the overlapping speaker label according to the nearest neighboring speaker, similarly to [117]. The other competing technique assigns the overlapping label to the most talkative speaker [119]. If the most talkative speaker has already been picked by the diarization system, the second most talkative speaker is selected in such case. In general, the differences between DERs of the three labeling techniques are small, but it can be seen that the results of the technique proposed in this thesis are competitive, in single-site scenario in particular.

Proposed labeling technique integrated in Viterbi decoding delivers competitive results compared to alternative strategies

In [113], another assignment strategy relying on posterior speaker probabilities was proposed and relative improvements of 5.1% and 2.3% for single- and multi-site AMI sets, respectivelly, were presented. However, these testing sets are not exactly the same as in our experiments because multi-channel data is not available for some recordings (refer to Chapter 5 for more details).

Table 10: Improved speaker diarization with labeling of simultaneous speech detected by different systems on AMI evaluation data, missed speaker-time error (MS), false alarm error (FA), speaker error (SPKE), DER (score when a collar of 0.25 s is applied in parentheses), and relative improvements over the new baseline (in %).

		SINGLE-SITE					MULTI-SITE				
	OVERLAP DET.	MS	FA	SPKE	DER	IMP.	MS	FA	SPKE	DER	IMP.
	Baseline + Beam. + TDOAs	18.7	0.0	17.0	35.7 (27.3)		17.1	0.0	15.4	32.5 (23.6)	
+ Ov. Labl.	Spct	15.3	0.6	17.9	33.8 (25.3)	+5.3	12.8	2.0	16.9	31.6 (23.2)	+2.7
	Spct+Spat (PCA) CDdT	14.8	0.6	18.1	33.5 (25.2)	+6.2					
	Spct+Spat (PCA) CdT						12.9	1.6	16.9	31.4 (22.8)	+3.4
	Spct+Spat MLP	14.7	0.8	18.1	33.6 (25.3)	+5.9	12.8	1.8	16.9	31.5 (23.0)	+3.0
	Spct+Spat LDA	15.6	0.5	17.9	34.0 (25.5)	+5.0					
	Spct+Prosod 20	14.2	1.1	18.1	33.4 (25.2)	+6.5	11.4	3.1	17.5	32.0 (24.1)	+1.6
	Including real SAD										
	Baseline + Beam. + TDOAs	21.8	6.7	16.7	45.1 (33.5)		19.6	6.0	16.5	42.2 (30.5)	
+ Ov. Lb.	Spct	18.4	7.3	17.8	43.5 (31.7)	+3.7	15.3	8.0	17.9	41.3 (30.1)	+2.1
	Spct+Spat (PCA) CDdT	17.9	7.3	17.9	43.1 (31.5)	+4.5					
	Spct+Spat (PCA) CdT						15.5	7.7	17.8	41.0 (29.6)	+2.7

Figure 36: Improvements of speaker diarization by different labeling strategies on AMI single- and multi-site data.

Table 11: Comparison of different labeling strategies on single- and multi-site data. Relative DER improvements over the new baseline (in %) by the attribution of second speaker labels according to Viterbi-decoding (Vit.), nearest-neighboring-speaker (NN), and most-talkative-speaker (MT) scheme.

	SINGLE-SITE			MULTI-SITE		
OVERLAP DET.	Labl. Vit.	NN	MT	Vit.	NN	MT
Spct	+5.3	+5.1	+4.8	+2.7	+3.2	+2.3
Spct+Spat (PCA) CDdT	+6.2	+5.9	+5.6			
Spct+Spat (PCA) CdT				+3.4	+3.9	+3.0

7.3.2 *Addition and Effect of Overlap Exclusion*

The improvements obtained by using overlap exclusion and beamforming with TDOA features are similar as if used standalone. Table 12 shows the performance of the extended diarization system if overlapping speech segments are labeled and also excluded. What is obvious from these results is that in both scenarios the relative DER improvements are either not better than with the application of overlap labeling only, or they are actually worse (compare to Table 10). Moreover, in some cases the results are even worse than the DER of the system without any overlap handling.

Addition of overlap exclusion could not improve new diarization performance

This performance difference is especially contrasting with the *Spct+Prosod 20* setup on multi-site data. After having a very good relative improvement of 11.1% over baseline diarization in Table 9, and yield-

Table 12: Improved speaker diarization with exclusion and labeling of simultaneous speech detected by different systems on AMI evaluation data, DER (score with 0.25 s forgiveness collar given in parentheses) and relative improvements over the new baseline (in %).

		SINGLE-SITE		MULTI-SITE	
	OVERLAP DET.	DER	IMP.	DER	IMP.
Baseline + Beam. + TDOAs		35.7 (27.3)		32.5 (23.6)	
+Ov. Ex. and Lb.	Spct	34.0 (25.4)	+4.9	32.8 (24.6)	-0.9
	Spct+Spat (PCA) CDdT	35.6 (27.4)	+0.3		
	Spct+Spat (PCA) CdT			34.1 (25.9)	-5.0
	Spct+Spat MLP	33.7 (25.1)	+5.6	34.3 (26.4)	-5.4
	Spct+Spat LDA	34.9 (26.6)	+2.3		
	Spct+Prosod 20	33.9 (25.8)	+5.0	35.5 (28.3)	-9.1

ing some small improvement by labeling in the extended diarization system (Table 10), here it exhibits a significant degradation of performance. In conclusion, the exclusion technique in this case was not very successful in further improvement of the diarization system.

It seems that overlap exclusion and beamforming with TDOAs in speaker diarization are not complementary techniques, or that there exists some sort of improvement redundancy between them. In a speech overlap situation, where the speech comes from several locations, the TDOA behavior might be either erratic or it can probably focus on the acoustic source with higher energy. The latter situation clearly benefits the diarization task. Both the cleaned MFCC parameters derived from the beamformed signal and the process of filtering of the interfering speakers yield to improved speaker segmentation results. In this sense, this approach is close to the strategy of overlap exclusion. Another reason might also be the not very stable behavior of exclusion which was visible in Figure 34.

These observations are actually in concordance with the observations made by Otterson and Ostendorf in [5] and also with the results published by Huijbregts et al. in [117]. The purification of clusters by overlap exclusion was not working well with the use of spatial information stream in the diarization system.

NIST RT Experiments

Finally, in a series of preliminary experiments, some effort was spent to obtain results on the NIST RT '09 conference meetings. We selected the overlap hypotheses presented in Figure 30 for the same OIPs as were already selected for the AMI corpus, i. e., no penalization for

overlap exclusion and OIP −100 for labeling. The baseline diarization performance with the improved system utilizing beamforming and TDOAs is 32.5%. The application of the overlap handling techniques reduced the diarization error in this case to 30.6% for the spectral overlap detection setup, and to 30.2% for the combined spectral and spatial setup. Again, these DER results are computed without any scoring collar. When the standard collar of 0.25 s is used, the corresponding error reduction is from 19.6% to 17.5% DER for the latter of the two overlap setups. The DER 95% confidence interval radius for RT '09 data is also ±0.1%.

Overlap handling experiments on NIST RT data show good potential

In preliminary experiments using a real SAD instead of reference speech segments, the baseline DER of 43.3% (23.0% with collar) could be decreased to 42.9% (22.8% with collar) with overlaps from the combined system. It seems that the system has a good potential for improvements, but the preliminary obtained results are not statistically different.

7.3.3 *Performance Analysis on Individual Meetings*

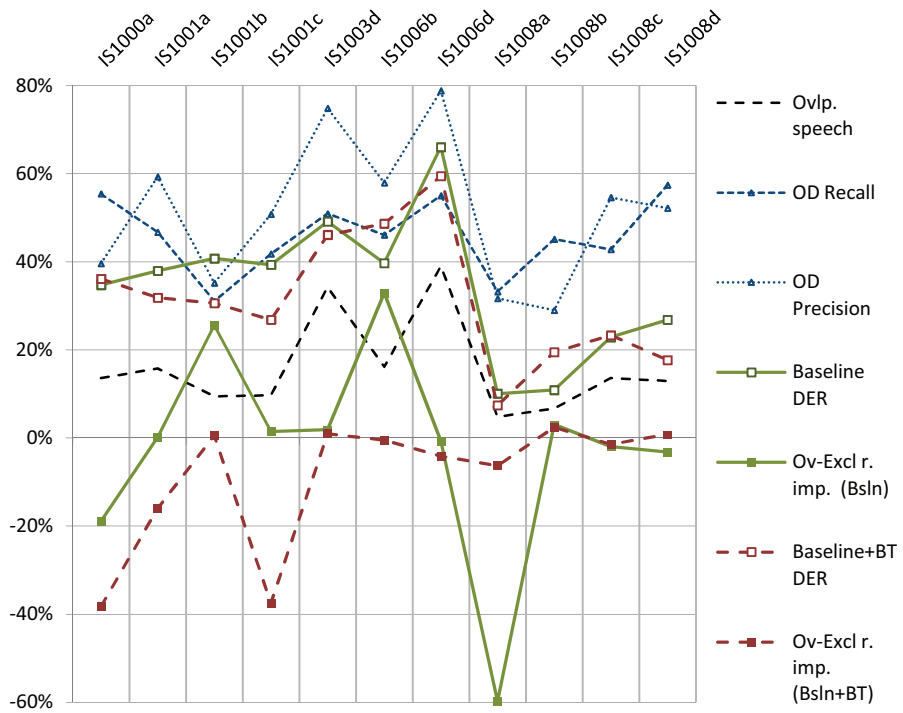
Our experience that overlap exclusion failed to further improve the diarization which uses also a spatial feature stream, on evaluation data as a whole, motivated us to do a more closer analysis. Figures 37 (a) and (b) detail the performance of the diarization system on individual meeting recordings from AMI single-site data, and also demonstrate the corresponding relative improvements by handling simultaneous speech.

There are several observation that can be made. First of all, the DER values among different meetings are highly variable, they range between approximately 10% and 66%. A simple explanation for such high variability is not at hand, since all meetings comprise the same amount of speakers, i. e., four (refer to Chapter 5), the acoustic conditions are also the same, and not to forget that the speech/non-speech segments are considered perfect. However, note that the recordings with a high amount of overlapping speech also exhibit a worse diarization performance. The amount of speaker overlap is derived from reference annotations.

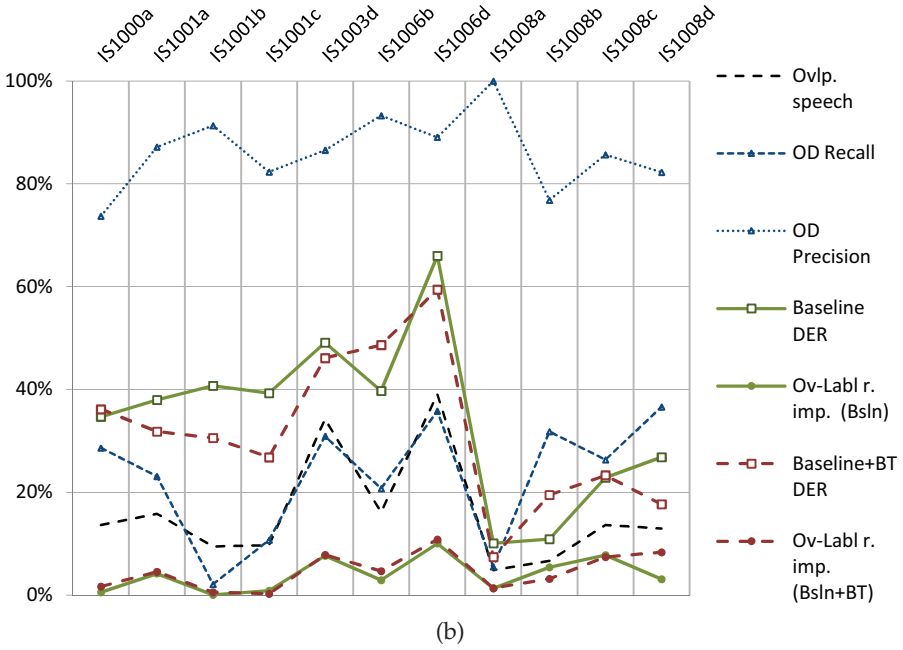
Considerable variability in diarization error among individual meeting recordings

The introduction of beamforming and a second, TDOA, feature stream generally improves the diarization results, but not for all recordings. For instance, in the IS1006b and IS1008b meetings the extended diarization system (in Figure 37 denoted as *Baseline+BT*) yields higher DERs than the original baseline system.

As regards the overlap detection in Figure 37 (a), overlapping speech segments were detected by *Spct+Spat CDdT* system with no penalization. Interesting fact is that OD precision seems to be somehow



(a)



(b)

Figure 37: Performance analysis on individual meetings from the AMI single-site evaluation data. DERs of baseline diarization and diarization using beamforming and TDOAs together with corresponding relative improvements by (a) exclusion and (b) labeling of detected overlapping speech with *Spct+Spat CDdT* system.

correlated with the amount of overlapping speech in meetings in this case.

High improvements by exclusion apply only to some recordings

Although overlap exclusion improves the baseline DER on evaluation data as a whole, the decomposition of relative improvements according to particular meetings reveals substantial differences between them. Some recordings are improved a lot, with some the effect of exclusion is only moderate, and on some recordings the DER is actually increased. The extensive relative degradation in case of IS1008a meeting is due to a very low baseline DER value. No particular interrelation between improvement by exclusion on one side and OD recall, OD precision, the amount of overlap, or baseline DER on the other side is visible. For instance, consider the recordings IS1001b, IS1008a, and IS1008b. We can see that all of them have similar amounts of overlap, and also comparable OD precisions were achieved. Two of them, IS1001b and IS1008a, obtained similar recalls and other two, IS1008a and IS1008b, have almost the same baseline DER. Nevertheless, all three recordings exhibit very different relative improvements by overlap exclusion over the diarization baseline. It remains unclear what is the factor that influences the potential to improve a particular meeting DER. Perhaps, these results can be put in context with the not easily predictable behavior of exclusion in Figure 34.

Unclear relationship between exclusion performance and other factors

Exclusion, in this case, did not manage to reduce the DERs of the extended diarization system for almost any meeting. However, it would be wrong to deduce that improvement can never be achieved. Although such results are not directly presented here, but when we take into account also recordings from multi-site data or when we use overlap hypotheses from another overlap detection system, it was observed that on some few recordings the extended baseline could be improved. Sometimes even in cases in which the original baseline DER was not reduced. In Figure 37 (a) this can be partly visible on IS1008a meeting where the degradation, despite smaller DER of the extended system compared to the baseline, is much lower. In conclusion, it only confirms the uncertainty concerning the exclusion technique stated before.

For completeness, Figure 37 (b) provides a similar performance analysis per meeting for overlap labeling. Overlap detection precisions are higher since overlap hypotheses at OIP -100 are used for labeling. The recall is lower, though the line has almost the same shape as in Figure 37 (a).

Overlap labeling normally exhibits much easier predictable behavior

Considering the relative improvements and the amount of overlapping speech in recordings, a certain correlation between them can be recognized. Furthermore, a relationship also exists between DER improvements and recall in combination with sufficiently high precision. Note that sometimes a superb (IS1001b) or even a 100% (IS1008a) precision cannot assure high DER reduction when there is a small room

for improvement caused by either low amount of overlap or eventually low recall.

Overlap labeling within the baseline and within the diarization system relying also on beamforming and TDOAs shows only small differences. Its application within the extended system yields slightly better results for several meeting recordings, but not always (IS1008b, for instance).

8 CONCLUSIONS

This chapter gives a brief summary of this thesis and pinpoints the most important outcomes of the presented experiments. Accomplished work and proposed techniques are reviewed with regard to the thesis objectives formulated in Chapter 1. Finally, some suggestions for the future work on this topic are outlined.

8.1 SUMMARY

This thesis deals with the issues of overlapping speech in the context of speaker diarization on distant microphone channels. In order to locate the regions where multiple speakers are speaking simultaneously an overlap detection system was built. We have found that spatial information can be utilized to perform this detection and proposed three novel cross-correlation-based features. The problem of high and variable dimensionality of spatial feature space was addressed with the application of a per-site-specific PCA, LDA, or an MLP neural network. Furthermore, we have also introduced features based on prosody and their long-term statistics for the detection of overlapping speech. The final subset out of all candidate prosodic features was selected according to mRMR criterion and a successive hill-climbing wrapper selection method.

Honestly speaking, the performance of the simultaneous speech detection is in general not very good. The task to distinguish single-speaker speech from speech including multiple speakers proves to be extremely challenging for an automated system. As a matter of fact, in some cases it is difficult even for humans to decide what can and what cannot be considered overlapping speech, for instance, loud breathing or nonverbal sounds. Such ambiguities can also have an impact on reference transcriptions which are used either for training or scoring the output of the detection system.

Proposed spatial and prosodic parameters contributed to the detection of overlapping speech

Nevertheless, in several experiments on AMI single- and multi-site meeting data, we showed that overlap detection involving the use of spatial or prosodic parameters outperformed the baseline system. In this work the baseline system relies on spectral-based features only, such as MFCCs, LPCRE, SF, and first-order deltas. From tests with various combinations of spatial parameters, we can conclude that the PCA-fused dispersion ratio is well suited for the single-site condition. The system using MLP score has a good detection performance in this scenario, but for high OIPs its precision drops bellow the one of the

baseline system. On the contrary, the system relying on LDA fusion exhibits the highest precision in all experiments, but at the cost of detection recall.

For the multi-site scenario, however, the mentioned PCA-fused dispersion ratio seems to lack robustness. The possible reason for the worse performance of feature setups involving this parameter is its dependency on the spatial distribution of microphones, which might be an issue in case of using multiple recording rooms. Moreover, the limited ability to compensate for the variability of this scenario can most likely be attributed to the simplicity of the PCA technique. In that case the better performing combinations included spatial coherence and delta TDOA, but the distinction in performance between setups including and not including dispersion ratio becomes evident only at higher penalization values. The MLP technique combines all three spatial parameters in this scenario more effectively and outperforms, or at least equals, the baseline system at all instances. In general, the less precise multi-site models need a higher amount of overlap penalization to arrive to the lowest detection errors. The complexity of the NIST RT data in the sense of the number of involved meeting rooms was probably the reason why the detection on this alternative corpus was worse than on AMI data, even with higher penalization.

The addition of prosodic features decreased the overlap detection error in both scenarios either due to higher precision for low penalties or due to improved recall in high penalty region. Despite our initial concerns that the HMM model of overlapping speech will be prone to detect unrelated laughter, we discovered that the presence of such segments in the training data is not affecting the actual overlap detection much.

Speaker diarization could be improved by handling detected overlapping speech

By handling of the detected simultaneous speech segments, we managed to improve the baseline speaker diarization system. With the objective to build more precise speaker models, the speech frames including overlapping speech were excluded from the training process. In addition, we reduced diarization's missed speech by assigning second speaker labels for speaker overlap segments.

The most successful overlap detection setups in terms of successive diarization improvement was on single-site data the combination of spectral and all three PCA-transformed spatial parameters. A good result was also obtained with the combination of spectral and prosodic features. In the multi-site scenario the relative improvements were higher, particularly on account of overlap exclusion. Here, taking advantage of the mentioned overlap detection performance, the best observed result was with the combination of spatial coherence and delta TDOA. Another successful system was the one using MLP for the spatial parameter fusion. Considering overlap labeling only, the comparison of DER reductions between the two scenarios shows much

better performance on single-site data than on multi-site data. Such results are not surprising since the detection error and precision of the multi-site overlap detection do not attain the quality of those in the single-site detection.

The extension of the baseline diarization system with beamforming and TDOA features improved not only the clustering process, but also resulted in an increased performance of overlap labeling. The reason is that the more effectively working clustering causes the mechanism for picking the second speaker to be mistaken fewer times, in general. A further study has shown that our labeling technique, integrated in the Viterbi segmentation algorithm, delivers competitive results to alternative simple strategies for the assignment of overlapping speaker labels, especially in the single-site scenario. Overlap exclusion in this case did not result in further improvement of the new system on evaluation sets as a whole.

Overlap labeling works on top of the baseline system, exclusion affects the diarization algorithm basics

Interestingly enough, these two techniques demonstrated quite distinct behavior regarding DER improvement. The performance of overlap labeling is closely related to overlap detection error, and, in fact, recall, with a requirement for a high detection precision. Overlap labeling can be considered as an addition on top of the baseline system. Due to the necessity for high precision, and the consequently rather low recall of the overlap detection system, the potential improvement is developing in a limited range.

Overlap exclusion, on the other hand, affects the core elements of the diarization algorithm. This may be one of the reasons why it exhibits an unpredictable nature to some extent. In some instances it achieves very high improvements of the baseline DER, but on others it can actually cause performance degradation. Such performance variability was observed both between different setups of the underlying overlap detection (different settings of the OIP, for instance), as well as across different meeting recordings. High variability among the DER scores of individual meetings is, however, also typical of the baseline diarization system. Poor diarization performance on some meeting recordings is often correlated with a high amount of present overlapping speech. Based on our experiences, we suggested that it is reasonable to use independent overlap detection hypotheses for exclusion and for labeling.

The application of a scoring forgiveness collar in some cases mitigates the gains in segmentation precision introduced by overlap handling, nevertheless, the DER improvements remain consistent. A similar observation can be made for the use of a real SAD system. The relative improvements are lower, but basically follow the same pattern as with a perfect speech/non-speech segmentation.

Preliminary diarization experiments on NIST RT data with overlap handling demonstrated a reduction of the baseline DER for both

spectral and combined spatial overlap detection system. Further experimentation with a real SAD revealed that the system has a good potential for improvements, but the preliminary obtained results were not statistically different.

8.2 FUTURE PROSPECTS

One of the drawbacks of current overlap detection systems are the issues with robustness. Dealing with this problem should be among the priorities of future research. We saw in the past in other fields that robustness of systems can be improved by employing another modality, in the case of meetings this could be the video information. Since meeting participants are usually looking at the current speaker, reliable head orientation information could also be beneficial assuming that an interrupting speaker will draw attention by some of the participants.

More inspiration could also be taken from source separation approaches, such as CASA. For instance, when performing a subband pitch estimation, the detection of different pitches in two subbands may be an indication of speaker overlap.

The acoustic signal from one speaker arrives to a couple of distant microphones with different delays as the speech by a concurrent speaker. Given that there is a critical distance between the microphones, further information sources on the presence of overlapping speech could be found by considering the time-frequency-space diversity of multi-channel signals.

In the context of overlap handling in diarization, the crucial task for the future is to increase the stability of overlap exclusion operation. This might, however, be related to the general problem in speaker diarization, where there are recordings that exhibit unusually high DER (called "nuts") and others that are over-sensitive to tuning (referred to as "flakes").

A

SPEAKER DIARIZATION OF BROADCAST NEWS IN ALBAYZIN 2010 EVALUATION CAMPAIGN

Objective evaluations became a valuable part of research and development in the field of spoken language processing. The comparison of performance of different approaches (systems) to a specific task helps setting new trends and stimulates the progress in a particular line of research. The Albayzin 2010 is the third in the series of evaluation campaigns (2006, 2008) organized by RTTH[1] and held under the FALA 2010 workshop [149]. Largely inspired by the NIST Rich Transcription evaluations [118], the Albayzin 2010 campaign focuses among others on the task of speaker diarization of broadcast news.

In this appendix we present as the co-organizers of Albayzin 2010 responsible for speaker diarization section an overview of the evaluation and report the results achieved by five submitted speaker diarization systems. The evaluation was performed on Catalan broadcast news data. Although the presented systems have several features in common (e. g., MFCCs, agglomerative clustering), there are also many differences among them, e. g., online optimized processing, speaker factor analysis, dot-scoring similarity, or acoustic fingerprinting. Based on the observed results, we try to derive the most successful system features and outline promising investigation directions. The diarization performance is analyzed in the context of the diarization error rate, the number of detected speakers and also the acoustic background conditions.

Broadcast news is a challenging domain, because such shows contain an unpredictable number of different speakers speaking for a very variable amount of time and speakers sometimes talk simultaneously. However, overlapping speech issue was not very significant in this case. Broadcast news data often contain a large amount of music and commercial breaks.

A.1 SPEAKER DIARIZATION TASK AND SCORING

The organized evaluation campaign aims at evaluating the performance of automatic algorithms for speaker diarization task. The participants could submit more than one system output, but only the primary hypothesis was considered here.

1 RTTH is the Spanish acronym for "Red Temática en Tecnologías del Habla" (the Spanish Speech Technologies Thematic Network)

Table 13: Distribution of speakers

Gender	# Speakers	Duration [h]	# Segments
male	1239	44:23:41	12869
female	507	25:43:54	7559
unknown	270	07:50:38	2579
overlapped	68	00:12:38	241

The minimum silence duration separating two utterances was set to 0.5 s, since pauses smaller than this value were not considered to be segmentation breaks in a speaker's speech (it is also complementary to the scoring collar discussed later). The Diarization Error Rate (DER) defined by NIST [118] is the primary metric.

A scoring "forgiveness collar" of 0.25 s around each reference segment boundary is used. This accounts for both the inconsistent annotation of segment times by humans and the uncertainty when does speech begin for word-initial stop consonants.

A.2 EVALUATION DATABASE

Broadcast news audio data recorded from Catalan 3/24 TV channel

The database contains broadcast news channel recordings, i. e., announcements, reports, interviews, discussions and short statements recorded from Catalan 3/24 TV channel throughout the program. Its original video recordings were supplied by a stationary digital video broadcasting (DVB-T) receiver. Their original audio tracks were extracted being available at 32 kHz sample rate, 16 bit resolution, but were downsampled to 16 kHz sample rate.

The annotated recordings comprise a total duration of 88 hours, but for the Albayzin 2010 speaker diarization evaluation a subset of 8 recordings totaling approximately 30 hours was selected. Although TV3 is primarily a Catalan television channel, the recorded broadcasts contain a proportion of roughly 8.5% of Spanish speech segments.

Catalan language, mainly spoken in Catalonia, exhibits substantial dialectical differences and is divided into an eastern and western group. The eastern dialect includes northern Catalan (French Catalonia), central Catalan (the eastern part of Catalonia) and Balearic. The western dialect includes north-western Catalan and Valencian (south-west Catalonia) [150]. Presumably, the majority of recorded Catalan speakers features the central Catalan dialect.

A first annotation pass segmented the recordings with respect to background sounds, channel conditions, and speakers as well as speaking modes. Table 13 shows the speaker distribution. Since segments of overlapping speakers did not receive a gender tag, they form also

Table 14: Distribution of recording channels and background conditions (number of segments in parenthesis)

Channel	Background [h]			
	None	Speech	Music	Noise
None	04:27:10 (2451)	00:18:54 (131)	04:36:06 (1945)	01:15:30 (1113)
Studio	15:04:24 (4752)	01:36:16 (594)	08:40:47 (1407)	00:57:12 (2067)
Telephone	00:00:40 (11)	00:00:10 (2)		00:06:47 (10)
Outside	14:49:44 (6558)	03:55:29 (1319)	01:52:52 (557)	18:55:19 (4342)

a subset of the "unknown" gender account. The gender conditioned distribution indicates a clear misbalance in favor of male speech data. The number of speakers per recording ranges from 30 to 250 with some speakers appearing in several recordings (newscaster, journalists). However, the majority of speakers are related only to a particular news and account to only a short duration.

The total durations of audio segments of specific conditions are given in Table 14. Besides, there are a few conditions featuring an overlap of all noted background sounds, but only with minor duration and are therefore omitted. Few segments are indicated to originate from telephone speech. The recorded speech within these segments can be considered band-limited to frequencies from 300 Hz to 3.4 kHz.

A second annotation pass provided literal transcriptions and acoustic events of segments that feature planned and spontaneous speech, but no long term background noises. The non-speech acoustic events were furthermore tagged with time stamps indicating their beginning and end.

Because of the fact that silences were not manually annotated, the transcriptions were extended by passing the signal through the hierarchical audio segmentation described in [151]. This involved a simple low-energy silence detector to estimate regions with non-speech (silence). Furthermore, to avoid too short segments, a smoothing constraining the minimal non-speech duration to 0.5 s was applied.

Table 15: Participating teams in the Albayzin 2010 speaker diarization section

Team ID	Research institution
AhoLab	University of the Basque Country (EHU)
GSI	University of Coimbra (UC)
GTM	University of Vigo (UVigo)
GTC-VIVOLAB	University of Zaragoza (UZ)
GTTS	University of the Basque Country (EHU)
ATVS-UAM	Autonomous University of Madrid (UAM)

A.3 PARTICIPANTS

Six teams from five research labs submitted their systems to the Albayzin 2010 speaker diarization evaluation. The list of participants is given in Table 15.

Six submitted systems (but only five considered) from five reserach labs

After submitting evaluation results one of the teams discovered that in half of the recording sessions their system was reading corrupt audio input. Therefore, their evaluation results cannot be considered representative and only five systems are presented here. The original descriptions of the speaker diarization evaluation can be found in [152].

Several teams participated also in another category of the Albayzin 2010 evaluation, in the audio segmentation section, where five acoustic classes were defined to segment the audio data [153]. The classes were as follows: music, clean speech, speech with music, speech with noise and other (e. g., noise, silence). Since audio segmentation normally constitutes a part of speaker diarization systems, we are referring in latter system descriptions to these five acoustic classes.

3.1 System 1

The algorithmic concept of System 1 facilitates an online execution, i. e. the complete process is performed in a single iteration. The initial SAD employs a Viterbi segmentation of the audio signal and distinguishes five acoustic classes. Each class is modeled with a GMM and signal parameterization involves static MFCCs with first and second order derivatives.

Online clustering algorithm, growing window speaker-change detection relying only on voiced speech

Subsequently, the speaker change detection employs a growing window approach and BIC to measure the dissimilarity of two adjacent windows. The BIC metric estimates if windowed audio data is better modeled with two distributions or with only a single one. In general, a change point is detected at positions where the BIC value is greater

than zero. Even though the growing window scheme has higher computational cost, the authors of System 1 report its better performance compared to fixed-size sliding window approach and implemented a number of adjustments in order to decrease the computational load (skipping improbable places, window length limit). At this stage of the process, only static MFCC features with no derivatives are used. In system development experiments it was possible to reduce the diarization error by discarding unvoiced frames. Therefore, the speaker change detection of this system relies only on voiced audio data.

During the online clustering algorithm, every time a speaker change is detected, the BIC value of the recent speech segment against all known clusters is computed. If the lowest BIC value falls below a certain threshold the segment is assigned to the given cluster. Otherwise, a new cluster is created. The theoretically suboptimal online algorithm can in practice benefit from the fact that it is prone to combine adjacent segments rather than segments far apart and consecutive segments are likely to come from the same speaker.

3.2 *System 2*

System 2 incorporates audio segmentation prior to the diarization to determine speaker turns and discard non-speech segments like silence and music. It uses a set of 16 MFCCs, 8 other features (e. g., energy, zero-crossing rate, spectral measures) and their derivatives. Segmentation is based on a hybrid ANN/HMM Viterbi decoder and discriminates between five acoustic classes.

To classify speakers, the algorithm begins with training a UBM with data of the entire audio file. Subsequently, the decoder determining the most likely mixture sequence detects (with high mixture transition penalization) the speaker turns . Homogeneous segments with speech of only one speaker tend to produce sequences with few mixtures turns.

UBM decoding where each Gaussian mixture corresponds to one cluster, clustering based on audio fingerprinting and BIC

Two passes of verification are then applied to the labeled speaker segments to test whether every pair of segments is homogeneous or not. The first pass involves an audio fingerprint system and the other is based on BIC. If two segments are classified as similar, then the corresponding speaker labels are equated.

Acoustic or audio fingerprinting refers to a condensed representation of an audio signal that can be used to identify an audio sample or quickly locate similar items in audio streams. A binary representation of spectral patterns computed by the convolution of spectrogram with a mask is used. This technique is convenient to discover repeated segments with high confidence. Labels are determined according to a majority voting scheme in order to deal with classification inconsistencies in repeated segments.

3.3 System 3

System 3 employs recent improvements in speaker segmentation of two-speaker telephone conversations using eigenvoice modeling and the traditional agglomerative hierarchical clustering.

The Joint Factor Analysis (JFA) received a lot of attention in the context of speaker verification over the last few years. The idea is to extract and model the desired sources of variability which are present among different speakers. The JFA-based speaker segmentation was originally designed for two-speaker telephone conversations, thus it works with a given number of speakers. Therefore, after separating the speech frames, every recording is split into 5 minute slices and every slice is processed individually. The segmentation system is forced to find 10 speakers in every slice.

Eigenvoice factor analysis, BIC-based agglomerative clustering

Every speaker GMM is adapted from a background model using an eigenvoice approach. Given a sequence of feature vectors consisting of 18 MFCCs, 20 speaker factors are estimated for every time point, and then transformed with the within-class covariance normalization (WCCN) in order to compensate for the intra-session variability. Afterwards, a 10-Gaussian GMM is estimated to model the stream of speaker factors, where each Gaussian will be assigned to a single speaker. Once there are 10 clusters for every 5-minute slice, clustering over the whole recording is performed to merge those clusters belonging to the same speakers. For this purpose, BIC is considered as both a clustering metric and a stopping criterion. Clusters are modeled with a single full-covariance Gaussian function using MFCCs.

3.4 System 4

System 4 consists of three decoupled elements: speech/non-speech segmentation, acoustic change detection and clustering of speech segments. All of them rely on 13 static MFCC features, while the MFCCs for clustering are additionally augmented with their first and second order derivatives.

Speech/non-speech segmentation makes use of an ergodic continuous HMM with 5 states (one per acoustic class). In order to detect speaker change points, speech segments were further segmented by means of a conventional metric-based approach evaluating the likelihood of the acoustic change in the center of a sliding window using normalized Cross-BIC (XBIC) metric. The authors of the system state that with this approach, besides many additional acoustic changes, almost all the speaker changes were detected.

XBIC-based speaker segmentation, clustering employing GMM sufficient statistics and dot-scoring similarity

The clustering employs linear dot-scoring, a fast and simple technique for scoring test segments against target models which employs the first-order Taylor-series approximation to the GMM log-likelihood.

For each speech segment a GMM was MAP-adapted from a universal background model, and zero- and first-order sufficient statistics are computed. The similarity between different segments is then estimated with TZ-normalized dot scores. Finally, an agglomerative clustering algorithm is used until no pair of clusters exceeds a similarity threshold.

3.5 System 5

The front-end parameterization of the speaker diarization System 5 involves the extraction of 19 static MFCCs with their deltas, followed by Cepstral Mean Subtraction (CMS), RASTA filtering and feature warping. All speech data detected by a preceding audio segmentation step is used to train a UBM. Given this UBM, sufficient statistics are extracted for every segment. The next steps involve a factor analysis to model the total variability subspace resulting in so-called iVectors and a LDA transformation of the computed iVectors.

Two-step clustering relying on iVectors and Viterbi decoding

The MFCC feature stream is divided into 90-second audio slices. LDA-projected iVectors in each slice are clustered based on their cosine distance. Cluster centroids represent candidate speakers. Candidate speaker models are accumulated over all the slices in the test session together with the frequency of appearance of their clusters.

Speakers presumably appear in several slices, thus a secondary clustering merges the initial centroids, obtaining an enhanced set of candidate speakers. A prior probability is assigned to each of the candidate speakers according to its presence in the entire session. Likelihoods for each candidate speakers are estimated in a second pass over the iVector stream using the cosine distance and the prior probability of each candidate speaker. Finally, the output diarization labels are obtained by a Viterbi decoding of so-calculated speaker scores.

A.4 EVALUATION RESULTS

The DER results for five submitted systems in Albayzin 2010 are given in Table 16. Furthermore, a decomposition considering missed-speech detection, false alarms and false speaker labeling is also depicted in Figure 38. The best result of 30.4% DER was obtained by System 1, followed by similar performances of System 4, 3 and 5.

Rankings according to overall DER from the best to the worst: System 1, 4, 3, 5, and 2

Figure 38 indicates incorrect assigned speaker labels as the most significant proportion of the DER. The challenge seems to be the fact that many speakers speak only short segments of time, while a speaker may feature different background conditions.

The speaker error achieved by the first system is very remarkable, since all the clustering happens in only a single iteration. Furthermore,

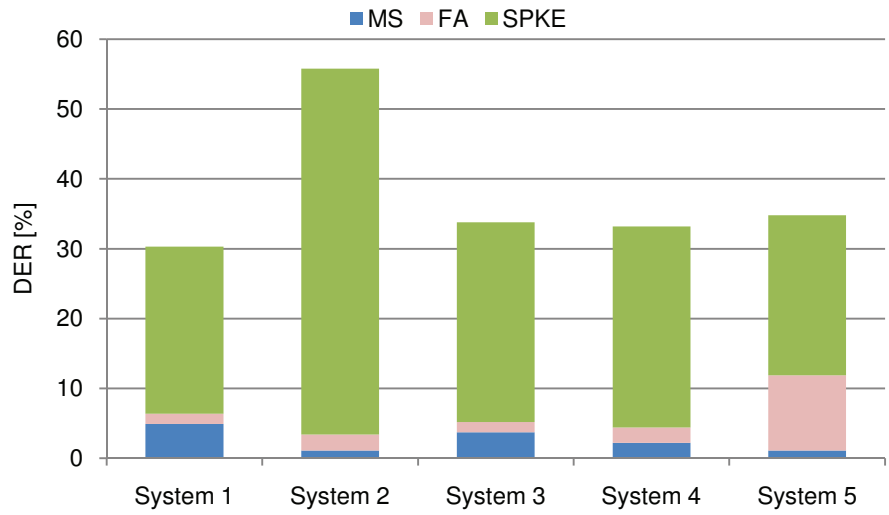


Figure 38: Overall speaker diarization results. DER distribution of missed speech rate (MS), false alarm rate (FA) and speaker error rate (SPKE).

Table 16: Speaker diarization results for all participants in terms of missed speech rate (MS), false alarm rate (FA), speaker error rate (SPKE) and diarization error rate (DER). All values are in given in (%).

Team	MS	FA	SPKE	DER
System 1	4.9	1.5	23.9	**30.4**
System 2	1.1	2.3	52.4	**55.8**
System 3	3.7	1.5	28.6	**33.8**
System 4	2.2	2.2	28.8	**33.2**
System 5	1.1	10.8	22.9	**34.7**

System 1 relies on the most popular approaches of the state-of-the-art diarization systems. Even though it is not possible to directly derive a conclusion from this result, the strategy to discard unvoiced frames in speaker change detection may have been the crucial factor of the best performance. The SAD of System 1 was tuned for hypothesizing more misses than insertions (false alarm).

The balanced and reliable SAD of System 4 and robust techniques applied for speaker segmentation resulted in the second best DER according to Table 16. System 3 also relied on a good operating SAD and the factor analysis technique used in speaker segmentation proved to be well-suited for this task. The overall DER and speaker error rate in particular were very similar for Systems 3 and 4.

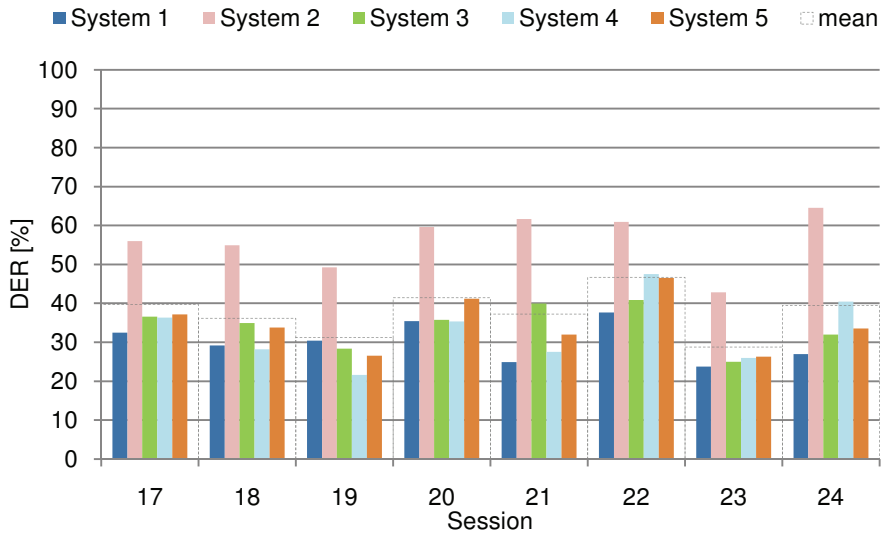


Figure 39: Speaker diarization performance for each of the eight testing recordings from 3/24 TV broadcast news corpus in terms of DER.

The factor analysis approach was also employed in System 5, which achieved the lowest speaker error with Viterbi decoding of iVector-stream scores over candidate clusters. It remains an open question, how the score normalization according to cluster appearance probability impacts the error rates.

System 2, with its hybrid ANN/HMM approach, displays the lowest error accounting to speech/non-speech detection, but it cannot benefit from this advantage in the overall performance. It is unclear what was the major reason for the higher overall DER score. It may have been the very simple initial speaker change detection, or the fingerprinting technique, which was observed to work well for audio segmentation [153], is not so appropriate for clustering speaker segments. Eventually, using the same set of acoustic features (and deltas) in all three stages of the process may not have been the optimal choice.

A more detailed analysis of the DER for each testing session shows (see Figure 39) that the recording hardest to diarize was the session 22, where almost all the evaluated systems obtained the worst result. Otherwise, the performance of the systems was rather stable. The DER standard deviation over the eight test recordings for each system lies between 4.6 and 8.0% DER. All systems were operating well (with respect to their average performance) in the test session 23. The absolutely lowest error of 21.6% DER was achieved by the System 4 on session 19.

Background-condition-specific evaluation

The speech signal can be divided according to acoustic background conditions into three categories: clean speech, speech over noise and speech over music. A particular difficulty of the diarization task is

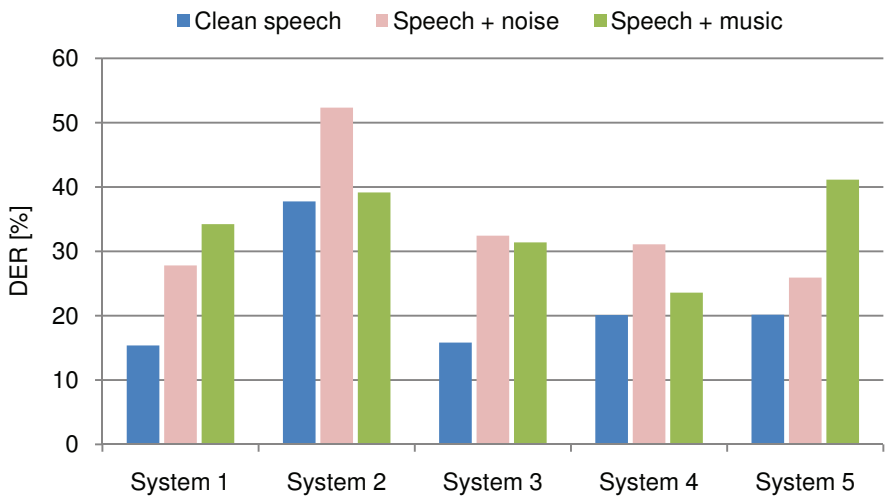


Figure 40: Speaker diarization performance in three acoustic background conditions: clean speech, speech with noise, and speech over music.

due to the nature of broadcast news data, which may exhibit different background conditions for one and the same speaker. It makes it very challenging for the clustering algorithm to put such speaker segments with different background conditions into the same cluster. By creating continuous chunks, which include only the segments of one speech class and non-speech segments (music, noise, silence), and computing their total duration, we can estimate how these three classes roughly contribute to the overall diarization error. Clean-speech, speech-over-noise and speech-over-music segments are influencing the DER by 36%, 46%, and 18%, respectively. Looking at the individual DER performances (evaluating each speech class independently), given in Figure 40, it is not surprising that the DERs of clean speech are usually the lowest.

Very different clustering stopping criteria between Systems 1, 2, 3 and Systems 4, 5

The operation of the systems in terms of detected speaker count is shown in Figure 41. Here, the Systems 5 and 4 exhibit the highest number of true detected speakers, but at the same time suffer from even higher counts of false speakers. The System 1, for instance, though detecting less correct speakers, maintains a significantly lower number of false speakers. Similar observation applies also for the operation of System 3.

The possible reason for the high number of false speakers of System 4 could be the substantial initial over-segmentation (reported in Section A.3) in a combination with a too strictly defined merging threshold of the dot-scoring similarity. Nevertheless, since the overall

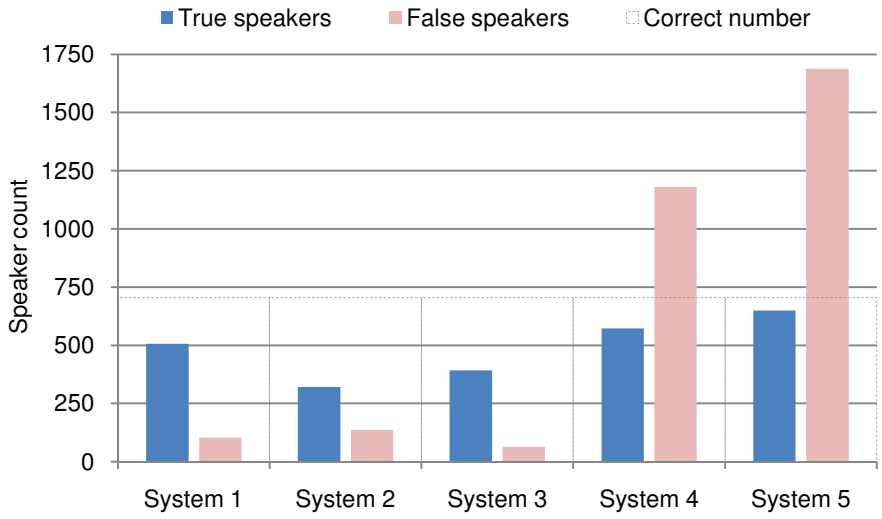


Figure 41: Number of correctly detected (True) and falsely introduced (False) speakers by the evaluated systems.

DER is not much different from System 1 or 3, the affected speaker segments were probably very short.

In the case of System 5, the probable cause of the high number of falsely detected speakers lies in the substantial false alarm rate (see Table 16) of the speech/non-speech detection rather than clustering algorithm, because the speaker error rate is very good compared with other systems.

A.5 DISCUSSION AND CONCLUSIONS

The analysis of speaker diarization results and the characteristics of the submitted systems revealed several observations which can be summarized as follows:

- The use of only voiced frames for performing speaker segmentation, which was implemented in one of the systems, seems a very interesting step in context of the very good speaker error result of that particular system.
- The speaker factor analysis technique, which received attention in the field of speaker verification, was successfully adopted in two presented diarization systems. Both of them delivered competitive results compared to the best system. This approach has the potential to become popular in speaker diarization also in the future.
- Almost all systems rely exclusively on MFCC features (13-19 coefficients) and for clustering also the derivatives can be used.

MFCCs are the very standard features for almost any kind of speech-related recognition task. One of the systems also included additional features, but the resulting performance has not proven them to be successful.

- BIC maintains as the most popular and effective cluster merging metric and/or clustering stopping criterion. It can be accompanied with other segmentation passes applying other metrics, but in all the cases the BIC is present at some point.
- All the systems used the conventional bottom-up agglomerative clustering approach. Even though it can sometimes suffer from merging instability or stopping criteria difficulties, it is usually robust and is also the most popular in other state-of-the-art systems.

The Albayzin 2010 speaker diarization evaluation results were presented for five of the six teams from four Spanish (EHU, UVigo, UZ, UAM) and one Portuguese (UC) university. The system which obtained the best result was also designed to run online and relies on modified growing-window BIC-based speaker-change detection and on a BIC-based clustering algorithm.

Conditions were probably more difficult than in comparable evaluations (NIST RT'04, ESTER)

The evaluation data turned out to be relatively challenging, since the DER results in other comparable evaluations, e. g., the NIST RT'04 evaluation [154] or the ESTER evaluation on French broadcast news [155], were considerably lower than in this case. The high number of speakers in Catalan TV 3/24 broadcast news corpus was perhaps also the reason why no system managed to determine the correct speaker count in neither recording.

BIBLIOGRAPHY

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] X. Zhu, C. Barras, L. Lamel, and J. Gauvain, "Speaker diarization: From broadcast news to lectures," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science, S. Renals, S. Bengio, and J. Fiscus, Eds. Springer Berlin / Heidelberg, 2006, vol. 4299/2006, pp. 396–406.

[3] S. E. Tranter and D. A. Reynolds, "Speaker diarisation for broadcast news," in *Proc. Odyssey-2004 The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp. 337–344.

[4] X. Anguera, C. Wooters, B. Peskin, and M. Aguiló, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science, S. Renals and S. Bengio, Eds. Springer Berlin / Heidelberg, 2006, vol. 3869, pp. 402–414.

[5] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. ASRU '07*, Kyoto, Japan, 2007, pp. 683–686.

[6] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *Proc. Interspeech '07*, Antwerp, Belgium, 2007.

[7] X. Anguera, C. Wooters, and J. M. Pardo, "Robust speaker diarization for meetings: ICSI RT06S meetings evaluation system," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science, S. Renals, S. Bengio, and J. Fiscus, Eds. Springer Berlin / Heidelberg, 2006, vol. 4299/2006, pp. 346–358.

[8] J. Psutka, *Komunikace s počítačem mluvenou řečí*. Praha: Academia, 1995.

[9] M. Gales and S. Young, "The application of hidden markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[10] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, 1st ed. Prentice Hall, May 2001.

[11] F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*, 1st ed. Springer, Dec. 2007.

[12] L. 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.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book, version 3.4*. Cambridge University Engineering Department, 2006.

[14] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[15] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2000.

[16] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[17] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

[18] J. Gauvain, L. F. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *Proc. 5th International Conference on Spoken Language Processing (ICSLP '98)*, Sydney, Australia, 1998, p. paper 0084.

[19] 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, Feb. 1997, pp. 97–99.

[20] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, St. Thomas, VI, USA, 2003, pp. 411–416.

[21] S. Meignier, J. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *Proc. 2001: A Speaker Odyssey - The Speaker Recognition Workshop (Odyssey-2001)*, Crete, Greece, 2001, pp. 175–180.

[22] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, "Segmentation of speech using speaker identification," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '94)*, vol. 1, Adelaide, Australia, 1994, pp. I/161–I/164.

[23] D. A. Reynolds, R. B. Dunn, and J. L. McLaughlin, "The lincoln speaker recognition system: NIST eval2000," in *Proc. ICSLP '00*, vol. 2, Beijing, China, 2000, pp. 470–473.

[24] D. Moraru, L. Besacier, and E. Castelli, "Using a priori information for speaker diarization," in *Proc. Odyssey-2004 The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp. 355–362.

[25] S. Meignier, D. Moraru, C. Fredouille, J. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 303–330, 2006.

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

[27] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–32, 1994.

[28] C. Nadeu, J. Hernando, and M. Gorricho, "On the decorrelation of Filter-Bank energies in speech recognition," in *Proc. 4th European Conference on Speech Communication and Technology (Eurospeech '95)*, vol. 20, Madrid, Spain, 1995, pp. 1381–1384.

[29] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. 2001: A Speaker Odyssey - The Speaker Recognition Workshop (Odyssey-2001)*, Crete, Greece, 2001, pp. 213–218.

[30] X. Zhu, C. Barras, S. Meignier, and J. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Proc. Interspeech '05*, Lisbon, Portugal, 2005, pp. 2441–2444.

[31] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP '03*, vol. 2, Hong Kong, China, 2003, pp. II–53–6.

[32] "Workshop 2002 - SuperSID: exploiting high-level information for high-performance speaker recognition." [Online]. Available: http://www.clsp.jhu.edu/ws2002/groups/supersid/

[33] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at

ICSI," in *Proc. 1st International Conference on Human Language Technology Research*, San Diego, CA, USA, Mar. 2001, pp. 1–7.

[34] D. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. ICASSP '05*, vol. 5, Philadelphia, PA, USA, 2005, pp. v/953–v/956.

[35] J. Žibert and F. Mihelič, "Fusion of Acoustic and Prosodic Features for Speaker Clustering," *Lecture Notes in Computer Science*, vol. 5729/2009, pp. 210–217, 2009.

[36] G. Friedland and O. Vinyals and 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, Jul. 2009.

[37] D. Imseng and G. Friedland, "Tuning-Robust initialization methods for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2028–2037, Nov. 2010.

[38] M. Yamaguchi, M. Yamashita, and S. Matsunaga, "Spectral Cross-Correlation features for audio indexing of broadcast news and meetings," in *Proc. Interspeech '05*, Lisbon, Portugal, 2005.

[39] O. Vinyals and G. Friedland, "Modulation spectrogram features for improved speaker diarization," in *Proc. Interspeech '08*, Brisbane, Australia, 2008, pp. 630–633.

[40] G. Lathoud and I. McCowan, "Location based speaker segmentation," in *Proc. International Conference on Multimedia and Expo (ICME '03)*, vol. 3, Baltimore, MD, USA, 2003, pp. III–621–4.

[41] J. Ajmera, G. Lathoud, and L. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Proc. ICASSP '04*, vol. 1, Montreal, Canada, 2004, pp. I–605–8.

[42] 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. Interspeech '06*, Pittsburgh, PA, USA, 2006, p. paper 1337.

[43] J. Schmalenstroeer and R. Haeb-Umbach, "Online speaker change detection by combining BIC with microphone array beamforming," in *Proc. Interspeech '06*, Pittsburgh, PA, USA, 2006, p. paper 1078.

[44] J. Luque, C. Segura, and J. Hernando, "Clustering initialization based on spatial information for speaker diarization of meetings," in *Proc. Interspeech '08*, Brisbane, Australia, 2008, pp. 383–386.

[45] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Proc. ICASSP '08*, Las Vegas, NV, USA, 2008, pp. 4133–4136.

[46] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[47] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *Proc. ICASSP '00*, vol. 3, Istanbul, Turkey, 2000, pp. 1423–1426.

[48] T. Hain, S. Johnson, A. Tuerk, P. Woodland, and S. Young, "Segment generation and clustering in the HTK broadcast news transcription system," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, 1998, pp. 133–137.

[49] J. F. López and D. P. W. Ellis, "Using acoustic condition clustering to improve acoustic change detection on broadcast news," in *Proc. ICSLP '00*, vol. 4, Beijing, China, 2000, pp. 568–571.

[50] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Springer Berlin / Heidelberg, 2008, vol. 4625/2008, pp. 509–519.

[51] H. S. M. Beigi, S. H. Maes, and J. S. Sorensen, "A distance measure between collections of distributions and its application to speaker recognition," in *Proc. ICASSP '98*, vol. 2, Seattle, WA, USA, 1998, pp. 753–756.

[52] B. Fergani, M. Davy, and A. Houacine, "Speaker diarization using one-class support vector machines," *Speech Communication*, vol. 50, no. 5, pp. 355–365, May 2008.

[53] G. Schwarz, "Estimating the dimension of a model," in *Annals of Statistics*, 1978, vol. 6, no. 2, pp. 461–464.

[54] ——, "A sequential student test," in *The Annals of Mathematical Statistics*, Jun. 1971, vol. 42, no. 3, pp. 1003–1009.

[55] S. Chen and P. Gopalakrishnan, "Clustering via the bayesian information criterion with applications in speech recognition," in *Proc. ICASSP '98*, vol. 2, Seattle, WA, USA, 1998, pp. 645–648.

[56] S. S. Chen, E. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanvesky, and P. Olsen, "Automatic transcription of broadcast news," *Speech Communication*, vol. 37, no. 1-2, pp. 69–87, May 2002.

[57] P. Delacourt, D. Kryze, and C. J. Wellekens, "Detection of speaker changes in an audio document," in *Proc. Eurospeech '99*, Budapest, Hungary, 1999, pp. 1195–1198.

[58] P. Delacourt and C. J. Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32, no. 1-2, pp. 111–126, Sep. 2000.

[59] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649–651, 2004.

[60] A. Tritschler and R. A. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Proc. Eurospeech '99*, Budapest, Hungary, 1999, pp. 679–682.

[61] X. Anguera, J. Hernando, and J. Anguita, "Xbic: nueva medida para segmentación de locutor hacia el indexado automático de la señal de voz," in *III Jornadas en Tecnología del Habla*, Valencia, Spain, 2004.

[62] M. Roch and Y. Cheng, "Speaker segmentation using the MAP-Adapted bayesian information criterion," in *Proc. Odyssey-2004 The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp. 349–354.

[63] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "MultiBIC: an improved speaker segmentation technique for TV shows," in *Proc. Interspeech '10*, Makuhari, Japan, 2010, pp. 2670–2673.

[64] A. Willsky and H. Jones, "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Transactions on Automatic Control*, vol. 21, no. 1, pp. 108–112, 1976.

[65] H. Gish, M. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proc. ICASSP '91*, vol. 2, Toronto, Canada, 1991, pp. 873–876.

[66] S. Kullback, *Information Theory and Statistics*. New York: Dover Publications, 1968.

[67] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Proc. ICASSP '98*, vol. 2, Seattle, WA, USA, 1998, pp. 757–760.

[68] "A tutorial on clustering algorithms." [Online]. Available: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/

[69] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-Multiple speaker clustering using HMM," in *Proc. ICSLP – Interspeech '02*, Denver, CO, USA, 2002, pp. 573–576.

[70] J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, "Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in On-Line broadcast," in *Proc. ICASSP '06*, vol. 5, Toulouse, France, 2006, pp. V–521–4.

[71] M. Betser, F. Bimbot, M. Ben, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs," in *Proc. Interspeech '04*, Jeju Island, Korea, 2004, pp. 2329–2332.

[72] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, vol. 40, no. 3, pp. 351–363, 2003.

[73] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, USA, 2004.

[74] S. E. Johnson and P. C. Woodland, "Speaker clustering using direct maximisation of the MLLR adapted likelihood," in *Proc. ICSLP '98*, vol. 5, Sydney, Australia, 1998, pp. 1775–1779.

[75] D. Moraru, S. Meignier, L. Besacier, J. Bonastre, and I. Magrin-Chagnolleau, "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation," in *Proc. ICASSP '03*, Hong Kong, China, 2003, pp. II–89–92. [Online]. Available: http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/fich_art/339-mor-icassp2003.pdf

[76] S. Bozonnet, N. Evans, C. Fredouille, D. Wang, and R. Troncy, "An integrated Top-Down/Bottom-Up approach to speaker diarization," in *Proc. Interspeech '10*, Makuhari, Japan, 2010, pp. 2646–2649.

[77] L. Canseco-Rodriguez, L. Lamel, and J. Gauvain, "Speaker diarization from speech transcripts," in *Proc. Interspeech '04*, Jeju Island, Korea, 2004, pp. 601–604.

[78] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of Multi-Party conversation," in *Proc. Eurospeech '01*, Aalborg, Denmark, 2001, pp. 1359–1362.

[79] E. Shriberg, "Spontaneous speech: how people really talk and why engineers should care," in *Proc. Interspeech '05*, Lisbon, Portugal, 2005, pp. 1781–1784.

[80] Özgür Çetin. (2007) Overlap in Meetings: ASR Effects and Analysis by Dialog Factors, Speakers, and Collection Site. [Online]. Available: http://videolectures.net/mlmi06_cetin_aeadf/

[81] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, Sep. 2006.

[82] M. Kashino and T. Hirahara, "One, two, many—Judging the number of concurrent talkers." *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 2596–2603, Apr. 1996.

[83] J. Cardoso, "Blind signal separation: Statistical principles," *Proceedings of the IEEE*, vol. 86, pp. 2009–2025, Oct. 1998.

[84] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," in *Neural Networks*, Jun. 2000, vol. 13, pp. 411–430.

[85] E. Weinstein, M. Feder, and A. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 405–413, 1993.

[86] D. Yellin and E. Weinstein, "Multichannel signal separation: methods and analysis," *IEEE Transactions on Signal Processing*, vol. 44, no. 1, pp. 106–118, 1996.

[87] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Kyoto, Japan, 1996, pp. 423–432.

[88] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[89] F. R. Bach and M. I. Jordan, "Blind one-microphone speech separation: A spectral learning approach," in *Advances in neural information processing systems*. Cambridge, MA, USA: MIT Press, 2005, vol. 17, pp. 65–72.

[90] L. Gu and R. M. Stern, "Single-Channel speech separation based on modulation frequency," in *Proc. ICASSP '08*, Las Vegas, NV, USA, 2008, pp. 25–28.

[91] S. T. Roweis, "One microphone source separation," *Advances in Neural Information Processing Systems*, no. 13, pp. 793–799, 2001.

[92] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *The Journal of the Acoustical Society of America*, vol. 60, no. 4, pp. 911–918, Oct. 1976.

[93] D. Morgan, E. George, L. Lee, and S. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 407–424, 1997.

[94] A. J. W. van der Kouwe, D. Wang, and G. J. Brown, "A comparison of auditory and blind separation techniques for speech segregation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 189–195, 2001.

[95] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *Proc. ASRU '01*, Madonna di Campiglio, Italy, 2001, pp. 107–110.

[96] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Feature selection for the classification of crosstalk in Multi-Channel audio," in *Proc. Eurospeech '03*, Geneva, Switzerland, 2003, pp. 469–472.

[97] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.

[98] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based fo extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, April 1999.

[99] K. Krishnamachari, R. Yantorno, J. Lovekin, D. Benincasa, and S. Wenndt, "Use of local kurtosis measure for spotting usable speech segments in co-channel speech," in *Proc. ICASSP '01*, vol. 1, Salt Lake City, UT, USA, 2001, pp. 649–652.

[100] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under Co-Channel conditions," *IEEE International Symposium Intelligent Sig. Process. and Comm. Systems*, pp. 710–713, 2000.

[101] R. E. Yantorno, K. R. Krishnamachari, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "The spectral autocorrelation peak valley ratio (SAPVR) – a usable speech measure employed as a Co-Channel detection system," in *Proc. IEEE International Workshop on Intelligent Signal Processing (WISP)*, Budapest, Hungary, 2001.

[102] M. A. Lewis and R. P. Ramachandran, "Cochannel speaker count labelling based on the use of cepstral and pitch prediction derived features," *Pattern Recognition*, vol. 34, no. 2, pp. 499–507, Feb. 2001.

[103] T. Nwe, M. Dong, S. Khine, and H. Li, "Multi-speaker meeting audio segmentation," in *Proc. Interspeech '08*, Brisbane, Australia, 2008, pp. 2522–2525.

[104] K. Laskowski, Q. Jin, and T. Schultz, "Crosscorrelation-based multispeaker speech activity detection," in *Proc. Interspeech '04*, Jeju Island, Korea, 2004, pp. 973–976.

[105] K. Laskowski and T. Schultz, "Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings," in *Proc. ICASSP '06*, vol. 1, Toulouse, France, 2006, pp. 993–996.

[106] G. Lathoud, I. A. McCowan, and D. C. Moore, "Segmenting multiple concurrent speakers using microphone arrays," in *Proc. Eurospeech '03*, Geneva, Switzerland, 2003, pp. 2889–2892.

[107] S. Otterson, "Use of speaker location features in meeting diarization," Ph.D. dissertation, University of Washington, 2008.

[108] K. Yamamoto, F. Asano, T. Yamada, and N. Kitawaka, "Detection of overlapping speech in meetings using support vector regression," in *Proc. IWAENC 2005*, Eindhoven, Netherlands, 2005.

[109] M. Kepesi, F. Pernkopf, and M. Wohlmayr, "Joint Position-Pitch tracking for 2-Channel audio," in *Proc. International Workshop on Content-Based Multimedia Indexing (CBMI '07)*, 2007, pp. 303–306.

[110] K. Boakye, "Audio segmentation for meetings speech processing," Ph.D. dissertation, University of California, Berkeley, 2008.

[111] B. Trueba-Hornero, "Handling overlapped speech in speaker diarization," Master's thesis, Universitat Politècnica de Catalunya (UPC), Barcelona; International Computer Science Institute (ICSI), Berkeley, May 2008.

[112] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization

in multiparty meetings," in *Proc. ICASSP '08*, Las Vegas, NV, USA, 2008, pp. 4353–4356.

[113] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech," in *Proc. Interspeech '08*, Brisbane, Australia, 2008, pp. 32–35.

[114] O. Ben-Harush, I. Lapidot, and H. Guterman, "Entropy based overlapped speech detection as a Pre-Processing stage for speaker diarization," in *Proc. Interspeech '09*, Brighton, UK, 2009, pp. 916–919.

[115] N. Sundaram, R. Yantorno, B. Smolenski, and A. Iyer, "Usable speech detection using linear predictive analysis – a model based approach," in *Proc. ISPACS*, Awaji Island, Japan, 2003, pp. 231–235.

[116] D. A. van Leeuwen and M. Huijbregts, "The AMI speaker diarization system for NIST RT06s meeting data," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science, S. Renals, S. Bengio, and J. Fiscus, Eds. Springer Berlin / Heidelberg, 2006, vol. 4299/2006, pp. 371–384.

[117] M. Huijbregts, D. van Leeuwen, and F. de Jong, "Speech Overlap Detection in a Two-Pass Speaker Diarization System," in *Proc. Interspeech '09*, Brighton, UK, 2009, pp. 1063–1066.

[118] NIST. (2009) The NIST Rich Transcription evaluation project website. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/

[119] D. A. van Leeuwen and M. Konečný, "Progress in the AMIDA Speaker Diarization System for Meeting Data," *Multimodal Technologies for Perception of Humans*, vol. 4625/2008, pp. 475–483, 2008.

[120] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1-3, pp. 117–132, Aug. 1998.

[121] A. Zolnay, R. Schlüter, and H. Ney, "Robust speech recognition using a Voiced-Unvoiced feature," in *Proc. ICSLP – Interspeech '02*, Denver, CO, USA, 2002, pp. 1065–1068.

[122] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, Sep. 1978.

[123] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on*

Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2011–2022, 2007.

[124] P. Svaizer *et al.*, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. ICASSP '97*, Munich, Germany, 1997, pp. 231–234.

[125] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. ICASSP '97*, Munich, Germany, 1997, pp. 375–378.

[126] T. Gustafsson and B. Rao and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 791–803, 2003.

[127] J. Luque, X. Anguera, A. Temko, and J. Hernando, "Speaker diarization for conference room: The UPC RT07s evaluation system," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Springer Berlin / Heidelberg, 2008, vol. Volume 4625/2008, pp. 543–553.

[128] S. Araki and M. Fujimoto and K. Ishizuka and H. Sawada and S. Makino, "Speaker indexing and speech enhancement in real meetings/conversations," in *Proc. ICASSP '08*, vol. 1, Las Vegas, NV, USA, 2008, pp. 93–96.

[129] M. Zelenák, C. Segura, and J. Hernando, "Overlap Detection for Speaker Diarization by Fusing Spectral and Spatial Features," in *Proc. Interspeech '10*, Makuhari, Japan, 2010, pp. 2302–2305.

[130] 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, Special Issue on New Frontiers in Rich Transcription*, 2011, accepted for publication.

[131] S. Otterson, "Improved location features for meeting speaker diarization," in *Proc. Interspeech '07*, Antwerp, Belgium, 2007, pp. 1849–1852.

[132] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 125–141, May 2008.

[133] A. Levy and M. Lindenbaum, "Sequential Karhunen-Loeve basis extraction and its application to images," *IEEE Transactions on Image Processing*, vol. 9, no. 8, pp. 1371–1374, 2000.

[134] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32/2000, pp. 1177–1207, 2000.

[135] E. Shriberg, A. Stolcke, and D. Baron, "Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation, disfluencies, and overlapping speech," in *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, USA, 2001, pp. 13–16.

[136] M. Zelenák and J. Hernando, "The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization," in *Proc. Interspeech '11*, Florence, Italy, 2011, pp. 1041–1044.

[137] M. Zelenák and J. Hernando, "Speaker Overlap Detection with Prosodic Features for Speaker Diarization," *IET Signal Processing*, 2011, in review.

[138] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[139] X. Anguera, C. Wooters, and J. Hernando, "Automatic cluster complexity and quantity selection: Towards robust speaker diarization ," in *Proc. Speaker Odyssey '06*, San Juan, Puerto Rico, 2006.

[140] A. Adami *et al.*, "Qualcomm-ICSI-OGI features for ASR," in *Proc. ICSLP – Interspeech '02*, Denver, CO, USA, 2002, pp. 21–24.

[141] J. Flanagan, J. Johnson, R. Kahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," in *J. Acoust. Soc. Am.*, vol. 78, No. 5, 1985, pp. 1508–1518.

[142] A. Temko, D. Macho, and C. Nadeu, "Enhanced SVM Training for Robust Speech Activity Detection," in *Proc. ICASSP '07*, Honolulu, HI, USA, 2007, pp. 1025–1028.

[143] G. Fung and O. Mangasarian, "Proximal Support Vector Machine Classifiers," in *Proc. KDDM*, 2001, pp. 77–86.

[144] The Augmented Multi-party Interaction project. (2011) Ami meeting corpus. [Online]. Available: http://corpus.amiproject.org

[145] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[146] A. Stolcke, X. Anguera, K. Boakye, O. Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "The SRI-ICSI spring 2007 meeting and lecture recognition system," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Springer Berlin / Heidelberg, 2008, vol. 4625, pp. 450–463.

[147] M. Zelenák, H. Schulz, and J. Hernando, "On the Improvement of Speaker Diarization by Detecting Overlapped Speech," in *Proc. FALA 2010*, Vigo, Spain, 2010, pp. 153–156.

[148] R. Barra-Chicote, J. M. Pardo, J. Ferreiros, and J. M. Montero, "Speaker Diarization Based On Intensity Channel Contribution," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 754–761, 2011.

[149] TALP. (2010) FALA 2010 Proceedings. [Online]. Available: http://fala2010.uvigo.es/images/proceedings/

[150] M. W. Wheeler, *The Phonology of Catalan*. Oxford, UK: Oxford University Press, 2005.

[151] M. Aguiló, T. Butko, A. Temko, and C. Nadeu, "A hierarchical architecture for audio segmentation in a broadcast news task," in *Proc. Workshop on Speech and Language Technologies for Iberian Languages*, Porto Salvo, Portugal, 2009, pp. 17–20.

[152] M. Zelenák, H. Schulz, and J. Hernando, "Albayzin 2010 Evaluation Campaign: Speaker Diarization," in *Proc. FALA 2010*, Vigo, Spain, 2010, pp. 301–304.

[153] T. Butko, C. Nadeu, and H. Schulz, "Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[154] J. Fiscus, A. Le, and G. Sanders. (2004) MDE Tasks and Results. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/2004-fall/rt04f-mde-nist.pdf

[155] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Proc. Interspeech '05*, Lisbon, Portugal, 2005, pp. 1149–1152.

DECLARATION

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Barcelona, Spain, October 2011

Martin Zelenák