Robust Speaker Diarization for Single Channel Recorded Meetings

Rong Fu

PhD

The University of York

Computer Science

January, 2011

Abstract

This thesis describes research into speaker diarization for recorded meetings. It explores the algorithms and the implementation of an off-line speaker segmentation and clustering system for meetings that have been recorded using one microphone.

Speaker diarization is defined as a process of partitioning a spoken record into speaker-homogeneous regions. The meeting record contains different kinds of noise and the length of the noise varies significantly. The average speech-turn is short and the number of speakers is unknown.

To reduce the influence of these aural characteristics on the performance of the speaker diarization system, this thesis proposed four new algorithms. First, a new speech activity detection method, which adjusts the non-speech model complexity according to the noise length ratio. Second, a new speaker change point detection measure was derived based on the Fisher Linear Discriminate Analysis to help detect short speaker turns. Third, the Equal Weight Penalty Criterion was formulated as a new model complexity selection criterion to train both the speakers' models and the Universal Background Model (UBM). It contains two penalty terms, one penalizes the model dimensions and removes mixtures with small mixing probability, the other penalizes the Kullback Leibler divergence between the prior and posterior distribution of the mixing parameters, removing those components that share the same location. This criterion can be adjusted

by the prior distribution parameter δ, which controls how many components are used in the model. Fourth, a weight and mean adaptation method was developed to adapt potential speaker models from the UBM. In addition, a potential speaker merging termination scheme, based on the Normalized Cuts, was introduced into the system.

Combining all the new techniques derived in this thesis together, the error rate of the baseline system was reduced from 18.61% to 9.24% on the development set, 18.89% to 10.50% on the evaluation set from AMI corpus, and 21.35% to 15.48% on the evaluation set from ISL corpus. When using the Normalized Cuts based potential speaker merging termination scheme, the error rate of the baseline system was reduced 18.61% to 10.33% on the development set, 18.89% to 9.99% on the evaluation set from AMI corpus, and 21.35% to 13.70% percentage points on the evaluation set from ISL corpus.

Contents

List of Tables

List of Figures

Acknowledgement

I would like to gratefully acknowledge the enthusiastic supervision of Dr. Ian Benest during my Ph.D study, without his supports, knowledge, and advices I would never have finished. In addition, I am grateful to my colleagues and friends who made my time at the University of York much easier and enjoyable.

Finally, I am deeply indebted to my husband, my parents and all families for their understanding, endless patience and encouragement when it was most required.

Declaration

Two papers have been published on this subjects, one in the proceeding of the "IEEE International Conference on Signal Processing and Multimedia Applications" (Fu and Benest, 2007a), and the other is in the proceeding of the "International Conference on INTERSPEECH 2007" (Fu and Benest, 2007b), both publications were peer reviewed conferences.

Chapter 1

Introduction

As processing power, storage capacity and network bandwidth increase, so grows the quantity of available information that can be stored and accessed by machines. That information can take the form of text (Salton, 2000)), audio (Tranter et al., 2004), graphics (Wang et al., 2008) and multimedia (Cortes, 2008). For humans to be able to cope with, and exploit, this 'information explosion' it is necessary for it to be indexed for ease of future retrieval, processed for different search strategies, and re-used so as to bring together fragments which have not hitherto been juxtaposed but which together can offer further insights into a topic. ASCII-based text has long been the target for indexing and retrieval techniques (Yu et al., 2004) and today that knowledge is helping to index multimedia material (Xu and Chang, 2008) (Bruno et al., 2008).

Arguably speech is the most popular form of expressive and exchangeable communication: used to perpetuate stories, to consolidate episodic memory, to bind people together. But it is not just the overt message contained in the speech that is important, but the hidden information that identifies the individual, their emotional state and the environment in which the message is spoken.

Speaker diarization is a process by which speaker information is extracted from an audio stream. In particular it attempts to identify who spoke when in a

conversation between two or more people. As such, the result of the diarization offers a pre-process for speech recognition, enabling the right template to be used in identifying the words spoken so as to enhance the recognition rate.

This chapter defines what is meant by speaker diarization (Section 1.1), how it is applied (Section 1.2) and outlines the problems that hinder the achievement of one hundred percent success (Section 1.3). Section 1.4 identifies the different types of discussions for which diarization may be usefully applied and finally section 1.5 introduces the strategic issues covered in the remaining chapters.

1.1 Acoustic diarization and speaker diarization

In general, a spoken document is a single-channel recording of a continuous speech stream that contains multiple audio sources (people and noise). Audio diarization is defined as the process of segmenting a spoken document into several clusters according to their different acoustic sources. The types and details of the acoustic sources vary according to the application. If the focus is to find the speech part in a spoken document, it will be segmented into speech and non-speech (silence, noise, music, etc) regions (Saunders, 1996). If the bandwidth (a measure of the width of a range of frequencies) of the conversation or the gender of the speakers need to be known, the spoken document will be divided according to the gender of each speaker or their conversation channels (Sinha et al., 2005). The most complicated application is to partition a spoken document into speaker-homogeneous regions. Within NIST Rich Transcription (NIST-RT) evaluation framework (Fiscus et al., 2005), this is what is meant by speaker diarization (Martin and Przybocki, 2001).

Speaker diarization provides the answer to the ‘Who spoke when’ question. That is why it is referred to as ‘unsupervised speaker segmentation and cluster-

ing' in some early documents (Zhou and Hansen, 2000), (Siegler et al., 1997). It consists of three subtasks. The first subtask is to detect where speaker changes occur in the given spoken document. The second subtask is to group the speech segments (a segment is a section of speech bounded by two speaker change points) from the same speaker together (speaker clustering). The third subtask is to estimate the number of speakers that contributed to the spoken document (the final number of the clusters). It is hoped that there is only one speaker's speech involved in each cluster, and a cluster contains all the speech of the corresponding speaker.

Usually there is no prior information provided about the speakers; for example, the number of speakers, their names, their gender, their speech samples, or their adjacency in the audio stream. This is what classifies the processing of the audio stream as 'unsupervised' and makes the speaker diarization task especially difficult.

1.2 Applications of speaker diarization

Early research focused on the audio transcription, derived from automatic speech recognisers. Later on, research concentrated on other aspects of audio information. Speaker information was extracted to facilitate the indexing and retrieval of audio documents, while non-speech information was detected to identify the structure of the spoken document. Beyond that, information linked to the spontaneous nature of speech was studied to understand speech communication behaviour.

Speaker diarization concerns speaker information, such as speaker turns, the number of speakers and the speakers' identities (to associate the 'relative' speaker label as 'speaker 1' or 'speaker 2', not the true speaker name). Speaker

diarization has six main applications.

- It helps to improve speech recognition performance. Speaker diarization provides speakers' locations and boundaries in a spoken document, which could be used within speaker adaptation and vocal track length normalization in speech recognition systems (Tranter et al., 2004) (Gupta et al., 2008). Furthermore, speaker information makes transcripts easier to read, since it identifies speech that enables the transcript to be turned into oral paragraphs.

- Speaker diarization enables speaker-based indexing and retrieval of a spoken document, as described in (Johnson and Woodland, 2000). It is also helps with determining other information, such as the speaker's gender and their true identity.

- Although speaker diarization usually deals with only one audio file with no prior information of the speakers, it facilitates other speaker indexing tasks such as speaker tracking (Tranter, 2006) and speaker tying (Tsai et al., 2007). Speaker tracking tries to explore all the occurrences of a particular speaker in an audio stream. Speaker tying is a classification process consisting of finding the number of speakers present in a collection of audio documents, then segmenting and clustering all the documents according to the speakers.

- Speaker diarization supplies useful information for detecting disfluencies and speaker overlaps, which directly link to the spontaneous nature of speech (Boakye et al., 2008) (Hung et al., 2008).

- Combined with speech recognition, high-level linguistic information, such as the speaker's name, the conversation topic and speaker's view, can be discovered (Tranter, 2006) (Ma et al., 2008).

- Speaker diarization, combined with various image processing techniques, helps to analyse video content, such as scene segmentation and classification, target object discrimination, etc (Liu et al., 1998) (Quenot et al., 2003).

1.3 Difficulties arising for speaker diarization

Depending on the nature and the environment of a spoken document, the speaker diarization process will encounter several difficulties. Because of the spontaneous nature of speech, hesitations, repetitions and overlaps always happen. The overlaps between speakers will confuse the recognizer system, and the hesitations in the speech will contaminate the speaker model. The number of speaker turns, and the length of each speech segment will also affect the speaker diarization results. When the speaker change frequency is high and the speech segments of each speaker are short, the speaker diarization task becomes more difficult. If some speakers talk much more or much less than others in an oral stream, it is hard to estimate the number of speakers present. The audio environment may also include music, non-verbal sounds such as paper shuffling and other extraneous sounds; all of which have a negative impact on performance. Finally, the more speakers present, the more difficult is the diarization process.

1.4 Different types of spoken documents

There are large volumes of spoken documents, including radio and television broadcasts, interviews, answer machine messages, telephone conversations, voice mails, meetings, etc. Among them, broadcast news, recorded meetings, and telephone conversations are the three primary domains used for speaker diarization research and development.

The data from these domains differ in the quality of the recordings, the environment where the speech happened, and the style of the speech. Telephone conversation is often recorded with a narrow bandwidth. The noise level is affected by the recording channel. Except for telephone meetings (including two speakers), the number of speakers involved is unknown. Broadcast news has various kinds of programming, usually containing commercial breaks and music. The recordings alternate between studio and outside broadcast, with different bandwidths. The speech in broadcast news is always well presented, with less overlap between two speakers. The number of speakers is unknown, and usually high. Sometimes there exist a few anchor speakers, but no dominant one (Gales et al., 2006) (Leeuwen, 2005). Only single channel recordings are available for broadcast news.

Meetings are recorded using table-top microphones, lapel microphones, or headset microphones. If a meeting is recorded with one microphone for each participant, the number of speakers is known and each microphone mainly captures the voice of a particular speaker. But the speaker diarization cannot be accomplished by a simple energy-based approach applied to each individual microphone because there is cross-talk between microphones (Pfau et al., 2001). Sometimes, recordings from each individual microphone can be combined and used to enhance the speaker diarization performance (Anguera et al., 2005).

This thesis is focused on the single channel recorded meeting using only a table-top microphone. Such meeting data contains several distortions arising from the microphones being distant from the speakers (Meignier et al., 2005). Moreover, the recorded meetings include informal, natural, and even impromptu meetings. The natural style of talking leads to plenty of speaker overlaps and frequent changes in speakers each with short segments. The number of speakers present in recorded meetings is also unknown, although it is limited by the size

of the meeting room. The noise contained in the recorded meetings is always impulsive, including laughing, breathing, clapping, coughing, doors shutting, pens falling, speakers touching their microphones, and so on.

Each domain presents unique diarization challenges and Table 1.1 summarizes the various difficulties encountered in each spoken document type.

	Telephone	**Broadcast news**	**Meeting**
Number of speakers	known	unknown, but high	unknown, limited by the room size
Length of segments	usually short	usually long	some really short
Changes in speaker	medium	low	high
Types of non-speech	noise	noise, music, commercial	various impulsive noises
Overlap	little	little	a lot
Quality of recording	low bandwidth	headset mic	distant tabletop mic
Disfluency	rarely	rarely	sometimes
Bandwidth	different setting	different setting	same setting

Table 1.1: Difficulties encountered with the three types of spoken document

In this thesis, the most difficult problem is of interest: the speaker diarization of single-channel recorded meetings, with no prior information of the number of speakers, their gender, etc. The meeting types include both formal meetings and natural meetings. Although sometimes prior knowledge enhances the speaker diarization performance, to make the system more robust and portable, no information in addition to the audio itself will be used in the proposed system. The implementation proposed in this thesis works towards creating a speaker diarization system that is insensitive to noise and to changes in the dataset; that is changing the value of the parameters slightly has no impact on system performance.

1.5 Thesis overview

This thesis is split into seven main chapters.

The primary literature on speaker diarization systems are reviewed in Chapter 2 to scope the research area of this thesis, provide a basic knowledge of acoustic feature extraction and speaker modelling techniques. A well-regarded system that is based on deep-rooted theory and adopts state-of-the-art techniques is adopted as a baseline system.

In Chapter 3, the shortcomings of each part of the speaker diarization system are identified, from the speech activity detection to the Universal Background Model training. The specifics of the meeting data that contribute to the difficulties incurred in speaker diarization are explained and several measures are developed to quantify the influence of these difficulties.

A new speaker change detection algorithm is developed in Chapter 4. Its performance is compared with some traditional speaker change detection measures, and the improvements are discussed.

In Chapter 5, a new criterion for model complexity selection will be developed. This new criterion can reduce intra-speaker variance when building speaker models or maintain inter-speaker variance during the Universal Background Model training by adjusting the prior distribution of the mixing parameters. The model complexity selection criterion proposed by Figueiredo and Jain (2002) can integrate the selection of the number of components into the EM training. This is applied at the model adaptation step to adjust the mean and weight value simultaneously from the UBM.

The experimental procedure that assesses all these novel technologies is described in Chapter 6. Their effectiveness, evaluated by comparing their results to the baseline system separately and in combination and the improvements, will also be presented. The results are analysed, and give a hint as to future work.

Finally, Chapter 7 summarizes the major conclusions and contributions obtained in this thesis and proposes some improvements and future work.

1.6 Toward a contribution

The objective of the thesis was to explore speaker diarization mechanisms with a view of contributing towards achieving perfect performance. The intention was to pinpoint the weaknesses in some of the current strategies and introduce alternative strategies with variations, all based on sound argument.

In this thesis, four new algorithms were proposed to improve the performance of the speaker diarization system. First, a new speech activity detection method was developed to cope with various impulsive noises in meetings. Second, a new speaker change point detection measure was derived to help detect short speaker turns. Third, the new model complexity selection criterion, Equal Weight Penalty Criterion, was formulated to train both the speakers' models and the Universal Background Model (UBM). The new criterion could reduce the model complexity to reduce intra-speaker variability and allow more model complexity in the UBM to capture more inter-speaker variability. Fourth, a weight and mean adaptation method was developed to adapt potential speaker models from the UBM. In addition, a potential speaker merging termination scheme, based on the Normalized Cuts, was introduced into the system.

Chapter 2

Literature Review

For more than a decade, speaker diarization processing has been used to facilitate speech recognition. Today, it is adopted as a means of indexing large speech databases. The requirement for enhanced recognition accuracy, a robustness to extraneous noise and adaptability in a variety of conditions, have all served to increase the difficulty in processing audio files successfully.

This chapter is a literature review of related research into speaker diarization that has been conducted in the last few years. First, background information about speaker recognition will be introduced. Various acoustic features used in speaker diarization will be explained in section 2.1 and speaker modelling techniques will be presented in section 2.2. Then the main steps of a speaker diarization system will be introduced. They are speech activity detection (SAD) (section 2.3), speaker change detection (SCD) (section 2.4), and potential speaker clustering (section 2.5). Next the strategies to combine the results of different diarization systems will be given in section 2.6. Finally, the baseline system that was used in the research described in this thesis will be illustrated in section 2.8.

2.1 Front-end processing

The front-end is a generalized term that refers to the initial stages of a process. In speaker recognition, front-end refers to the part that converts a continuous speech stream into a sequence of acoustic feature vectors. In this section, the speech production mechanism will be first introduced (section 2.1.1). Then how to extract the speaker-dependent information from the speech waveform will be described in section 2.1.2. Next, the most popular acoustic parameters used for speaker diarization processing, the Mel-Frequency Cepstrum Coefficients (MFCC), will be explained in section 2.1.3. Some other acoustic parameters that are usually applied in combination with MFCC will be given in section 2.1.4. Finally, the features used in speaker diarization systems will be reviewed in section 2.1.5.

2.1.1 Speech production mechanism

Speech is produced as a result of the acoustic excitation of the vocal tract. The excitation comes from a series of nearly periodic pulses generated by the vocal cord or the turbulent flow of air. Then it is constrained by the vocal tract, which can be thought of as an acoustic tube which continually changes its shape during speech production. Finally the produced speech is radiated from the lips, or from the nostrils in the case of nasal consonants. The resulting speech can be described by a waveform, plotting the instantaneous amplitudes of a periodic quantity against time.

2.1.2 Speaker characteristics and their representation

For speaker recognition, it is necessary to find in the speech those factors which convey speaker-dependent information. First, the anatomical details of the vocal tract vary considerably from one person to another. Such differences result from

the fixed structural differences such as the mass of the vocal cord, the size of the mouth, the shape of the tongue, the position of the pharynx, etc. Second, the differences in speaking habits of different individuals are an important source of inter-speaker variation. The differences in the speaking habits result from the manner in which people use their speech mechanism, such as intonation patterns, speaking rates, and so on. Such differences are produced by the acoustic move and are seen in the temporal variations. In speaker recognition, both anatomical and speaking habit differences are exploited to distinguish the speech of one speaker from another.

To extract speaker-dependent parameters that reflect fixed anatomical properties of the vocal tract, the time-invariant parameters are ideal because of their independence of the spoken message. On the other hand, idiosyncrasies in the speaking habits of individuals by nature vary from one sound to another, and hence cannot be represented in a time-invariant style. For most sounds, the shape of the vocal tract changes slowly compared to the excitation vibrations, so the speech production can be considered to be in a quasi-stationary mode. As a result, when examined over a sufficiently short period of time (between 5 and 100 milliseconds), speech characteristics stay fairly constant. However, over longer periods of time (0.2s or more), they change to reflect high-level characteristics, in the form of linguistic information. Consequently, it is possible to carry out a spectral analysis over a short period (20ms-30ms), which determines speech characteristics in the frequency domain. This efficient way to describe all the acoustic characteristics of speech is called a short-time spectrum. It provides a three-dimensional representation of the speech signal, the coordinates being time, frequency, and energy. While the short-time speech characteristics are presented by the spectrum of each short time interval, the time-varying characteristics can be obtained by averaging over time.

A wide range of features that are related to some property of the short-time power spectrum, such as Linear Predictive Coding Coefficients (Atal and Hanauer, 1971), MFCC (Mermelstein, 1976), principal components of the spectra (Bridle and Brown, 1974), Perceptual Linear Prediction (Hermansky, 1990b), Representation Relative Spectra (Hermansky, 1990a) and so on, have been investigated in automatic speaker recognition application.

2.1.3 Mel-Frequency Cepstrum Coefficients

By approximating the human auditory system's response, MFCC is perhaps the best-known and most popular set of acoustic parameters for both speech recognition (Zheng et al., 2001) and speaker diarization (Davis and Mermelstein, 1976). Instead of the linearly-spaced frequency bands, MFCC extracts acoustic parameters on the Mel-scale.

After being read by the computer, the audio stream is sampled at regular time intervals, forming a sequence. The sampling rate f_s (the number of samples obtained in one second) is fixed during a sampling process and is usually 16kHz. To transform this time-sampled, discrete waveform into a short-time spectrum, the sequence of discrete samples need to be divided into many overlapped short time frames. Every frame has the same time length, usually 20ms, with an overlap of 10ms with the prior block. The signals in each frame are multiplied with a Hamming window and then transformed into the frequency domain by applying a Fast Fourier Transform (FFT). The spectrum of each frame is then filtered by a collection of triangular filters and the log energy outputted by each filter is calculated. Transforming all the log energy back into the time domain using the Discrete Cosine Transform (DCT), the MFCCs are obtained. For each frame, the dimension of the MFCCs is determined by the number of filters. These filters are spaced according to the Mel-scale (Beranek, 1949), in which a linear frequency

spacing is adopted below 1000 Hz and a logarithmic spacing above 1000 Hz. Equation 2.1 shows how to approximate the frequency in the Mel-scale f_{mel} using the normal frequency f_c; and Figure 2.1 displays a Mel-scale filter bank that contains 30 triangular filters.

$$f_{mel} = 2595 \log_{10}(\frac{f_c}{700} + 1) \tag{2.1}$$

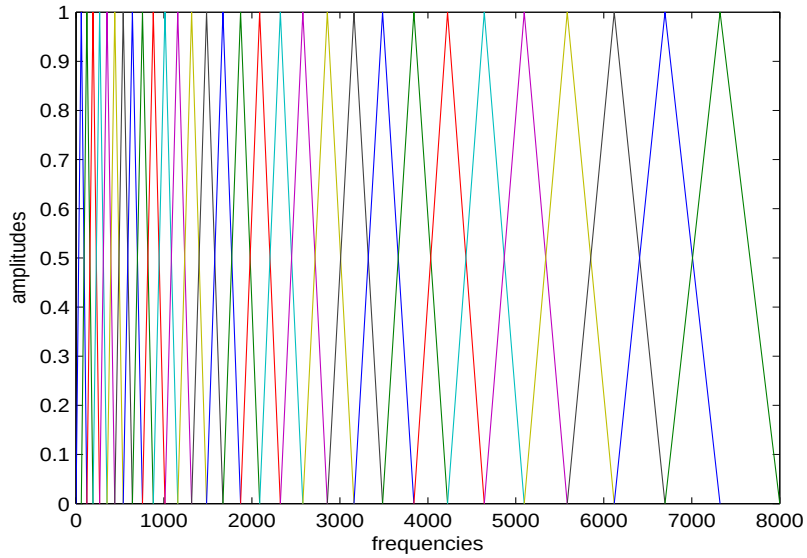


Figure 2.1: The mel-scale filter bank that contains 30 triangular filters which are spaced between 0Hz and 8kHz.

The whole process of extracting MFCC features is illustrated in Figure 2.2. Given the frame size and the overlap between frames, it is simple to compute how many frames are contained in a time interval. If the frame size is 20ms and the overlap is 10ms and the speech lasts one second, then there will be one hundred frames, and 100 MFCCs will be extracted.

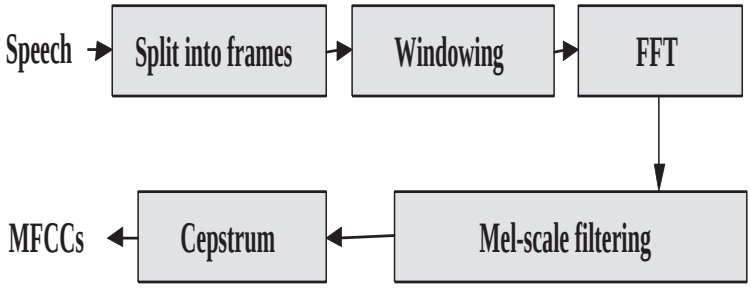


Figure 2.2: Block diagram of the MFCC processor

2.1.4 Other acoustic parameters

One of the simplest characteristics of any signal is its log-energy. For a frame n, the energy vector $e(n)$ is one-dimension and defined by Equation 2.2:

$$e(n) = \log(\sum_{t=1}^{T} o(t)^2). \tag{2.2}$$

where o(t) is the tth discrete signals in the frame n and T is the size of frame n.

The first and second differential coefficients of MFCC together with this log-energy feature are widely used as speaker acoustic features. And all have the same dimension as the features that are differentiated. The log-energy feature and the the first and second differential features are always used in combination with the MFCC.

Throughout the thesis, $\bar{D}$ is used to refer to the dimension of the feature vectors, and N is the size of the feature vectors. the frame size will be set to 30ms and the overlap will be set to 20 ms.

2.1.5 Features used in speaker diarization systems

The MFCC features are considered to be very effective for speaker recognition because they are obtained by spectrum analysis and the spectrum reflects speakers' predominant physiological characteristics (the vocal tract structure). The

number of MFCC features are generally different in speaker diarization systems.

For example, 12 MFCCs, the mean-normalized log energy and their first and second differential coefficients are extracted as the acoustic feature vectors in the HTK broadcast news transcription system. The dimension $\bar{D}$ of the feature vectors is 39. In contrast in the LIMSI broadcast news transcription system (Meignier et al., 2005), the log energy feature was not included, resulting in a dimension of 38. In (Anguera et al., 2006a), only 19 MFCCs were used without deltas (the divergence features). The PLP feature vectors were used in (Tranter et al., 2004). Recently, long term speaker features, like pitch, vocal source, and prosodic features, were applied for speaker diarization (Yamaguchi et al., 2006) (Chan et al., 2006) (Friedland et al., 2009). Sometimes, feature vectors are projected into a lower dimension space prior to the clustering step (Tsai and Cheng, 2006).

2.2 Speaker modelling

When two people utter the same words, the variations in the speech fundamentally originate from the difference between the speakers' voices. When a person utters two sequences of different words, the variations of the speech essentially come from the difference between the two sequences of phonemes. Even when the same speaker utters the same word twice, variations occur. This can be caused by many factors such as the speaking rate, the emotional state of the person of the person, and so on. These last two variations are referred to as intra-speaker variations. If two utterances with the same words are compared in order to determine whether they are from the same speaker or not, the task is called text-dependent speaker recognition. However, the most general speaker recognition task is to recognize a voice whatever is spoken and whenever it is said. This

task is more difficult because inter-speaker variations must be detected without being confused by intra-speaker variations. It is called text-independent speaker recognition. The speaker diarization task is text-independent.

The spectrum acoustic parameters convey not only the speaker-dependent information, but also phonetic information and environmental conditions. Therefore, various speaker-modelling techniques are introduced to represent speaker-dependent information over the long term. The more data from the same speaker that is included to build the model, the better it discriminates one speaker from another. The GMM (Hansen, 1982), the Vector Quantization (VQ) codebook, the tied GMM, the Radial Basis Function (Poggio and Girosi, 1990) and the Multilayer Neural Network (Rumelhart et al., 1986) have all been applied in modelling the speaker (Reynolds and Rose, 1995) (Matsui and Furui, 2004) (Reynolds, 2002) (Farell et al., 1994). GMM, which is the most popular and flexible, was used both in speaker recognition (Reynolds, 2002) and speaker diarization (Tranter and Reynolds, 2006). Recently, the hybrid systems of Support Vector Machine (SVM) (Boser et al., 1992) have been successfully adopted for both speaker verification and speaker recognition (Kharroubi et al., 2001) (Fine et al., 2001) (Wan and Renals, 2005a).

The GMM model will be described in the next section 2.2.1. Then, two principal motivations for using Gaussian mixture densities as a representation of the speaker characteristics will be given in the section 2.2.2 that follows. Finally, some relevant algorithm issues, such as parameter estimation, initialization, and how the model order is determined will be introduced in section 2.2.3.

2.2.1 Gaussian Mixture Model (GMM) description

GMM are the most widely used mixture model, and is a weighted mixture of a number of Gaussian components. With an appropriate number of components,

GMM has the ability of forming smooth approximations to arbitrarily-shaped densities (Reynolds and Rose, 1995). It can be described by Equation 2.3:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g_i(x) \tag{2.3}$$

where x is a given feature vector with dimension $\bar{D}$; λ contains all the the parameters in the model; $p(x|\lambda)$ is the probability of the appearance of x given the model. M is the number of components in the model and w_i is the weight of the component i, which must satisfy the conditions that $w_i \leq 1$ and $\sum_{i=1}^{M} w_i = 1$. $g_i(x)$ is a component of the GMM, and is a multivariate Gaussian function of the form 2.4:

$$g_i(x) = \frac{1}{(2\pi)^{\bar{D}/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right\} \tag{2.4}$$

for $1 \leq i \leq M$, where μ_i and Σ_i are the mean and covariance of the Gaussian component i. The parameters in a GMM can be collectively represented by the notation $\lambda = \{\mu_i, \Sigma_i, w_i\}$ where $i = 1, \cdots, M$. μ_i has the same dimension as x and is the mean vector of the component i. The mean vector controls a component's position among other components. Σ_i is a $\bar{D} * \bar{D}$ matrix, which is the covariance matrix of the component i. The shape of each component is decided by its covariance matrix.

2.2.2 Motivation Interpretation

The speech contains broad phonetic events. The production of speech can be divided into three classes: voiced sounds, unvoiced sounds and plosive sounds. They can be further separated as vowels, semivowels, voiced stops, nasals, voiceless stop consonants, stop consonants, and various fricatives. These phonetic events may characterize the sub-spaces of the acoustic space of a speaker's voice

(Reynolds and Rose, 1995). Because all the training or testing speech is unlabelled, these sub-spaces and acoustic classes are 'hidden' - and are therefore unknown. These sub-spaces cannot be directly mapped to their various monophones (Reynolds and Rose, 1995).

GMM is a semi-parametric probabilistic density and provides great flexibility and precision in modelling the underlying statistics of sample data. Assuming the independence of the feature vectors, the components contained in a GMM are suitable for modelling a wide range of hidden acoustic classes. Speaker characteristics, such as the shape of the vocal tract, are contained in these acoustic classes, and will be represented by the mean vector of the component and the intra-speaker variation will be captured by the covariance matrix (Reynolds and Rose, 1995). Also, because the component Gaussians are acting together to model the overall probability density function, any inaccuracy due to single components will be compensated by the whole model.

2.2.3 Algorithm issues

Given training feature vectors of a speaker, the goal of speaker model training is to estimate the parameters of the GMM, $\lambda = \{\mu_i, \Sigma_i, w_i\}$ where $i = 1, \cdots, M$. The estimated parameters, in some sense, need to make the GMM a best match to the true distribution of the feature vectors. To minimize the training errors is thought to be consistent with minimizing the difference between the model and the true distribution. There are several techniques available for estimating the parameters of a GMM. By far the most popular and well-established method is the Expectation-Maximization (EM) algorithm, which approximates the Maximum Likelihood (ML) estimates of the parameters.

The aim of ML estimation is to find the model parameters which maximize the likelihood of the training data, given the GMM. For a sequence of N feature

vectors, $X = x_1, x_2, \cdots, x_N$, their likelihood in a GMM model can be described by Equation 2.5:

$$p(X|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda) \tag{2.5}$$

where $p(x_n|\lambda)$ is given by Equation 2.3. Usually the ML estimation of the parameters can be obtained by solving Equation 2.6:

$$\partial \log(p(X|\lambda))/\partial\lambda = 0. \tag{2.6}$$

Unfortunately, this expression with respect to the covariance parameters is a nonlinear function so Equation 2.6 cannot be solved directly. The EM algorithm solves this problem iteratively, by monotonically increasing the value of $\log(p(X|\lambda))$ at each step.

The EM algorithm is widely used to obtain both the ML estimates and the maximum a posteriori (MAP) estimates in various applications, including the Hidden Markov Model (HMM) (McLanchlan and Basford, 1988). The detailed steps for the EM algorithm as it is applied in the GMM training process will be given in Chapter 5.

Two critical factors in training a Gaussian mixture speaker model are selecting the complexity M of the mixture (the number of components contained in the GMM) and initializing the model parameters λ. A random initialization method, which randomly chooses M vectors from a speaker's training data as the means of the components, and uses the identity matrix as the starting covariance matrix, is widely used. This method is thought to be simple and computational efficient. However, it does not guarantee a global optimum solution.

Determining the number of components M in a mixture that can model a speaker adequately is an important but difficult problem. There is no theoreti-

cally determined way to estimate the number of mixture components in a GMM. For speaker modelling, the objective is to choose the appropriate number of components to capture adequately the speaker's characteristics. Either too few or too many mixture components will affect the GMM's ability to capture the distinguishing characteristics of the speaker.

In order to train a GMM that reliably models the characteristics of a speaker, adequate training data is necessary. In speaker diarization, however, there is no labelled training data available. Moreover, some speaker utterances last less than one second and sometimes the GMM needs to be trained on small data collections. Hence the model complexity selection influences the success of the process.

2.3 Speech activity detection

The aim of SAD is to find the speech regions in an audio stream. The speech in a stream may overlap with other sounds, such as music and noise. During the speech activity-detecting process, all the portions containing speech will be retained, while the non-speech portions will be discarded. Removing the non-speech parts will reduce the processing time of speaker modelling, and improve speaker diarization performance because it increases the efficiency of speaker modelling. If the data obtains a number of different sorts of noise, the speaker models will be contaminated and distorted.

The non-speech in broadcast news could be categorized into three types, silence, music, and noise (Tranter and Reynolds, 2006). The noise class is composed of any event occurring in the signal that could not be categorized as silence, music or speech. Music is not a common type of non-speech in the meetings. Silence portions in the audio can be detected by energy-based threshold and zero-

crossing rate. When it comes to the other types of non-speech, more complicated methods are in need. The general approach used is maximum-likelihood classification with GMMs. GMMs are trained to represent different acoustic conditions. Usually, two GMMs are separately trained for speech and non-speech (Wooters et al., 2004). However, in some work, GMMs were trained separately for all kinds of non-speech(Gauvain et al., 1998) (Reynolds and Carrasquillo, 2004) (Sinha et al., 2005).

Given two trained GMMs, one for the speech and the other for the non-speech each feature vector x_n extracted from the audio file will be assigned to the model where it represents the maximum likelihood according to Equation 2.7.

$$\hat{k} = \arg\max(\log p(x_n|\lambda_{speech}), \log p(x_n|\lambda_{non-speech}))1 \leq n \leq N, \qquad (2.7)$$

where λ_{speech} and $\lambda_{non-speech}$ are the GMM models for speech and non-speech separately. $p(x_n|\lambda_{speech})$ is the probability of x_n present in the speech model calculated by Equation 2.3. $\hat{k}$ is the selected acoustic cluster of x_n, in this case the speech or the non-speech. Due to the continuity of speech, this classification result needs to be smoothed over several frames (Siegler et al., 1997) (Reynolds and Carrasquillo, 2004).

In some work, the detected speech and non-speech were passed through some heuristic rules so as to refine their boundaries (Reynolds and Carrasquillo, 2004). As well as the GMM maximum likelihood classifier with smoothing window, the HMM model is also widely used for acoustic classification; its transition parameters can be used to control the speech length (Tranter and Reynolds, 2006). A hybrid approach that combines the energy-based noise detector and GMM-based clusters was proposed to detect noise during meetings (Li et al., 2002); The speech and non-speech detected by an energy-based detector was then used to train the speech and non-speech GMMs.

Missing speech (MISS) and false alarm (FA) are the two measures to evaluate the speech activity detection performance. Missing speech corresponds to those portions of the audio that are speech, but recognized by the detection process as non-speech. False alarm, on the other hand, contains these portions of the audio that are non-speech, but recognized by the system as speech. Speech detection errors include both miss and false alarm errors. The detection error rate is the percentage of the time that all the error portions occupy in the whole audio. Generally, it is more important to minimize the missing speech, because they want to enhance the speed.

Sinha et al. (Sinha et al., 2005) and Zhu et al. (Zhu et al., 1998) applied a word recognizer to remove the non-speech parts. However, as many speaker diarization systems adopt speech activity detection to facilitate speech recognition, they are not available at this stage. For the speaker indexing task, it is unnecessary to include a complicated speech recognition system. If the audio is recorded in multiple channels by individual microphones, the recording of these channels can be combined to enhance the speech signal and remove non-speech portions (Pfau et al., 2001) and (Anguera et al., 2005). The meeting recorded by multiple microphones can also be used to detect the position of the speakers (Pfau et al., 2001) (Pertila and Parviainen, 2007) (Brutti et al., 2007) (Brutti et al., 2008a).

2.4 Speaker change detection

There are three essential subtasks contained in the speaker diarization process: SCD, clustering, and estimating the number of the speakers. The SCD (also referred to as speaker segmentation) produces a sequence of utterances with the same speaker within each one. The boundaries between such utterances, where the speaker changes, are called the speaker change points.

Traditionally SCD do not cut words in half and so most change points are hypothesized to happen within silence. Some energy-based change detectors analyse the energy waveform and use a threshold to find the points where a speaker change is most likely to exist (Kemp et al., 2000) (Nishida and Kawahara, 2003). A decoder-guided change point detector, in contrast, runs a full speech recognition process to obtain the change points by forced alignment (Liu and Kubala, 1999), (Kubala et al., 1997). However, there is no clear relationship between the existence of a silence in a recording and a change of speaker. The voice might be overlapped between different speakers and long pauses may happen during one person's speech. Moreover, music or commercial might be played as the background sound when speakers change, instead of silence.

Some systems detect the change in various acoustic conditions (telephone bandwidth, speaker gender, music/speech/noise) instead of speakers (Gauvain et al., 1998) (Ajmera et al., 2002) (Ajmera and Wooters, 2003). For this kind of system, prior information is required to train the models for different acoustic conditions and only some of speaker changes can be discovered; there is no guarantee that a speaker change happens when there are changes in the acoustic condition.

Other than the energy-based SCD and acoustic model-based SCD algorithm, a metric based SCD detects changes depending on the distance between two adjacent segments. To detect if the speaker changes at a point, a window is located around the point and the feature vectors in this window are separated into two parts, one before the point, and the other after it. Then the distance between these two parts are measured and a threshold is set. If the distance is larger than a specific threshold, this point is the change point, otherwise it is not. Various distance matrices, such as the Bayesian Information Criterion (BIC) (Schwarz, 1978), the Kullback-Leibler Divergency (KL) (Kullback and Leibler, 1951), the General-

ized Likelihood Ratio (GLR) (Willsky and Jones, 1976), the Gish distance (Gish et al., 1991), the Divergence Shape Distance (Lu and Zhang, 2002), the Cross Likelihood Ratio (CLR) (Mood et al., 1974), the Malalanobis distance (Mahalanobis, 1936) and the Kolmogorov-Smirnov test (Kolmogorov, 1933)(Smirnov, 1948), have been applied to detect the change points (Chen and Gopalakrishnam, 1998) (Siegler et al., 1997) (Zhou and Hansen, 2000), (Sinha et al., 2005), (Barras et al., 2004) (Gauvain et al., 1998) overview1-30-4 (Gish et al., 1991) (Lu and Zhang, 2002) (Anguera et al., 2005) (Wooters et al., 2004) (Campbell, 1997) and (Deshayes and Picard, 1986). One class-SVM and SVM supervised classification errors have also been used as distance measures between two segments. The optimum value of thresholds are usually selected depending on training data sets (Kadri1 et al., 2008) (Wan and Renals, 2005b).

Metric based SCD is probably the most used technique to date. Among them, the BIC distance and the KL2 distance are popular for their computational efficiency and good performance (Tranter and Reynolds, 2006). These two SCD algorithms will be introduced in the next section (2.4.1) and the evaluation of the task will be presented in the one that follows.

2.4.1 BIC and KL2

Bayesian Information Criterion (BIC) is a model selection criterion applied to choose one among a set of candidate models to represent a given data set (Schwarz, 1978). These models are trained maximizing the likelihood of the training data fitting the models, as computed by Equation 2.7. It is evident that when the number of parameters used in the model increases, the model fits the dataset better. However, when the parameters contained in the model become too large, there is over-training. BIC penalizes the model by its complexity - the number of parameters included in the model.

Let $L(X|\lambda_M)$ be the log likelihood of data set $X = \{x_n|n = 1, \cdots, N\}$ given the model whose model complexity is M, as describe by Equation 2.5. The BIC score of the model with model complexity M is calculated according to Equation 2.8:

$$BIC(M) = L(X|\lambda_M) - 1/2\phi M \log N \tag{2.8}$$

where ϕ is a constant modified by experiment. Among a series of models, the BIC criterion prefers the model that maximizes the BIC score.

As introduced in (Chen and Gopalakrishnam, 1998), for SCD, one Gaussian model is used for representing the data set and the window size is initialized at two seconds and located at the beginning of the feature vectors (Chen and Gopalakrishnam, 1998). For each point in the window, BIC is used to check if this point is a change point. Denote the Gaussian trained using the feature vectors before the point as λ_b, the Gaussian trained using the feature vectors posterior to the point as λ_p, and the Gaussian trained using all the feature vectors in the window as λ_f. If a change truly happens, the data is better to be represented by two models, λ_b and λ_p, otherwise a single model λ_f is preferred. To compare their BIC score, Equation 2.9 is applied:

$$\begin{aligned} \Delta BIC &= BIC(\lambda_b) + BIC(\lambda_p) - BIC(\lambda_f) \\ &= L(X|\lambda_b) + L(X|\lambda_p) - L(X|\lambda_f) - 1/2\phi \log N_f[M_b + Mp - M_f] \\ &= 1/2[N_f \log(|\Sigma_f|) - N_b \log(|\Sigma_b|) - N_p \log(|\Sigma_p|)] - \frac{1}{2}\phi \Delta M \log N_f] \end{aligned} \tag{2.9}$$

where N_b, N_p and N_f are the number of feature vectors used to train the parameters λ_b, λ_p, and λ_f and the Σ_b, Σ_p, and Σ_f are their covariance matrices. $\Delta M = (\bar{D}(\bar{D}+3)/2)$, where $\bar{D}$ is the feature vectors' dimension. If $\Delta BIC > 0$, this point is a change point. If a change point is discovered in the window, a new

window is started at the change point and the BIC test is done again. Otherwise, the window is enlarged to include another one second speech and the test is repeated. If the boundaries of the window are near a change point, it is difficult to detect this point because not enough data is available for model training. ϕ can be simply set as one, but it is better to be tuned by the development dataset.

Using the BIC criterion to search for the change point exhaustively is time consuming as it must be tested for each point. To speed up the algorithm, the test can be run every 30 feature vectors. To avoid the computation of three full covariance matrices, Hotelling's T^2 statistics were applied to accelerate the searching (Zhou and Hansen, 2000). Using only the mean value and a shared covariance matrix, T^2 statistics quickly select one candidate change point in a window, and BIC is applied to reject false candidates. Controlling the window size dynamically and overlooking the points near the window boundaries are other efficient ways to speed up the BIC based search.

The KL divergence (also referred to as relative entropy), is an unsymmetric measure of the difference between two probability distributions P_1 and P_2. The KL divergence of P_2 from P_1, denoted as $D_{KL}(P_1||P_2)$, is the expected value of their entropy with respect to the distribution of P_1. It is formulated in the following way:

$$D_{KL}(P_1||P_2) = \int_{-\infty}^{+\infty} p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \tag{2.10}$$

The larger this value, the greater the distance between probability densities of the two random variables. Because $D_{KL}(P_1||P_2)$ is not equal to $D_{KL}(P_2||P_1)$, a symmetric measure $KL2$ is introduced to measure the distance between the two

densities. $KL2$ is defined by Equation 2.11 as:

$$KL2 = D_{KL}(P_1||P_2) + D_{KL}(P_2||P_1) \tag{2.11}$$

Using $KL2$ to measure the model distance, a fixed window with two second length is used. Two Gaussian models are trained on this window. Model $\lambda_b = (\mu_b, \Sigma_b)$ is trained on the first half of the window and model $\lambda_p = (\mu_p, \Sigma_p)$ is trained on the second half of the window. Then KL2 is used as the distance measure between the two models to decide if a speaker change happened at the middle point of the window. It is described by Equation 2.12

$$KL2 = tr(\Sigma_b^{-1}\Sigma_p) + tr(\Sigma_p^{-1}\Sigma_{b)} + (\mu_b - \mu_p)^T(\Sigma_b^{-1} + \Sigma_p^{-1})(\mu_b - \mu_p) - 2\bar{D} \tag{2.12}$$

where $tr(\Sigma)$ takes the trace of the matrix Σ, $\bar{D}$ is the dimension of the features. Σ and μ are the Gaussian parameters, which are used in Equation 2.4. The window moves forward point by point, and at each step the KL2 distance (2.11) is calculated for the window. If the $KL2$ distance achieves the local maximum at a point, this point will be labelled as a change point (Siegler et al., 1997). Sometimes there are too many peaks in a window. To accelerate the searching, the peaks can be passed through some smoothing rules, and only those larger than a threshold will be treated as the change points (Zhu et al., 1998). KL divergence measures only the expectation of the log-difference between two distributions. The relative entropy of variance and skewness between the two parts can also been approximated and applied as the distance measure to detect the change points (Brutti et al., 2008b). Once all the speaker change points are detected, the feature vectors between two change points will be labelled as a section (the feature vector where the change happens will be included in the section after it).

As an initialization step, the speaker change detection needs to be computed quickly. Thus algorithms with a low computational burden are favoured. The speaker change detection can be evaluated by measuring the number of missed changes in speaker (missing turns) and the number of detected changes that are not true (false alarm). Reducing missing turns is important for the SCD because that speech section containing mixed speech from more than one speaker will contaminate the model trained by this speech section during the clustering step later. Although to reduce missing changes is important, if the resulting sections are too short to cover the main speaker characteristics, later processing will be affected as well.

2.5 Speaker clustering

After the SCD, the purpose of speaker clustering is to cluster the speech sections between speaker change points together according to their speakers. One cluster is produced for each speaker in the audio, and all speech sections from a given person are collected in a single cluster. The speech sections can be clustered in a agglomerative way (bottom-up framework) or using a splitting down scheme (top-down framework). The bottom-up framework will be presented in the next section. To use the new information relating to the speakers after the models are updated, the SCD and clustering steps can be integrated, as described in section 2.5.2. Some post-processing strategies will be introduced in section 2.5.3. The other clustering methods will be introduced in Section 2.5.4.

2.5.1 Bottom-up framework

Within the hierarchical agglomerative clustering framework, all the speech sections are organized in a tree structure, from the leaves to the root. It consists of

the following steps:

1. initialize the leaf speaker clusters of the tree, with each speech section assigned to a cluster;

2. a potential speaker model is trained for each cluster based on the speech sections assigned to it;

3. compute pair-wise distances between each pair of clusters;

4. select the closest pair of clusters, merge their segments to form a new cluster;

5. update the potential speaker model for the new cluster and the distances of the other clusters to it;

6. iterate the last two steps until the stopping criterion is met.

Usually, the results of the SCD step will be taken as the initialization leaves in a bottom-up framework (Tranter and Reynolds, 2006). Zhu et al. (Zhu et al., 1998) and Barras et al. (Barras et al., 2006) considered the cluster initialization problem to be less important and ignored the speaker change detection step by simply splitting them into small same-length speech sections. The number of initial clusters is set beforehand as a value that is much larger than the real speaker number (Barras et al., 2006), or is determined automatically depending on the length of the audio (Anguera et al., 2006a).

Moh et al. (Moh et al., 2003) and Barras et al. (Barras et al., 2004) represented the speaker using a full covariance Gaussian. Gauvain et al. (Gauvain et al., 1998), Meignier et al. Sinha (Meignier et al., 2005) and Moraru et al. (Moraru et al., 2003) used the GMMs because they model the speaker characteristics better. Tranter et al. (Tranter et al., 2004) adopted a single Gaussian model first when the speech sections were short, then used GMMs when the speech

sections became larger. Sinha et al. (Sinha et al., 2005) and Barras et al. (Barras et al., 2004) adopted diagonal GMMs to model the short speech sections while full covariance GMMs were used to model long speech sections. If the GMMs are used to model the speakers, the parameter complexity will affect the performance, as referred to in Section 2.2.3. Eight component GMMs were used in (Gauvain et al., 1998) to model a speech section and this number is unchanged in the whole process. In (Wooters et al., 2004) and (Ajmera and Wooters, 2003), when two clusters were merged, the model complexity of the new model was the sum of the two original models. This helps to remove the need for tuning the penalty weight ϕ in Equation 2.9. In (Anguera et al., 2006a), the complexity of the model was decided dependent on the speech section size and Cluster Complexity Ratio (CCR). In (Anguera et al., 2007), the model complexity is fixed, and the GMM is trained by cross-validation to improve model accuracy. The frame-level purification algorithm was presented in (Anguera et al., 2006c) to remove the components that are dominated by non-speech frames.

The distance metric used in step 3 can be KL2, GLR, the ΔBIC and normalized CLR (Lee et al., 2007) (Chen and Gopalakrishnam, 1998) (Zhou and Hansen, 2000). Vijayasenan et al. (Vijayasenan et al., 2007) proposed that the Jensen-Shannon divergence (Schutze and Manning, 1999) be adopted as the similarity measure between two segments, This depends on the the loss of mutual information caused by merging.

If the clustering process terminates, the remaining number of potential speakers in the tree determines the number of speakers. If the number of speakers in the speech is estimated in advance as K, the clustering tree will be pruned to obtain the K tightest clusters (Tranter and Reynolds, 2006). Some researchers terminate the clustering procedure if the distance measure is over a given threshold. Gauvain et al. (Gauvain et al., 1998) and Barras et al. (Barras et al., 2004) used

the likelihood of a model penalized by the weighted sum of the speech section number detected in the SCD and cluster number to judge if the cluster process should be continued. The BIC stopping criterion was provided by Moraru et al. (Moraru et al., 2003) and this has become the predominant approach. In this approach for two clusters waiting to be merged, their local ΔBIC value will be computed by Equation 2.9. If $\Delta BIC < 0$, they will be merged and the process continues, otherwise the clustering algorithm terminates. Han and Narayanan (Han and Narayanan, 2007) (Han and Narayanan, 2008) applied normalized log GLR as the stopping criterion. Vijayasenan et al. (Vijayasenan et al., 2008) adopted Minimum Description Length (Rissanen, 1989) and normalized mutual information to select the appropriate number of speakers.

2.5.2 Integrated speaker segmentation and clustering

To run the speaker segmentation and clustering separately lacks flexibility because once the SCD step has finished, there is no chance to correct the errors occurring in that step. Therefore, some work was undertaken on the speaker segmentation and clustering steps, with the results of the SCD only used as an initialization for the processing that follows.

The integration framework for iteratively combining speaker segmentation and clustering was first established in 1997 for LIMSI 1997 Hub-4E transcription system (Gauvain et al., 1998). It inserts a segmentation step each time two potential speakers are merged and a new speaker model is then trained. The segmentation is processed by both the maximum likelihood classifier (Gauvain et al., 1998) (Meignier et al., 2005) and the HMM (Ajmera et al., 2002) (Ajmera and Wooters, 2003) (Barras et al., 2004) (Barras et al., 2006). In this segmentation step, first all speech is clustered based on the speaker model for each potential speaker; and second, every potential speaker model is updated according to the

speech sections that are clustered together. This two-step process repeats until the speech sections assigned to all clusters stop changing. The advantages of this integrated speaker segmentation and clustering step is that the boundaries of speech sections that lie between two speaker change points are refined during every clustering round. However, the whole process is computationally expensive.

Another scheme which integrates the SCD and clustering steps, Evolutive HMM (EHMM), was first described by Meignier et al. (Meignier et al., 2000), and then developed in (Meignier et al., 2001) and (Meignier et al., 2005). At the start, an HMM with only one state is initialized and a potential speaker model λ_0 that is trained on the whole audio stream is used as the state's model. It represents all the speakers in the audio. Then several speech sections that have the least likelihood given the model λ_0 are selected to train a new model for a new potential speaker. This new model is added to the HMM as a new state and then all the feature vectors are re-assigned to these two models. All the existing potential speaker models in the HMM are adapted according to the current segmentation. The segmentation and updating process is repeated until the results stop changing. New potential speaker models are added one by one until the likelihood of the current solution is no more than the likelihood of the previous solution. Fredouille and Evans (Fredouille and Evans, 2008) introduced a confidence value to remove the influence of non-speech and overlapped speech portions in the EHMM system.

In (Anguera et al., 2006b), these two kinds of integrated SCD and clustering algorithm were combined. The speech sections were clustered into K_{ini} initial clusters by a method similar to the EHMM, and then they were agglomeratively clustered by the method introduced in (Barras et al., 2006).

2.5.3 Post processing

If the training data is not adequate, the speaker model may not cover the whole feature space. Therefore, complex speaker modelling approaches will fail to discriminate different speakers. After several iterations, the amount of data per cluster increases. Thus the state-of-the-art speaker recognition methods can be employed to improve the quality of the speaker clustering. A Universal Background Model (UBM) is a general speaker model, which is trained by plenty of data to cover all the speaker characteristics under arbitrary situations. The speaker model for a specific person can be created by adapting from the UBM. The adapted model is thought to represent speaker characteristics better, particularly when the training data for the specific speaker is insufficient. Maximum A Posteriori (MAP) estimation (mean-only) is applied to UBM adaptation. Under the Bayesian framework, a variable's posterior probability given a model is the normalized product of model's prior probability and the variable's likelihood given the model. As its name suggests, MAP estimation of the model parameters will select the value that increases the feature vectors posterior probability. Using the UBM as the prior model, the mean vector of the GMM can be obtained by Equation 2.13:

$$\tilde{\mu_i} = \frac{\rho\mu_i^{ubm} + \sum_{j=1}^{N} \tau_{ji} x_j}{\rho + \sum_{j=1}^{N} \tau_{ji}} \tag{2.13}$$

where μ_i^{ubm} is the mean vector of the component i in the UBM, and $\tilde{\mu_i}$ is the corresponding mean vector of the speaker model. $x_1, ..., x_N$ are the feature vectors and τ_{ji} is the posterior probability UBM component i given x_j. ρ is the fixed relevance factor which controls the balance between the speaker data and the prior (UBM) mean. Using UBM-MAP adaptation technology to create a speaker model has been shown to improve speaker recognition performance by

(Reynolds et al., 2000).

(Barras et al., 2004) provided a post-processing step for their speaker diarization system. After several iterations of the clustering process, when the amount of data per cluster increases, the UBM-MAP technique was applied to re-build the model of each cluster. Then the agglomerative clustering process was resumed with the Cross Log-likelihood Ratio (CLR) as the distance measure. The CLR of two segments X_1 and X_2 are calculated by Equation 2.14:

$$CLR(X_1, X_2) = \frac{1}{n_1} \log \frac{p(X_1|\lambda_2)}{p(X_1|\lambda_{ubm})} + \frac{1}{n_2} \log \frac{p(X_2|\lambda_1)}{p(X_2|\lambda_{ubm})} \tag{2.14}$$

where λ denotes the model and n is the number of feature vectors. The process was terminated when the CLR value larger was than a threshold, estimated from the development data sets. (Sinha et al., 2005) derived the segment model by applying two kinds of iterative-MAP adaptation. They also discussed the various approaches to build the UBM. It can be built using the test data itself (in an unsupervised fashion), using other training data, or concatenating the two types of data above. Barras (Barras et al., 2006) also applied a post-processing step in their diarization system. The UBM adopted in these systems was a 128 diagonal GMM. Feature warping (Barras and Gauvain, 2003a) (Barras and Gauvain, 2003b) was also applied to eliminate the acoustic differences of speaker models. If gender classification is applied, the post-processing will be operated separately for each gender, by using a gender-specified UBM (Barras et al., 2006) (Sinha et al., 2005).

The mean vectors of all the components contained in the UBM-MAP adapted speaker model are considered to represent well the speaker characteristics (Faltlhauser and Ruske, 2001) (Tsai et al., 2004) (Tsai et al., 2005) (Tsai et al., 2007). Tsai et al. (Tsai et al., 2005) and Tsai et al. (Tsai et al., 2007) adopted the normalized inner product of the concatenated mean vectors as the segment similarity

measure. Tsai et al. (Tsai et al., 2007) applied the segment purity based stopping criterion.

2.5.4 Other algorithms

Ajmera and Wooters (Ajmera and Wooters, 2003) proposed a top-down framework for speaker clustering, using the full covariances of the segments as the similarity measure. Tranter et al. (Tranter et al., 2004) applied the BIC criterion as the stopping criterion for the splitting procedure.

The use of proxy models in (Reynolds and Carrasquillo, 2004) were inspired by the ideas of anchor models and eigenvoices, which is similar to the method used in the speaker indexing system described by Akita and Kawahara (Akita and Kawahara, 2003). In this, a series of speaker models are built to represent different types of speaker. Then each segment is projected into another feature space by computing its likelihood against each proxy model. The dimension of the space is equal to the number of proxy speakers. The normalized likelihood scores are then treated as distance measures and the clustering process is performed.

2.6 Combination strategies

Each speaker diarization system is considered to have its own distinguishing features and advantages. They may be good at dealing with a particular situation or dataset. Therefore, combining methods used in different diarization systems could potentially improve performance over the best single one.

Moraru et al. (Moraru et al., 2003) performed a combination strategy called ‘piped’ in which two different systems used the results from one to initialize the other system. Then the two systems were applied one after the other to give

the results. Liu and Kubala (Liu and Kubala, 2004) adopted a 'plug and play' strategy to combine the steps of different systems. More integrated merging methods are described in (Meignier et al., 2005) and (Moraru et al., 2003), as the 'fusion' strategy. In this the results from two diarization systems are compared and all the segments whose labels conflict found. Then models are trained on them and the clustering step resumed. Tranter (Tranter, 2005) used a 'cluster voting' technique. This also collected those portions of the audio where the relative output labels are not agreed by all the systems, and then the candidate clusters that maximize the Cluster Voting Metric are selected. An external judge, BIC, is used to pick the optimum solution among them. Gupta et al. (Gupta et al., 2007) (Gupta et al., 2008) integrated systems using different feature vectors.

Figure 2.3 displays the main steps adopted in speaker diarization systems, and the main algorithms used for each step. In the Figure, the speech activity detection is referred to as SAD, and the speaker change detection step is labelled as SCD.

2.7 Evaluation Metrics

The main metric that is used for speaker diarization experiments is the Diarization Error Rate (DER) as described and used by NIST in the RT evaluations.

The NIST Rich Transcription diarization evaluations plan provides a Diarization Error Rate (DER) framework to analyse the performance of speaker diarization systems. It consists of missed data, false alarms and speaker errors. The final outputs of the speaker diarization system is a sequence of 'relative' speaker labels, which are referred to as the hypothesis speaker labels in DER. The 'true' speaker labels will be called the reference speaker labels. An optimal one-to-one mapping of the reference and hypothesis speakers need to be performed to

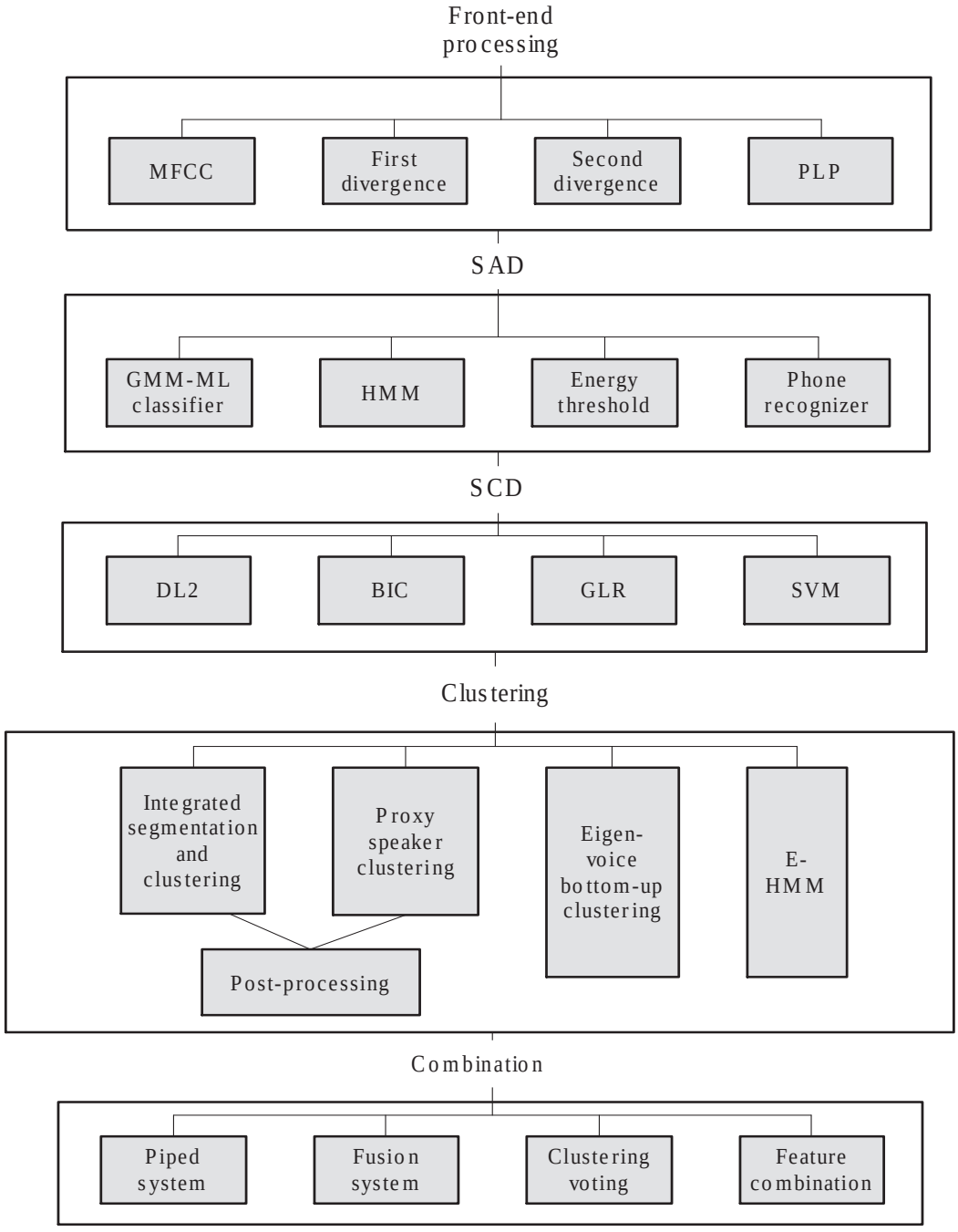


Figure 2.3: The main strategies adopted for diarization

maximize the overlap between their labels. This allows the scoring of different speaker tags between the two files. The Diarization Error Rate score is computed as

$$DER = \frac{\sum_{s=1}^{S} dur(s) \cdot max(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s)}{\sum_{s=1}^{S} dur(s) \cdot N_{ref}} \tag{2.15}$$

where S is the total number of speaker segments where both reference and hypothesis files contain the same speaker/s pair/s. It is obtained by collapsing together the hypothesis and reference speakers turns. The terms $N_{ref}(s)$ and $N_{hyp}(s)$ indicate the number of speaker speaking in segment s, and $N_{correct}(s)$ indicates the number of speakers that speak in segment s and have been correctly matched between reference and hypothesis. Segments labelled as non-speech are considered to contain 0 speakers. When all speakers/non-speech in a segment are correctly matched the error for that segment is 0.

The DER error can be decomposed into the errors coming from the different sources, which are:

- Speaker error: percentage of scored time that a speaker ID is assigned to the wrong speaker. This type of error does not account for speakers in overlap not detected or any error coming from non-speech frames. It can be written as

$$E_{spkr} = \frac{\sum_{s=1}^{S} dur(s) \cdot min(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s)}{\sum_{s=1}^{S} dur(s) \cdot N_{ref}} \tag{2.16}$$

- Missed speech: percentage of scored time that a hypothesized non-speech segment corresponds to a reference speaker segment. It can be expressed

as

$$E_{MISS} = \frac{\sum_{s=1}^{S} dur(s) \cdot (N_{ref}(s) - N_{hyp}(s))}{\sum_{s=1}^{S} dur(s) \cdot N_{ref}} \ \forall N_{ref}(s) - N_{hyp}(s) > 0 \tag{2.17}$$

- False alarm speech: percentage of scored time that a hypothesized speaker is labelled as a non-speech in the reference. It can be formulated as

$$E_{FA} = \frac{\sum_{s=1}^{S} dur(s) \cdot (N_{hyp}(s) - N_{ref}(s))}{\sum_{s=1}^{S} dur(s) \cdot N_{ref}} \ \forall N_{hyp}(s) - N_{ref}(s) > 0 \tag{2.18}$$

The DER is the sum of all these three types of errors.

$$DER = E_{spkr} + E_{MISS} + E_{FA} \tag{2.19}$$

2.8 Baseline system

When developing a new technique it is preferable to do it starting from a baseline system that has been proven to be successful and popular, and has been integrated into a well-rooted theory and state of the art technology. The difficulties met by this baseline system during the implementation will be discussed and a new algorithmic solution will be developed. Finally experiments will be set up to compare the results of the baseline system and the new system to show whether or not it has overcome the shortcomings of the baseline system.

The SAD phase, the SCD phase, the clustering phase and post processing are part of the baseline system. In the SAD phase, a model-based speech detection method is applied to remove the non-speech segments in the audio. Two GMMs are trained for speech and non-speech separately. In the SCD phase, the KL2

divergence is used as the metric to detect the speaker change points.

In the clustering phase, the detected speech sections between speaker change points produced by the SCD step are then used to train the speaker models. The Gaussian model is used to initialise potential speaker models, such that each potential speaker model is trained by a speech section. These potential speaker models are then clustered based on their similarity. ΔBIC (defined in Equation 2.9) is used as the measurement of similarity. The pair of potential speaker models with the lowest ΔBIC values are merged into one, and a new GMM is trained on all the sections assigned to them. In the new GMM, the number of components is the sum of the model complexities of the two GMMs being merged. The merging process terminates when the remaining potential speaker number is below a certain threshold. Then, every speech section detected between speaker change points is re-assigned to the remaining potential speaker model with the highest probability.

In the post-processing phase, a GMM with 128 components is trained by all the speech in the meeting as a UBM. Then mean-only adaptation is used to derive the speaker models of all remaining potential speakers from the UBM. The CLR is used as the similarity measure between the UBM-adapted speaker models, and the pair of potential speaker models with the largest CLR value are merged. The whole process is terminated when the CLR between all the pairs of potential speakers is below a certain threshold. Again, all speech sections between detected speaker change points are re-assigned to the remaining potential speakers. Finally, the non-speech segments detected in the SAD, the speech sections and their corresponding speakers will be output by the system as final results. The baseline system used in this thesis is illustrated in Figure 2.4. In the next chapter, data analysis will be done to help understand the nature of speaker characteristics, in order to derive new techniques to improve system performance.

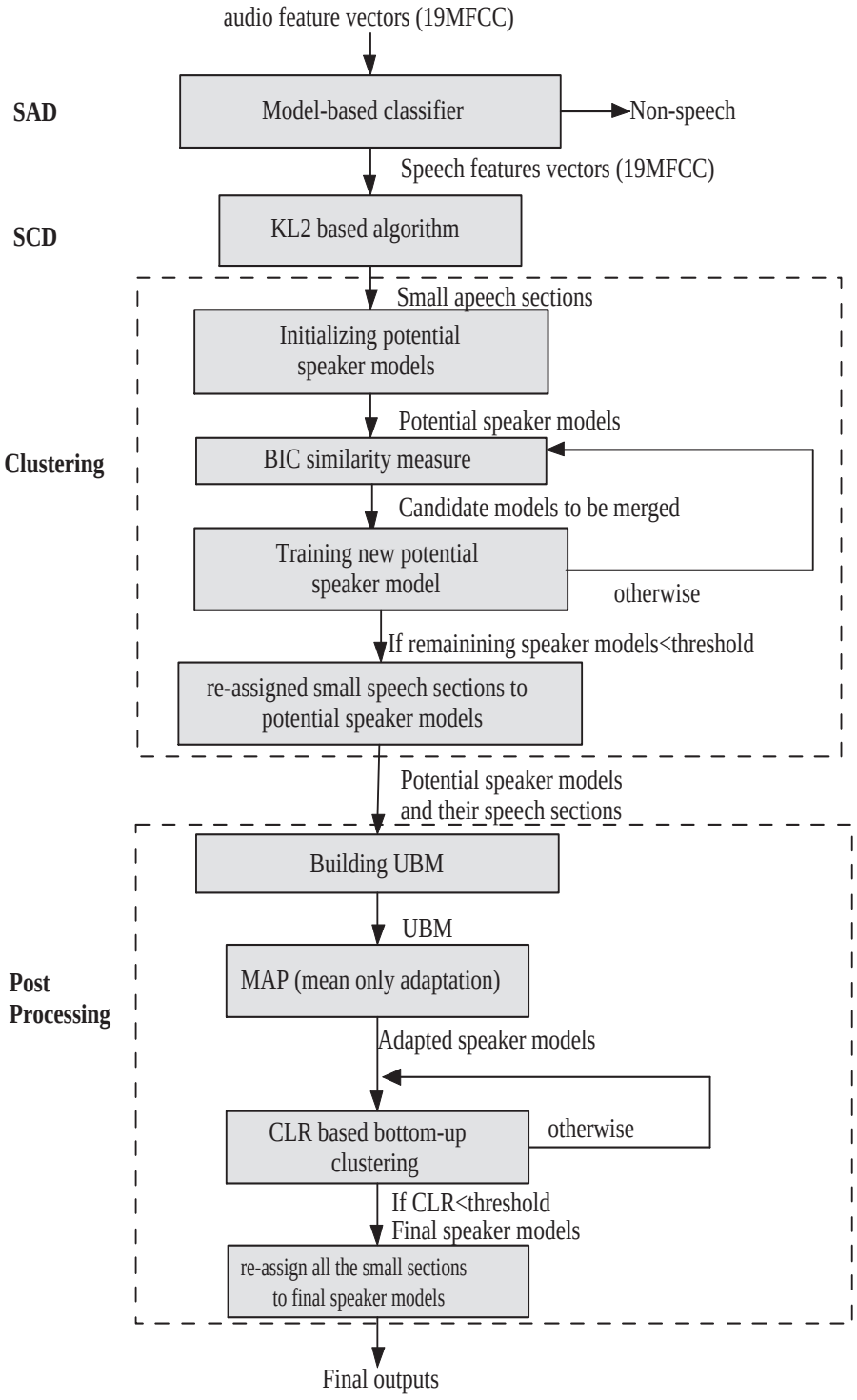


Figure 2.4: Block diagram of the baseline system

Chapter 3

Data Characteristics Analysis

To improve the performance of an existing system, it is necessary to identify those aspects that contribute to or detract from its success and then to exploit those characteristics in new algorithms to overcome the system's shortcomings. In the case of a speaker diarization system, no pre-training material is available for speakers, so the systems adopt an evolutionary strategy in which the speaker models are iteratively adjusted based on the accuracy obtained from the data. Therefore, the performance of the system depends heavily on the characteristics of the data, in this case, face-to-face meeting data. By analysing the specifics of the meeting and identifying their effects on the speaker diarization model, new algorithms can be proposed that improve the modelling accuracy.

In this chapter, we examine the shortcomings of each part of the baseline system in terms of the face-to-face meeting characteristics, from the SAD to the UBM. Section 3.1 explains in detail the specifics of the meeting data that will contribute to the difficulties incurred in speaker diarization. Several measures are developed to quantify the influence of these difficulties. We describe the AMI corpus, which was selected because it meets all the criteria of data selection. In Sections 3.2 to 3.4, a selected sample of AMI corpus data is split into subsets to test whether the meeting characteristics affect the baseline system performance.

In addition, the potential solutions are tested in 3 domains: the SAD, the SCD, and the application of the UBM. Finally, Section 3.5 summarises the conclusions.

New techniques will be developed in Chapters 4 and 5 based on the results from this chapter, and a new speaker diarization system will be proposed that focuses on the specifics of meetings while remaining robust to variations in the meeting characteristics, such as the number of participants.

3.1 Speaker diarization and data selection

The challenges for successful diarization of meetings were presented in Table 1.1. The details can be summarised along six dimensions:

- The number of speakers: the number of speakers in the set of meetings varies from three to ten. The rate of successful diarization decreases with the number of participants, particularly in the algorithm's stopping mechanism.

- Speaker turn length: in contrast with other types of dialogue, exchanges between speakers occur frequently during a meeting. Approximately half of the speaker turns last less than one second.

- Noise conditions: a significant amount of noise obstructs the generalisation of the non-speech training model and degrades the system's performance.

- Room characteristics: the quality of the walls, floor and ceiling, the room size, the arrangement of microphones, the positions of people and the reverberations of the room all affect the quality of the speech.

- Recording microphones: the conversations during meetings might be recorded by lapel microphones, headset microphones or table microphones. Each

type of microphone provides a different level of quality.

- Meeting types: natural meetings are meetings that happen in the real world, while artificial meetings are designed explicitly for research purposes. Artificial meetings can be controlled by a given scenario or a pre-arranged topic.

In addition to the individual challenges listed above, the situation is further complicated by interactions among them. The noise level could be affected by the room characteristics and the recording microphones. The meeting type affects the type of noise and speaker turn length. Artificial meetings may include a certain level of (pre-defined) noise. Informal meetings are more often interrupted by laughter, while intense discussion includes shorter speaker turns. Therefore, to test the influence of certain meeting characteristics on the speaker diarization performance, the main criteria for data selection should include meetings of different types, in different rooms, with different recording microphones and different numbers of speakers. Good reference data also contribute to the analysis of the dataset.

The AMI Meeting Corpus is selected in this thesis, as it meets all the required selection criteria. The AMI corpus is described in detail by (Carletta, 2007) (Hain et al., 2007) and basic information can be found at http://corpus.amiproject.org/. Briefly, the AMI corpus includes 100 hours of meetings, which were recorded in English using three different rooms. The corpus captures both natural conversations and those conducted in pre-designed meetings. Among the natural conversations, the number of speakers varies from three to five. In one type of artificial meeting, four speakers are involved, taking four pre-arranged roles (industrial designer, interface designer, marketing, and project manager). Other artificial meeting types also appear in the AMI corpus, such as a film club scenario.

The meeting rooms were the Edinburgh Room, the IDIAP Room and the

TNO Room, each with its own acoustic properties. Each participant had both a headset microphone and a lapel microphone in the Edinburgh and IDIAP Rooms. In the TNO room, only the headset microphone was provided. A circular microphone array was also provided in each room, either in the centre of the table or on the ceiling. Each meeting was down-sampled from 48 kHz to 16 kHz and recorded in the corpus as WAV files. For each meeting, all channels were provided in separate files unless the recording equipment was broken. The recordings from all microphones were synchronised into a common timeline. The headset and lapel recordings were mixed separately and provided as two single-channel recordings. The AMI Meeting Corpus includes a high quality transcription for each individual speaker, and word-level timings were derived using a speech recogniser in forced alignment mode. A simple energy-based technique was applied to process the speech/silence segmentation for each speaker in the channel derived from the lapel microphone. The meetings recorded by the headset microphone include more breath noise and cross-talking effect, and this part of the noise has not been efficiently labelled by the transcription. Because the new system proposed in this thesis is designed for single channel recordings, only the lapel microphone recorded meetings will be used for the data analysis. Due to the advantages described above, a dataset from the AMI including a variety of rooms and scenarios was selected for data characteristics analysis in this chapter. The meetings recorded in the TNO Room were not included because there was no lapel microphone recording for that room, and all meetings belonged to a single meeting type. In the Edinburgh Room, the meetings can be divided into two types, and in the IDIAP Room there were three meeting types. Three meetings of each type were extracted to form a test dataset. The number of speakers in the test dataset varies from 3 to 4. This dataset will be used for all experiments in this chapter. The meetings with 5 speakers and similar scenarios

can only be treated as a special case because of the limited meeting data (data are available for only two meetings).

Several measures were developed to record certain meeting characteristics, such as the number of turns, the percentage of short turns and the number of speakers. The following measures were applied to characterise the meetings:

- Average Speech to Noise Ratio (ASNR): this measure describes the ratio of the average speech energy to the noise energy. The higher the value, the easier it is to separate the speech from the noise:
  $ASNR = \frac{10*log(average\ speech\ power) - 10*log(average\ noise\ power)}{10*log(average\ noise\ power)}$, where the energy power is equal to the average square sum of the corresponding signals.
- Noise Length Ratio (NLR): this measure describes the length percentage of noise in the entire audio sample
- Speaker number: the number of speakers in the sample.
- Meeting room: the selected meetings occurred in two rooms: 'E' (Edinburgh Room) and 'I' (IDIAP room).
- Meeting type: N - natural meetings, S - artificial meetings under industrial scenarios, B - artificial meetings under other scenarios, such as club meetings.
- Average Turn Length: the average length of the speaker turns.

Table 3.1 lists the meeting room, meeting type, speaker number and the ASNR of all meetings in the test dataset. More measurements of these meetings will be given in the rest of this chapter.

3.2 Problems arising in Speech Activity Detection

As reviewed in the last chapter, speaker diarization systems usually begin with a speech detection step. MISS and FA are two types of errors that occur during

Meeting name	Meeting type	Meeting room	Speaker Number	ASNR
EN2002a	N	E	4	0.2247
EN2006a	N	E	3	0.2303
EN2009c	N	E	3	0.1562
ES2003a	S	E	4	0.2298
ES2009a	S	E	4	0.2321
ES2016c	S	E	4	0.1162
IB4001	B	I	4	0.4469
IB4002	B	I	4	0.4325
IB4005	B	I	3	0.2613
IN1001	N	I	3	0.1950
IN1002	N	I	4	0.1798
IN1005	N	I	4	0.2354
IS1001b	S	I	4	0.2507
IS1006a	S	I	4	-0.0411
IS1009a	S	I	4	0.1733

Table 3.1: Characteristics of the meetings used in experiments.

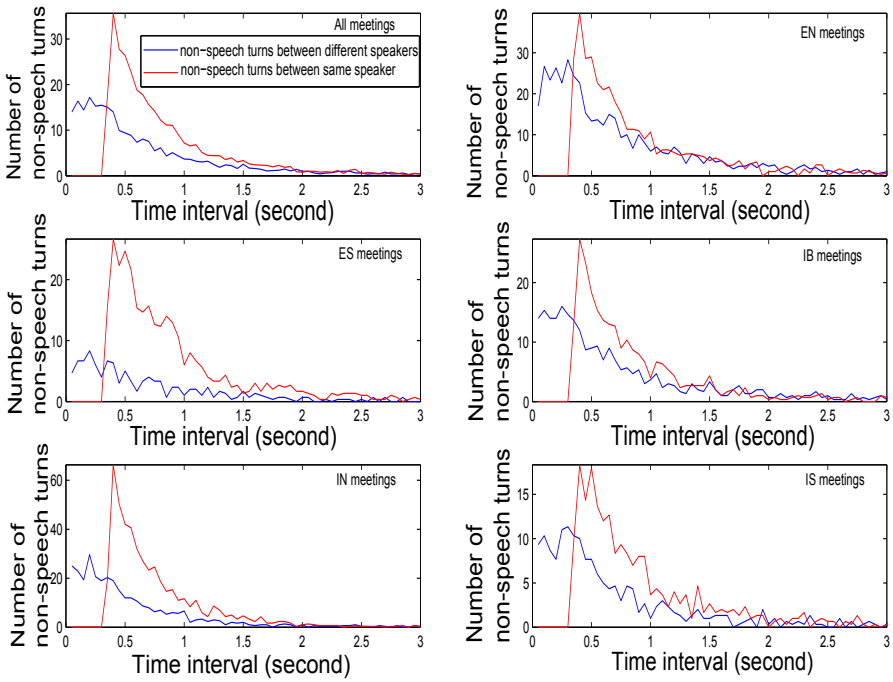
SAD. MISS measures the proportion of the length of speech that is judged to be non-speech, and FA is the proportion of the length of non-speech that is judged to be speech. In the baseline system, GMM models that have been pre-trained for non-speech and speech are used to cluster the audio into non-speech and speech. These models are trained with a small number of pre-labelled datasets using 12 MFCC and sum of squares of amplitude as acoustic features. During SAD, audio is split into small segments, which are then clustered into speech and non-speech using the pre-trained models. Three parameters need to be computed in this method: the length of the segments and the numbers of components used in speech GMMs and non speech GMMs. This section discusses the importance of the parameters (and their values) and the way in which the training material is selected.

3.2.1 Parameter determination

Non-speech segments appear at the intervals between different speakers and during one person's continuous speech. The non-speech segments between different speakers are usually much shorter than the segments from the same speaker. Although long segments promote clustering accuracy by averaging the influence of outliers, the classifier will be confused if a long segment contains both speech and non-speech features. Therefore, the length of the segments has to be long enough for good performance without frequently including both speech and non-speech. To determine the range of the non-speech segments, the distribution of the non-speech turns with a length less than 3 seconds is illustrated in Figure 3.1.

The majority of non-speech turns between speech of the same speaker have lengths from about 0.4 seconds and peak around 0.5 seconds in all meetings, and similar results are observed in various meeting rooms and meeting types (Figure 3.1, all meetings). As expected, natural meetings (Figure 3.1, EN and IN meetings) have more non-speech turns between different speakers than artificial meetings (Figure 3.1, ES, IB and IS meetings). To ensure the detection of most of the non-speech segments, the segment length should be 0.4 seconds.

More components are required to model the speech acoustic features because speech has a more complicated distribution, while in non-speech GMM, only four components are sufficient. To investigate how the number of GMM components affects the performance of SAD, Experiment 3.1 was conducted. In this experiment, a test dataset was divided into two groups, one for training data and the other for test data. The training data were used to train speech and non-speech GMMs with different numbers of components. For the speech GMM, the number of components varied from 2 to 7, while for the non-speech GMM, the Gaussian number varies from 1 to 3. The test data were separated into speech and



The distribution of non-speech turn length between different speakers is shown in blue, and the distribution of non-speech turn length from the same speaker is shown in red. In the first sub-figure, the distribution is averaged over all meetings. In the remaining five sub-figures, the distribution is averaged over meetings of EN, ES, IB, IN, and IS, respectively.

Figure 3.1: Distribution of averaged non-speech turn numbers.

non-speech. Then, each was split into a sequence of segments of equal length and clustered by the GMMs. The recognition rate was expected to vary with the number of components used in GMM, and some noise condition measurements, such as the ASNR or the NLR, might also affect the result. The process of the experiment is illustrated in Figure 3.2. Among the 15 meeting test sets described in Table 3.1, 5 meetings were used for training the speech and non-speech models, and the other ten meetings were used for testing the models. The first 10 minutes of audio were extracted from each meeting for this experiment. The following 10 minutes of speech will be used in the next experiment, where we will test whether the non-speech from different time sections within a meeting has an effect on the non-speech model construction.

The values of the NLR and ASNR measurements of the model training dataset are listed in Table 3.2, and the same measurements of the model testing dataset are listed in Table 3.3. In the meeting name, the number after "-" refers to the section of audio that was extracted. For example, 'EN2009a-1' denotes the first 10 minutes of audio from meeting EN2009a.

Meeting name	ASNR	NLR
EN2009a-1	0.326	24.0%
ES2016c-1	0.284	30.3%
IN1002-1	0.190	13.2%
IS1009b-1	0.213	15.6%
IB4002-1	0.358	37.1%

Table 3.2: Meetings used for non-speech model training and and their noise condition measurements: ASNR and NLR.

Figure 3.3 shows that the MISS error rate decreases when the number of components used in the speech GMM increases. When the model accuracy is improved by including more components in the GMM, less speech is misclassified as non-speech, as shown in Figure 3.3(a). Indeed, more non-speech is classified as speech, especially when the number of components in the non-speech model

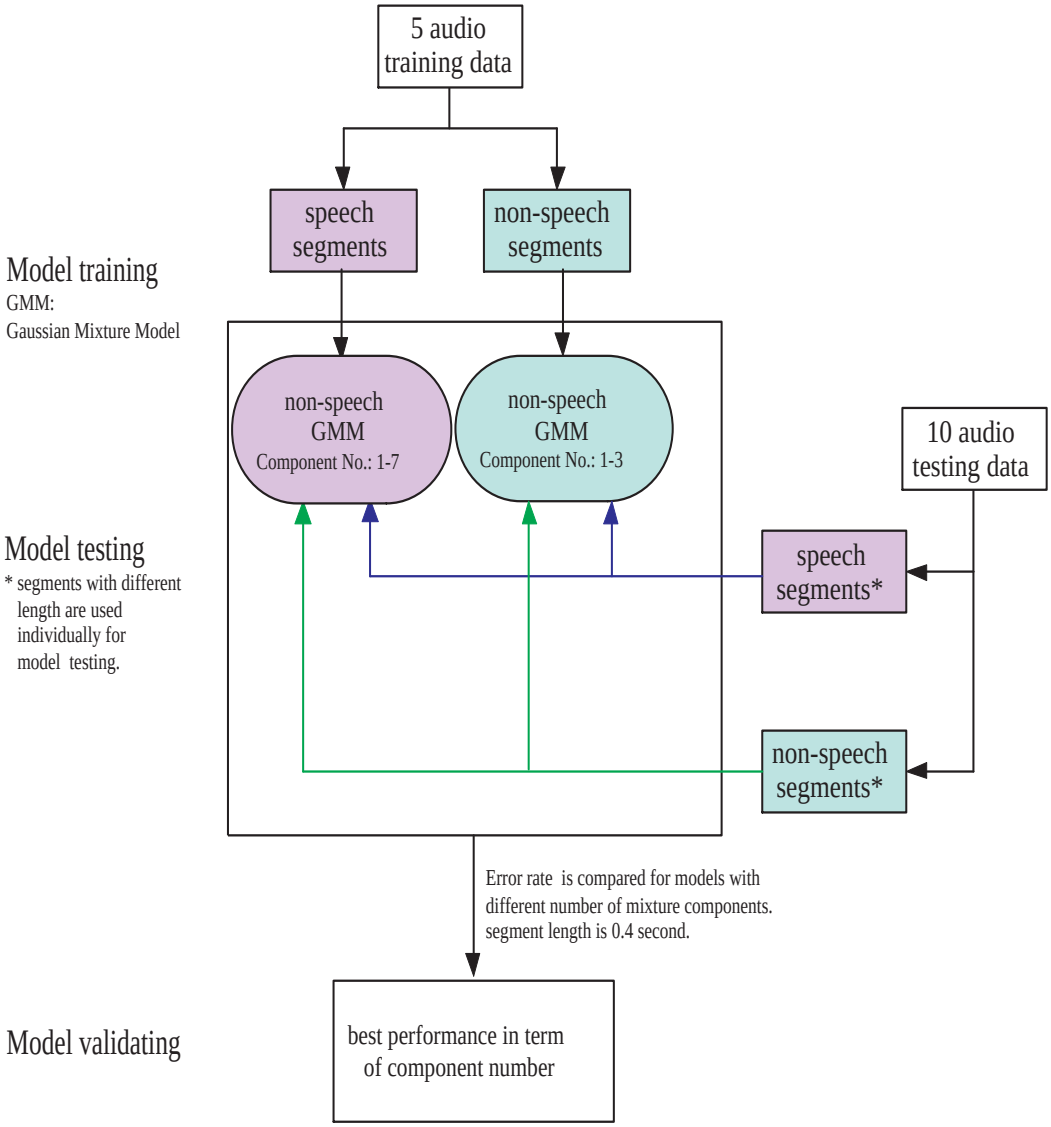


Figure 3.2: Process of Experiment 3.1

Meeting name	ASNR	NLR
EN2002a-1	0.158	22.8%
EN2006a-1	0.164	47.2%
ES2003a-1	0.229	53.1%
ES2009a-1	0.172	26.7%
IB4001-1	0.499	33.4%
IB4002-1	0.394	39.7%
IN1001-1	0.006	35.9%
IN1005-1	0.276	31.9%
IS1001b-1	0.231	43.3%
IS1009a-1	0.089	42.7%

Table 3.3: Meetings used for non-speech model testing and their noise condition measurements: ASNR and NLR.

is low. As shown in Figure 3.3(b), the increase in the number of non-speech GMM components leads to a decrease in FA and an increase in MISS. The total error rate is the sum of these two error measures, and it reaches its minimum value when the reduction in MISS is not cancelled out by the increase in FA. Figure 3.3(c) shows that the minimum total error rate occurs when the speech GMM number is five and the non-speech GMM number is two. No significant error reduction is observed when the speech GMM number increases to seven. Those two values are therefore used as the fixed values of NGMM and SGMM in the next experiment (Figure 3.4). In the experiment, the best results appeared when the Speech GMM number was 7 and the Non-speech GMM number was 1. Figure 3.3 shows how the MISS and FA values change with different numbers of GMM components.

To test whether the number of GMM components (fixed parameter) generated from the set of all meetings is consistent with each individual meeting, we introduced an optimum solution where the best GMM number was calculated as that which give the lowest error rate in each single meeting. The effects of different speaker numbers and the difference between SGMM and NGMM were also analysed. Figure 3.4 consists of four sub-figures (a-d) that show how the

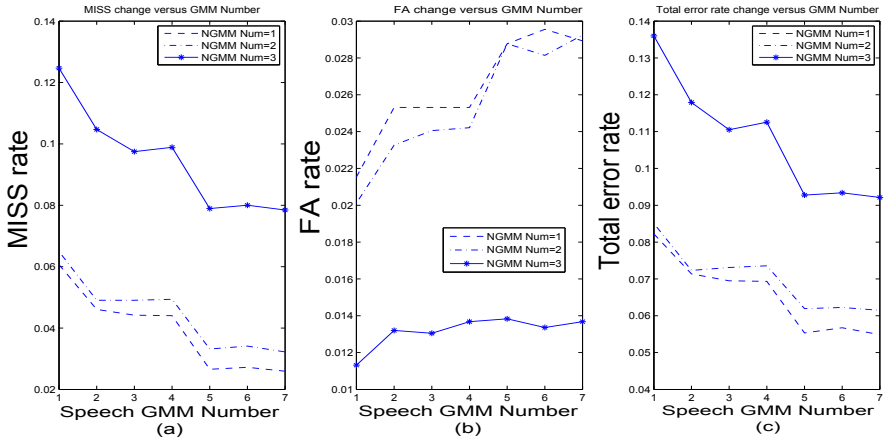


Figure 3.3: MISS, FA and total error rate change with GMM component number when segment length is 0.4 seconds.

total error rate of speech and non-speech clustering changes as a function of the energy measures and the speaker number.

Figure 3.4(a) shows the total error rate as a function of the ASNR. the ASNR is high if the speech is much louder than the noise. The optimum solution line shows the total error rate achieved by the optimum parameter setting for each particular meeting. The total error rate tends to decrease as SNR increases. The error rate line with fixed parameters shows the error rate variation when the speech GMM number and the non-speech GMM number are equal to their optimum values for the whole test set. It seems that the optimum parameters for the whole test set are not always those that produce the best performance in each meeting.

The difference between the fixed parameter and the optimum solution for the error rate is shown as a function of the NLR in Figure 3.4(b). The NLR is the non-speech length ratio of the meeting; its value increases when there is more non-speech in the audio. Figures 3.4 shows that the total error rate obtained using the optimising parameters for the entire test set and those for each individual meeting deviate as the NLR increases. Figure 3.4(c) shows the optimum param-

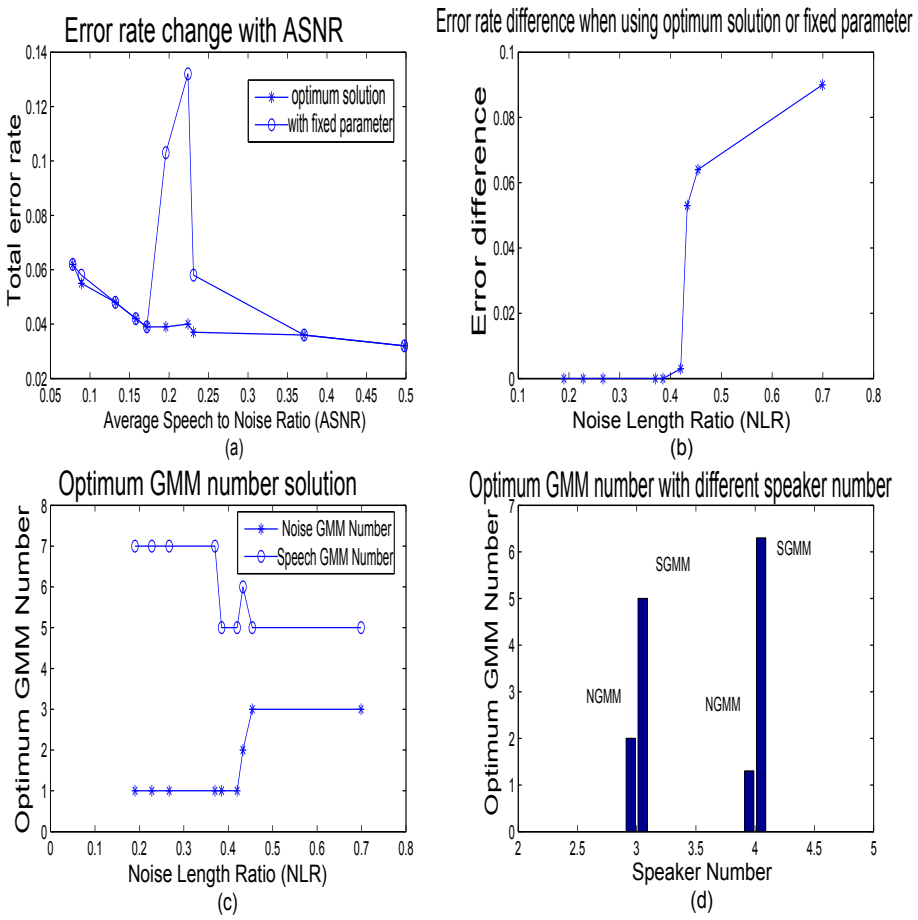


Figure 3.4: Total error rate difference between fixed parameter and optimized solution when segment length is 0.4 seconds. (a) shows the total error rate as a function of the ASNR. (b) shows the difference between the fixed parameter and the optimum solution for the error rate. (c) shows the optimum parameter setting for each meeting as a function of the NLR. (d) shows the total error rate is lower for NGMM when there are four speakers.

eter setting for each meeting as a function of the NLR. It shows that a higher non-speech GMM number and a lower speech GMM number are required when the NLR is high (e.g., 0.4). Two or three GMM components are required for the non-speech model when the non-speech is more than 40% of the audio. The difference between the total error rates is more significant when the noise length ratio is close to 50%, and to obtain the best performance, the non-speech model should include at least three components. In Figure 3.4(d), the total error rate is lower for NGMM when there are four speakers. However, because only two meetings in the test set have three speakers and one of them has a high NL value and a low SNR value, the decrease in performance is more likely caused by the noise length than the number of speakers.

This experiment suggests that the non-speech GMM number is better determined in terms of the NLR. The more non-speech appears in the audio, the more components should be applied in the GMM number. This can be achieved in two steps: using one component non-speech model in SAD to detect the non-speech and then calculating the NLR value depending on the detected non-speech. If the NLR value is higher than a given threshold, more components are used to re-train the model. Then SAD is run based on the new model.

3.2.2 Training material selection

The efficiency of the clustering depends on the similarity between the training and testing materials. It is difficult to train a non-speech model that can cope with all types of noise present in the meetings. In this sub-section, Experiment 3.2 is designed to analyse how the similarity between the training materials and testing materials affects the detection of speech activity. In Experiment 3.2, three different sources are used as training materials.

First, each testing audio is used to train speech and non-speech models for

itself (denoted as Self in this experiment). The testing audio samples are the same as those used in Experiment 3.1. The meeting names, Average Speech to Noise Ratio and Noise Length Ratio of those testing data are given in Table 3.3. Second, 11-20 minutes of speech from the same meeting as each testing audio are used to train the models (denoted as Semi-self). Third, training materials from different meetings are used (denoted as Different). Those training data from the different meetings are the same as the training materials used in Experiment 3.1, and their noise characteristics are given in Table 3.2. Self-training is expected to give the best performance.

The same process is used in this experiment as in Experiment 3.1, except that different training materials are used. The segment length is fixed at 0.4 seconds, and the component numbers used in the speech and non-speech models are the optimum solutions for each test sample according to the results of Experiment 3.1. The setup of Experiment 3.2 is illustrated in Figure 3.5, and the results are shown in Figure 3.6.

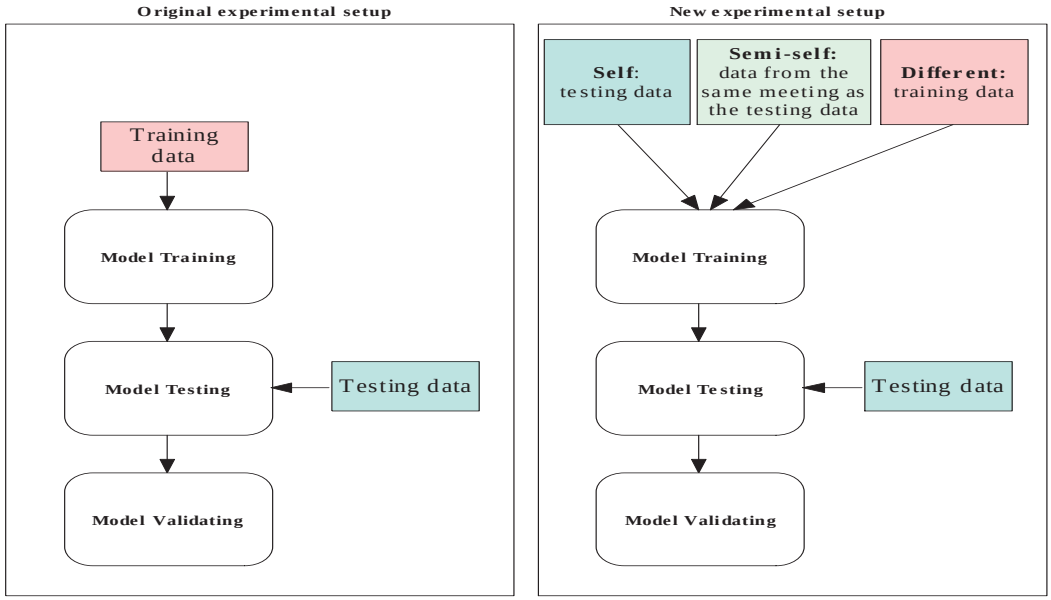
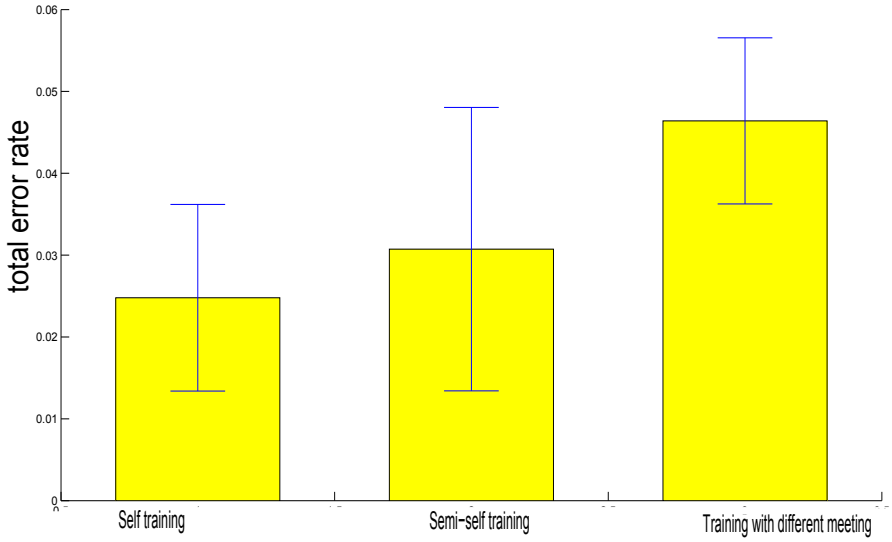


Figure 3.5: Experimental set up for different training materials

As expected, Figure 3.6 shows that the total error rate from either self-training or semi-self-training is different from the error rate from different-training. Interestingly, there is no significant difference between self-training and semi-self-training. Therefore, when constructing the training models, if the speech and non-speech information detected from test meetings can be included in the models, the speech activity detection performance will be improved.



Bar: mean of error rate; Error bars: standard deviation.

Figure 3.6: Comparison of using different training material in speech activity detection.

3.3 Measure of overlap between short speaker segments

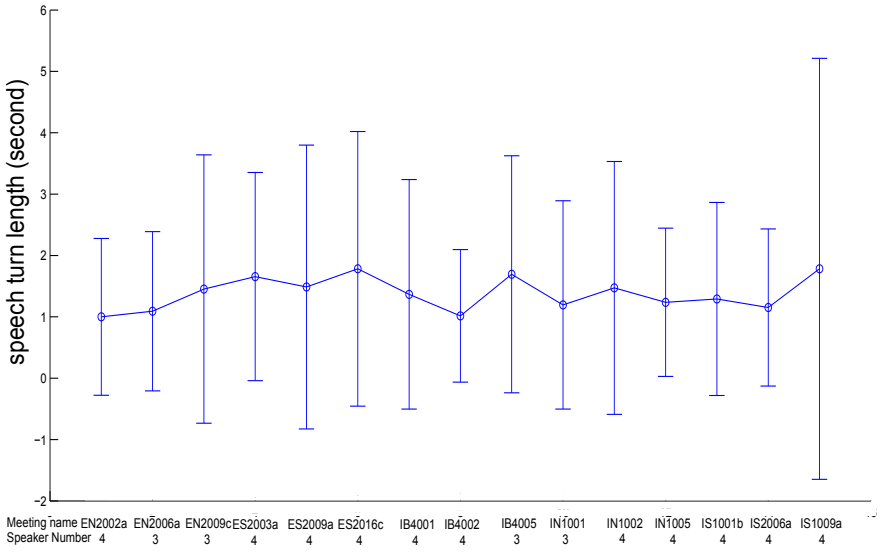
A metric based speaker change detector, which compares the similarity of the speech before and after each point in the meeting to identify change points, is widely applied in diarization of meetings (Miro, 2006). If the similarity is above

a given threshold, the point in question is identified as a change point. Two issues with metric based speaker change detectors are (1) the time constraint on the segment length and (2) the selection of the threshold. Long segments are preferable for speaker characteristics detection. However, the change point is difficult to recognise if multiple speakers appear in the two segments being compared.

Figure 3.7 shows the mean and standard deviation of the speaker turn length for the 15 meetings described in Section 3.2. As illustrated in Figure 3.7, the average speaker turn length for all meetings is in the range 1-2 seconds. The meeting room and meeting type show no influence on the speaker turn length distribution. The meeting with the largest standard deviation is meeting "IS2009a", showing that some long speaker turns appear in this meeting. Because most speaker turns are less than 3 seconds, the distribution of speaker turn lengths less than 3 seconds in these 15 meetings is displayed in Figure 3.8. From Figure 3.7, we see that there is no significant difference in the distribution of speaker turn lengths among different types of meetings recorded in different rooms. The majority of non-overlapping speech turns are under 1 second, and this is not affected by the meeting room or the meeting type. Hence, to ensure the detection of short speaker turns, the segment length should be set at 0.5 seconds.

Next, the similarity of 0.5 second short segments from different speakers and from the same speaker is analysed. Fisher's linear discriminant is a widely used technique in statistics, pattern recognition and machine learning. It can be applied for data classification, dimensionality reduction and feature characteristics description.

Assume X_1 are data from class 1 of size n_1 and X_2 are from class 2 of size n_2. The Fisher linear discriminant seeks to find an optimum hyperplane $< \psi^*, x > +b = 0$ (the notation $< \psi^*, x >$ represents the inner product of ψ^*



circle: mean; error bar: standard deviation

Figure 3.7: Comparison of speech turn length in different meetings

and x) that maximises the ratio of the inter-class distance and intra-class distance of the projections of X_1 and X_2. This Fisher Linear Discriminant Ratio (FDR) is denoted as $J_F(\psi^*)$, and the hyperplane that maximises it is given by Equation 3.1:

$$\psi^* = \arg\max_{\psi}(J_f(\psi)) = \arg\max_{\psi}(\frac{\psi^T((\mu_1 - \mu_2) * (\mu_1 - \mu_2)^T)\psi}{\psi^T(\Sigma_1 + \Sigma_2)\psi}), \quad (3.1)$$

where μ_1 and μ_2 are the means of class 1 and class 2, and Σ_1 and Σ_2 are their covariance matrices. Using the Lagrange method, this maximisation problem can be represented as a convex quadratic optimisation problem whose solution is given by Equation 3.2

$$\psi^* = (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2). \quad (3.2)$$

Figure 3.9 shows how the Fisher discriminant projects data of X_1 and X_2

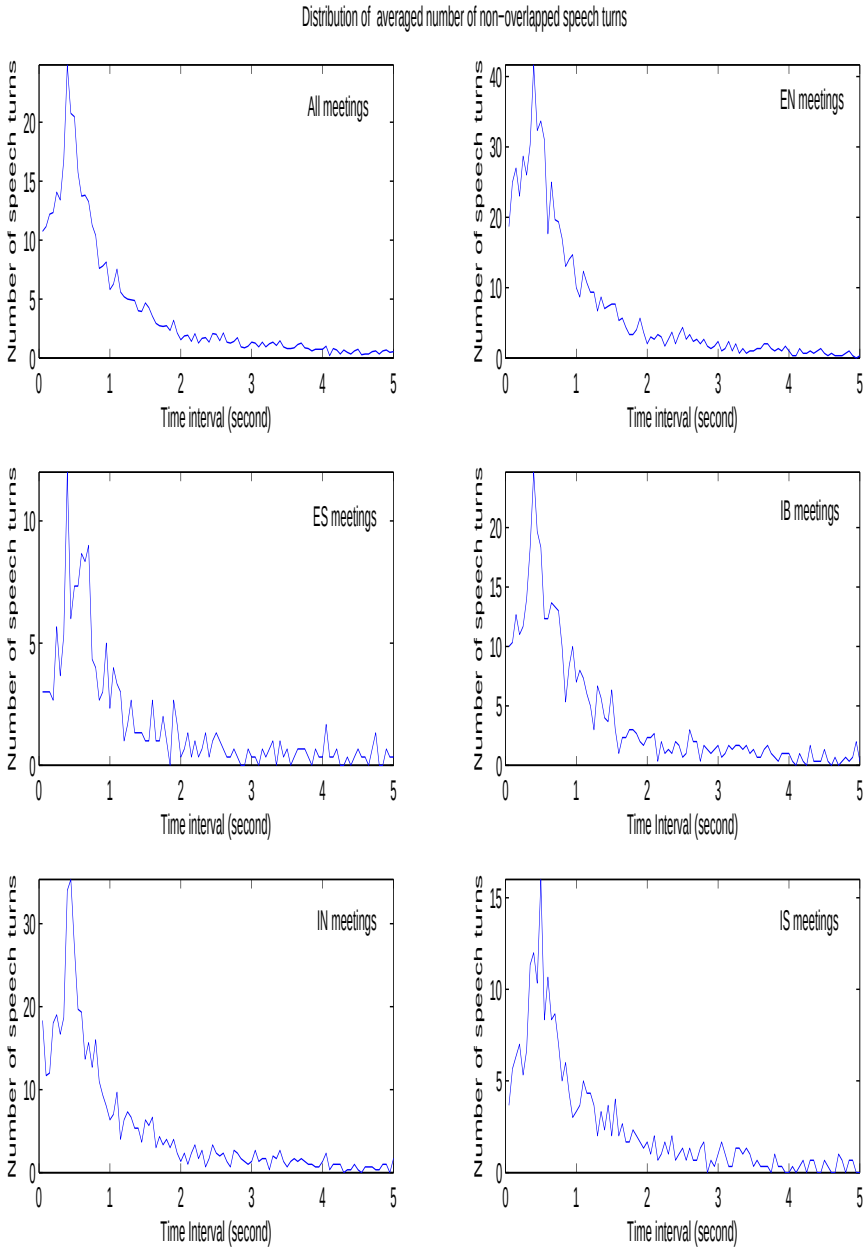
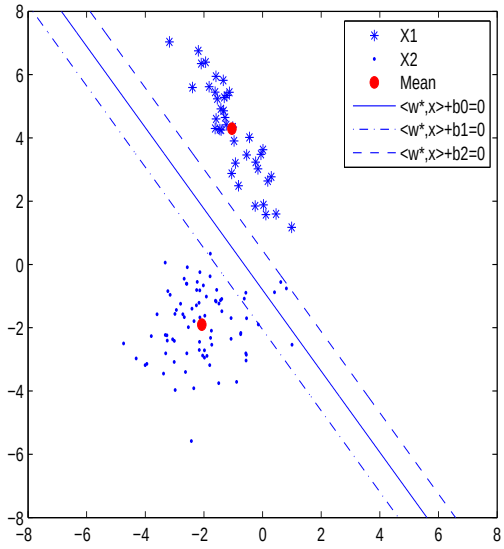


Figure 3.8: Distribution of averaged number of non-overlapped speech turns on 15 meetings.

onto the optimum hyperplane $< \psi^*, x > +b = 0$ It can be seen in Figure 3.9



X_1, X_2 represent the data from the two classes of data separately. $< \psi^*, x > +b_0 = 0$, $< \psi^*, x + b_1 >= 0$ and $< \psi^*, x > +b_2 = 0$ are three separating planes with different value of bias. $b_0 = (\mu_1+\mu_2)/2$, $b_1 = (\mu_1 * n_1+\mu_2 * n_2)/(n_1+n_2)$,and $b_2 = (\mu_1 * n_2 + \mu_2 * n_1)/(n_1 + n_2)$.

Figure 3.9: Fisher discriminant separating plane.

that the hyperplane $< \psi^*, x > +b = 0$ divides the features into two parts. This hyperplane is also called the Fisher linear discriminant classifier (FDC). The distance from a feature x to the hyperplane $< \psi^*, x > +b = 0$ is equal to the absolute value of $(< \psi^*, x > +b) \ / \ \|\psi\|$. If the two classes are separable by the hyperplane, as when $b = b_2$, any feature x from $X1$ will satisfy $< \psi^*, x > +b \geq 0$, while any feature x' from X_2 will satisfy $< \psi^*, x' > +b < 0$. We denote the class label of x as y, where $y = 1$ if x is from X_1; and $y = -1$ if x is from X_2. An error occurs whenever $y * (< \psi^*, x > +b) < 0$. When $b = (\mu_1 + \mu_2)/2$, the hyperplane is equidistant between the mean values of the two classes.

The FDR is the ratio of the inter-class distance and the intra-class distance of the projections of the two datasets onto the FDC. Therefore, it measures the

overlap between the two datasets. The higher the value, the less overlap there is between the two classes. It is assumed that the FDR from different speakers is much higher than from the same speaker because features from different speakers have less overlap.

Using the Fisher Linear Classifier to classify the features from a pair of segments, the classification error rate should be low if the segments are from different speakers because features from different speakers are more likely to stay on different sides of the classification hyperplane. When the two segments to be classified are from the same speaker, the error rate should be higher because the overlap between the segments is larger.

The average distance from errors to the classification hyperplane, another measure derived from the Fisher Linear Discriminant, can also be applied to analyse the data characteristics. If the average distance from the errors to the classification hyperplane is small, the errors appear at the classification boundary (near the classification hyperplane); otherwise, the errors are isolated from the rest of the features in the segment. In another type of classification, if two datasets are from different clusters, the average distance from the errors to the classification boundary should be short because the overlap between different clusters is small. However, the speaker features should be composed of several different mixtures, and some of the mixtures may be far from the others. Therefore, when two segments are from different speakers, the average distance from the errors to the boundary is more likely to be large. When using FDC to classify segments from different speakers, non errors can sometimes be detected; therefore, the distance from the errors to the classification boundary as discussed here is only applicable to the cases in which errors do exist.

The FDR, the classification error rate of the FDC, and the average distance from errors to the FDC can be applied to measure the overlap between segments

from different speakers and from the same speaker. The average distance from errors to the FDC measure can also be used to detect whether there are isolated mixtures in the feature distribution. Although the first two measurements have been widely applied for data characteristic analysis (Ho and Basu, 2002), no such usage has been found for speaker feature analysis.

All 15 meetings described in Section 3.1 will be used in Experiment 3.3. In this section, 19 MFCCs and energy vectors are extracted as acoustic feature vectors from the meetings. In Experiment 3.3, each audio sample in the test set is split into different speakers based on the transcription. The speech from each speaker in an audio segment is then split into small segments, and the distribution overlap between each pair of segments is measured. The experimental setup is illustrated in Figure 3.10. The length of the segments is 0.5 seconds. The overlaps between pairs of segments from different speakers or from the same speaker are shown in Figure 3.11, Figure 3.13, and Figure 3.14.

As expected, the upper panel of Figure 3.11 shows that the FDR values of segments from different speakers are much larger than those of segments from the same speaker. In other words, after projecting onto the FDC hyperplane, segments of different speakers have less overlap. Because the range of FDR of different speakers is much higher than that of the same speaker, a log scale is adopted to make the data more comparable. The minimum FDR of different speakers is approximately 4 (log(4) = 1.3868), and the maximum FDR of different speakers is approximately 500 (log(10) = 6.2146). On the other hand, the range of FDR of the same speaker is between 0 and 1 (log value less than 1). For each meeting, the minimum FDR of different speakers is larger than the maximum FDR of the same speaker. Moreover, the minimum FDR of different speakers is larger than the maximum FDR of the same speaker in all meetings. If a threshold is placed in the gap between the minimum FDR of different speakers

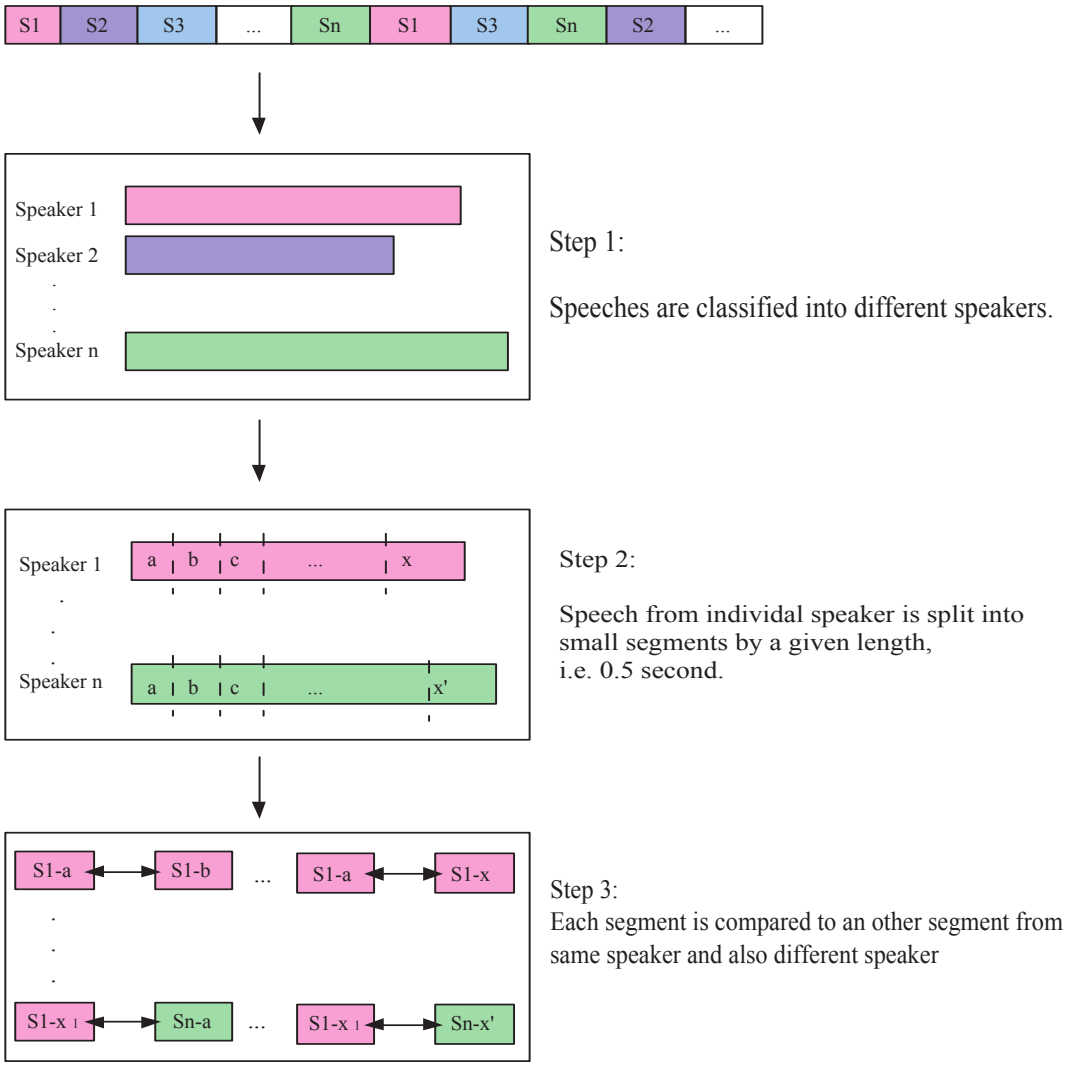
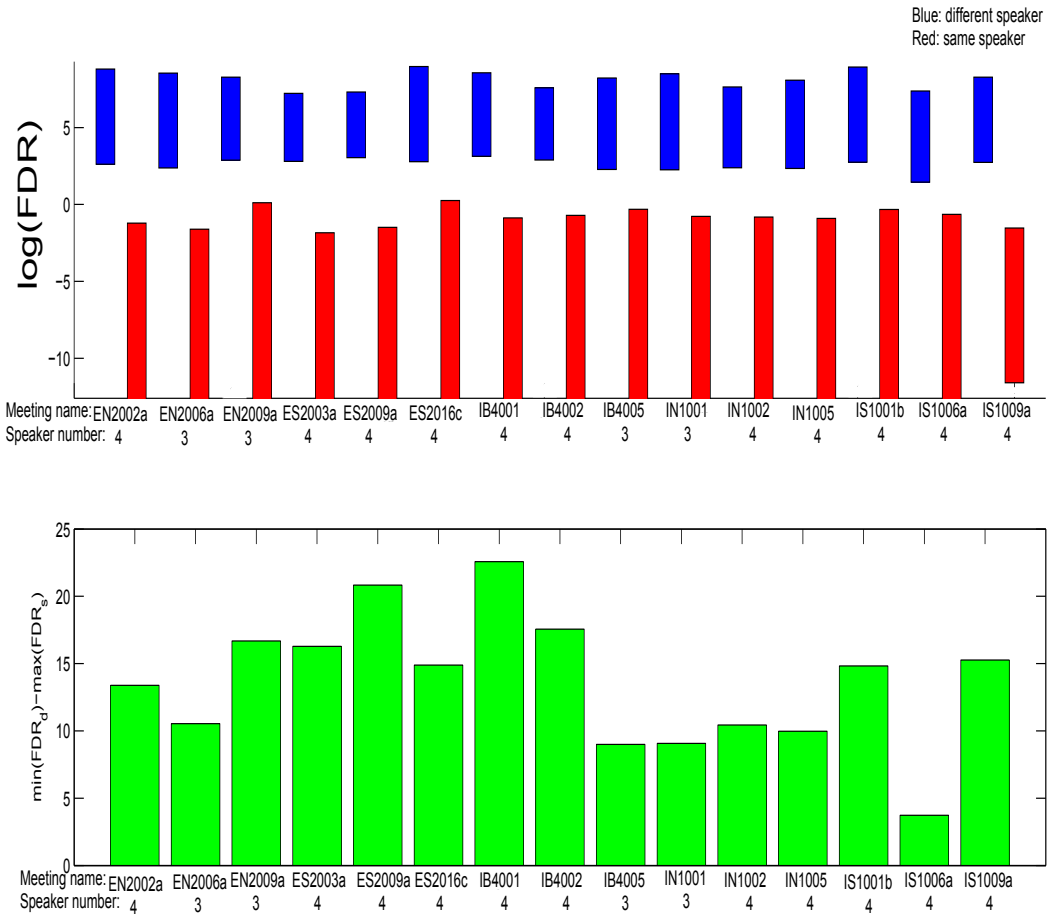


Figure 3.10: Process of Experiment 3.3



In the upper panel, the FDR of segments from different speaker or same speaker is determined in 15 meetings. To make the data comparable, FDR values are shown in log scale. For each meeting, bar presents the value range (minimum and maximum) of FDR, blue colour indicates different speaker, red colour indicates same speaker. In the lower panel, the FDR difference between the minimum value of different speaker ($min(FDR_d)$) and the maximum value of same speaker ($max(FDR_s)$) are displayed.

Figure 3.11: Overlap between short segments from different speaker or same speaker measured by FDR.

and the maximum FDR of the same speaker, all speaker change points will be identified.

The difference between the minimum FDR of different speakers and the maximum FDR of the same speaker is displayed in the lower panel of Figure 3.11. The difference is always positive, which is consistent with the result shown in the upper panel of Figure 3.11. The difference in FDR varies between different meetings. There is no evidence of a correlation between the difference and the meeting room, meeting type or number of speakers. It is clear that the meeting room and type do not affect the speaker characteristics, and there are only two speakers involved at a given speaker change point.

To investigate whether the noise condition of the audio has any effect on the overlap between short segments, the difference between the minimum FDR of different speakers and the maximum FDR of the same speaker is displayed as a function of the ASNR in Figure 3.12. The ASNR value of each meeting is listed in Table 3.1. There is no clear evidence that the SNR will affect the FDR difference between different speakers and the same speaker.

In Figure 3.13, the FDC error rate is applied to describe the overlap between pairs of segments. As expected, the range of FDC error rates of different speakers is higher than the range of FDC error rates from the same speaker, but there is no gap between the minimum FDC error rate of the same speaker and the maximum FDC error rate of different speakers for all meetings. From the lower panel in Figure 3.13, only three meetings have a positive difference between the minimum FDC error rate of the same speaker and the maximum FDC error rate of different speakers. There is no evidence that these differences are correlated with the meeting room, meeting type or number of speakers. These results suggest that FDC could partially identify change points.

In the upper panel of Figure 3.14, the average distance from the FDC errors

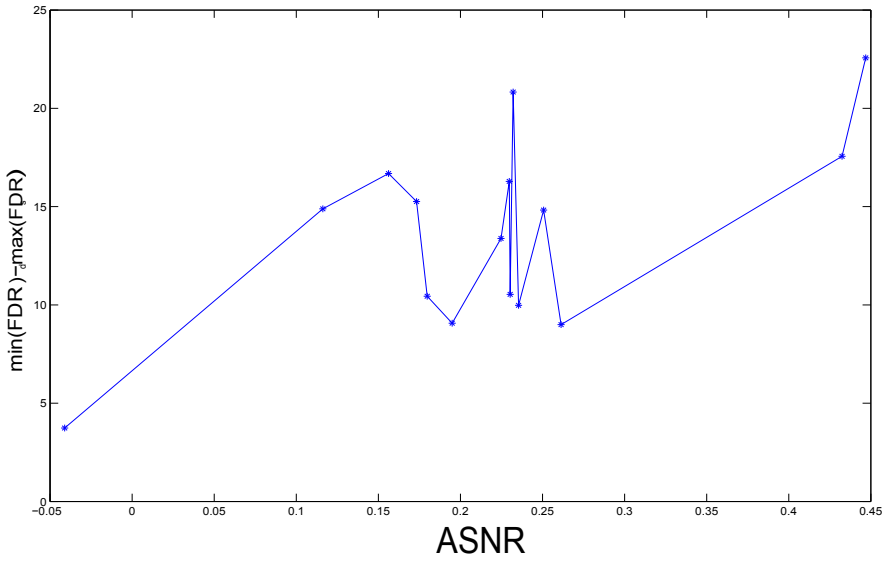
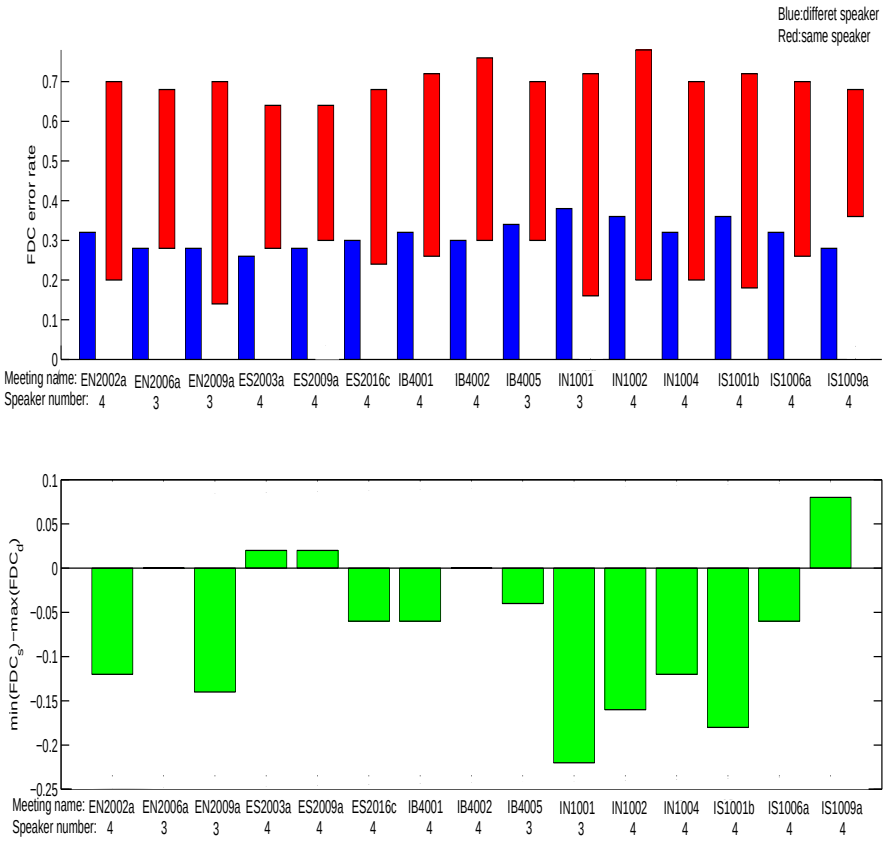


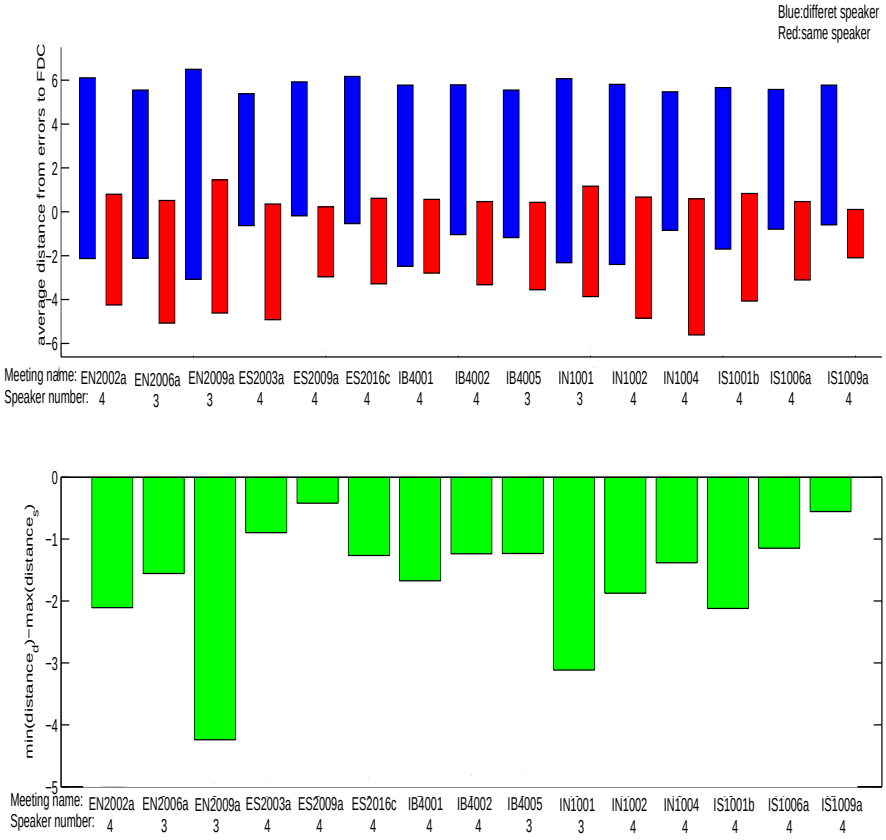
Figure 3.12: The effect of noise condition on the FDR difference between the minimum value of different speaker ($min(FDR_d)$) and the maximum value of same speaker ($max(FDR_s)$).

to the separating hyperplane is applied to analyse the data characteristics. Because the range of the average distances of different speakers is high compared to the average distance of the same speaker, a log scale is adopted to make the data more comparable. As expected, the range (minimum and maximum values of measures) of the average distance from errors to FDC of different speakers is higher than the average distance of the same speaker. However, as shown in the lower panel of Figure 3.14, the minimum average distance of different speakers is always larger than the maximum average distance of the same speaker. Therefore, although features from different speakers are more separable by FDR, some features are also isolated from the other features in the same segment.



In the upper panel, the FDC error rate of segments from different speaker or same speaker is determined in 15 meetings. For each meeting, bar presents the value range (minimum and maximum) of the FDC error rate blue colour indicates different speaker, red colour indicates same speaker. In the lower panel, the difference of FDC error rate between the minimum value of same speaker ($min(FDC_s)$) and the maximum value of different speakers ($max(FDC_d)$) are displayed.

Figure 3.13: Overlap between short segments from different speaker or same speaker measured by FDC error rate.



In the upper panel, distance from FDC errors to the classification hyperplane of segments from different speaker or same speaker is displayed in 15 meetings. To make the data comparable, the values are shown in log scale. For each meeting, bar presents the value range (minimum and maximum) of the average distance, blue colour indicates different speaker, red colour indicates same speaker. In the lower panel, the difference of distance from FDC errors to the classification hyperplane between the minimum value of different speakers ($min(distance_d)$) and the maximum value of different speaker ($max(distance_s)$) are displayed.

Figure 3.14: Overlap between short segments from different speaker or same speaker measured by average distance from errors to FDC classification hyperplane.

3.4 Data distribution in the Universal Background Model

A UBM is used in speaker recognition systems to represent general person-independent feature characteristics of speakers. Because GMM is used almost exclusively for text-independent speaker modelling, it is applied to the UBM to maintain the consistency and comparability of the models (Reynolds et al., 2000). The data used to train the UBM in speaker diarization may come from other sources (other speech corpus) or from the meeting itself (Sinha et al., 2005). The UBM is incorporated into the speaker diarization systems in two ways: 1) to use the UBM as an alternative hypothesis for the speaker model and 2) to derive speaker models by adapting the UBM. In the post-processing step of the speaker diarization system, to determine whether two segments are from the same speaker, the match score of each segment's model and the UBM are measured and compared. The match score of a segment is the likelihood ratio test between a speaker-specific model and an alternative model (in this case, the UBM) (Tranter and Reynolds, 2006). Instead of being trained independently, the speaker models can be derived by an adaptation approach that updates the parameters in the UBM iteratively toward particular speakers. The UBM-adapted speaker model provides a tighter coupling between the speaker's model and the UBM, which leads to better performance at lower computational expense (Reynolds et al., 2000).

The UBM is trained to represent the distribution of the speech features for all speakers in general; therefore, the data selected to train the UBM should be balanced in terms of all variables, such as channel, microphone, and speaker gender (Hasan et al., 2010). Because the task of this thesis is to improve the speaker diarization system performance for meetings recorded by a single type

of microphone, channel and microphone variability will not be discussed here. In addition, the variability concerns about speaker's information are irrelevant for two reasons: 1) data from other corpus and resources will not be applied in the system, and 2) no speaker information is provided for a target meeting. Without information to group the speech into subpopulations to balance their influence on the UBM, the distribution of acoustic features from different speakers and their intertwining will be investigated in Experiment 3.4 to improve the training of the UBM for speaker diarization systems.

The parameters in the process of training the UBM include the number of components in the GMM, the covariance of Gaussian models, and the initialisation method. Either increasing the number of components in the GMM or using a full rank matrix instead of a diagonal matrix as the covariance matrix will increase the model effectiveness. When a diagonal matrix is used as the covariance matrix in the GMM, the loss of accuracy in the model can be compensated for using more Gaussian components. The acoustic features distribution characteristics, which will be analysed in Experiment 3.4, can be applied to determine the parameters of the UBM. For speaker diarization, the initialisation of the UBM can take advantage of the results of the speaker change detection.

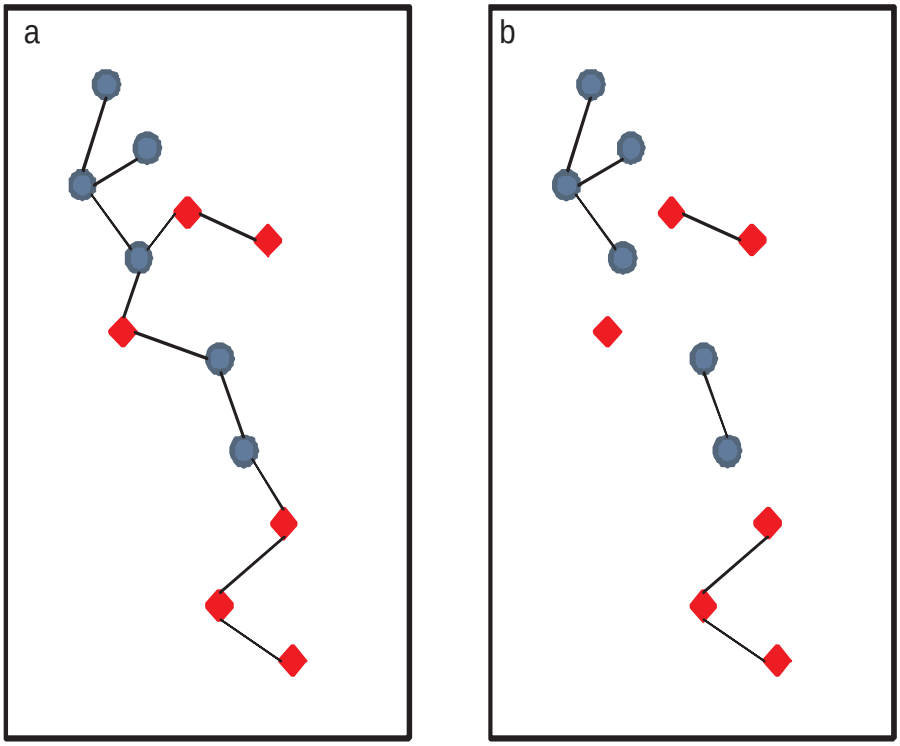
To analyse how the feature space correlates with the inter-speaker variability, Experiment 3.4 clusters the acoustic features according to their speaker and investigates how the clusters are intertwined. First, a Minimum Spanning Tree (MST) is build to connect all of the features extracted from a meeting. In the mathematical field of graph theory, a spanning tree is a subset of edges of a graph that form a tree spanning every vertex. A spanning tree connects all of the vertexes without forming any cycles. An MST is a spanning tree whose sum of edges has minimum total length; it is capable of representing a cluster with irregular boundaries. Refer to the works of Kruskal (1956) and Prim (1957) for the

definition of the problem and its first solution (Kruskal, 1956) (Prim, 1957). In the case of an acoustic feature space, the vertexes are all of the features, and there is an edge between each pair of features. The length of the edge is the Euclidean distance between the features. Therefore, the MST of the meeting connects each feature to its nearest neighbour and forms a tree traversing all of the features.

The algorithm to produce the MST is described by Dijkstra (Dijkstra, 1960). After producing the MST, if the two vertex features of a certain edge are from different speakers, the edge will be removed from the MST. Removing these edges produces a collection of connected components, which are sub-trees of the MST. Finally, the number of the sub-trees remaining in the MST shows how many subsets there are in the feature space, which is isolated by features from different speakers. An example of an MST is shown in 3.15(a), and the sub-trees remaining after removing all the edge connecting points from different clusters are illustrated in Figure 3.15(b).

The variance of the acoustic features comes from two sources, phonetic variance and speaker variance. Phonetic variance is based on different pronunciations of various syllables. Different speakers possess different speech/physiological characteristics, so that an increase in the number of speakers leads to an increase in the variance of the features. Instead of occupying disjoint spaces, features from different speakers are more likely to overlap. The speaker variability is likely to be mingled with phonetic variability, and as a result, they split the feature space into many small regions. The total number of sub-trees that are isolated from the features of the same speaker is expected to be high. Longer speeches include more vocabulary and hence more phonemes. More speakers will further divide the feature space. Thus, the number of isolated sub-trees is expected to increase with both the speech length and the number of speakers.

All 15 meetings described in Section 3.1 will be used in Experiment 3.4.



There are two classes of data set, one is denoted by circle and the other is denoted by diamond. In the sub-figure (a), the MST is built across the two class, shown by line '–'. In the sub-figure (b), the remaining subtrees after removing all the edges connecting data from different clusters are shown.

Figure 3.15: MST illustration.

Nineteen MFCCs and energy vectors are extracted as acoustic feature vectors from the meetings. Some characteristics of the meetings, such as the meeting type, meeting room, number of speakers and average speech to noise ratio, are given in Table 3.1. Other characteristics that may have an effect on the experiment, such as the average speech length and turn length in the meeting, are given in Table 3.4. The number of isolated sub-trees is displayed along with each meeting in Figure 3.16. The meeting type, meeting room and number of speakers are all labelled within the figure. Figure 3.16 shows that the number of isolated sub-trees has a high value for each meeting, ranging from 7000 to 35000. When the meeting type is natural, the number of isolated sub-trees is high, and the room type shows no clear influence on the number of isolated sub-trees. The effect of the number of speakers cannot be observed in Figure 3.16. However, because other meeting characteristics, such as speech length, have not yet been measured, the influence of the number of speakers may be concealed. Next, we show how the number of isolated sub-trees varies with the speech length in Figure 3.17. It

Meeting name	Speech Length (second)	Average Turn Length (second)
EN2002a	1367	1.5324
EN2006a	1586	1.9778
EN2009c	2174	2.8758
ES2003a	5251	3.1065
ES2009a	9663	2.1466
ES2016c	13201	2.4474
IB4001	10200	1.4340
IB4002	9635	1.7275
IB4005	15003	3.7777
IN1001	22993	2.4428
IN1002	18769	2.5710
IN1005	21109	2.2363
IS1001b	13819	2.8729
IS1006a	4667	1.9274
IS1009a	4913	2.3956

Table 3.4: Characteristics of the meeting used in experiments.

can be seen that when the speech length increases, the number of sub-trees tends to increase. However, several points fall outside the trend in the figure. This may be caused by two reasons: 1) an increase in the speech length in a meeting does not always represent an increase in the number of phonemes because the same words/phonemes can be repeated many times in an audio segment, and 2) the increase of the number of isolated sub-trees will be affected by the number of speakers. In Figure 3.17, it can be seen that the increase trend has been disturbed by a reduction of the number of speakers. However, there is no clear evidence of an influence of these two meeting characteristics on the number of isolated sub-trees.

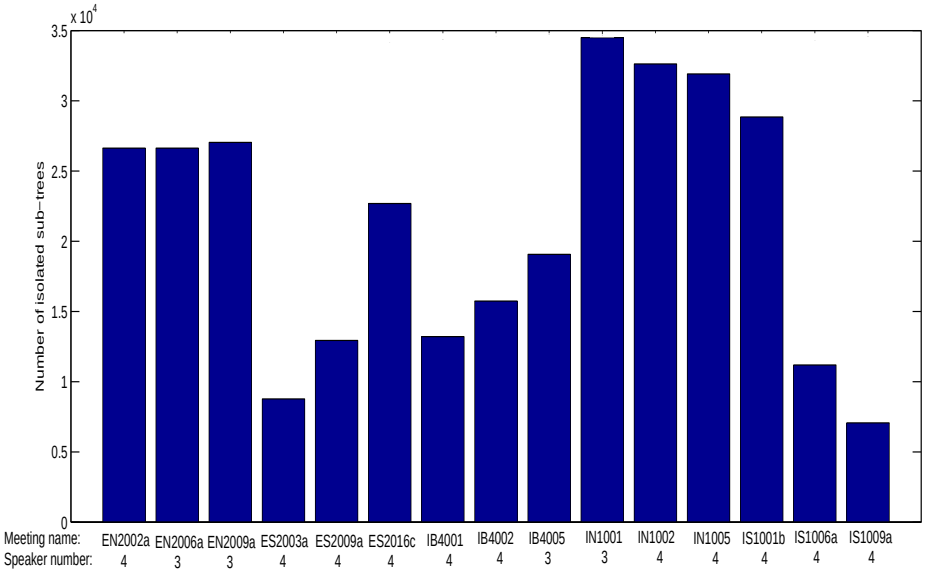


Figure 3.16: Number of isolated sub-trees in each meeting.

In Figure 3.18, how the number of isolated sub-trees changes with the ASNR and average speech length is illustrated. However, no obvious evidence can be observed showing the influence of the two meeting characteristics on the number of isolated sub-trees.

From Experiment 3.4, it can be concluded that the speaker variability is mingled with phonetic variability to divide the feature space into a huge number of

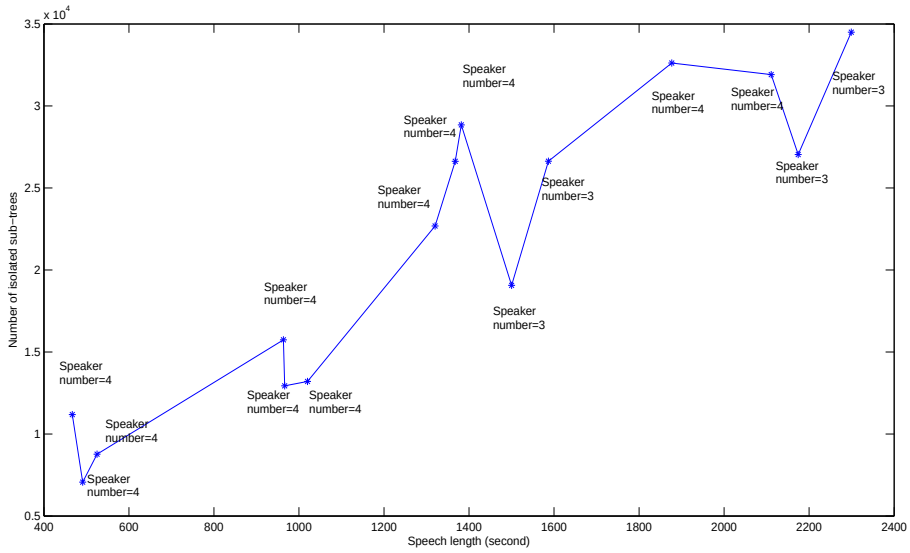


Figure 3.17: How the number of isolated sub-trees changes along with length of speech and number of speaker.

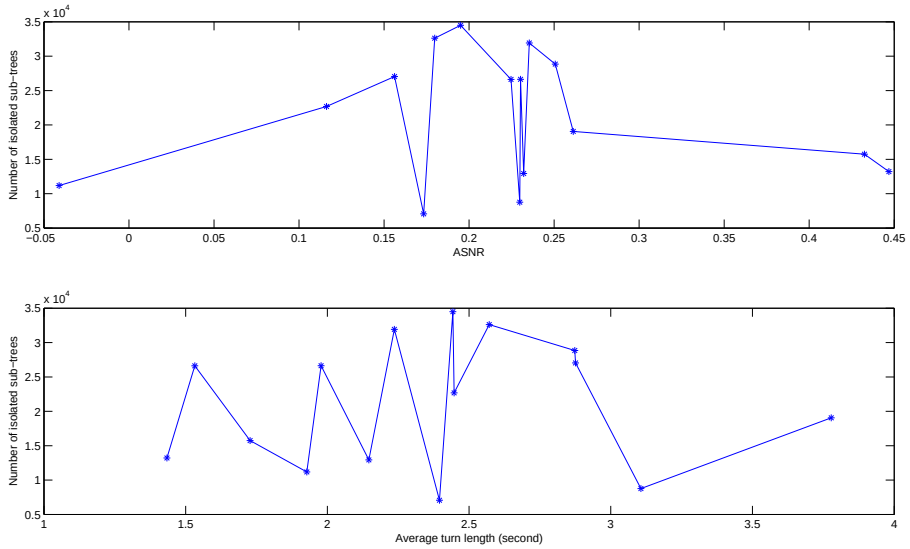


Figure 3.18: How the number of isolated sub-trees changes along with other meeting characteristics.

small sub-spaces according to the speaker. The number of isolated sub-spaces is affected by both the length of the speeches and the number of speakers in a given meeting. Because the UBM needs to capture as much inter-speaker variability as possible, more components must be included in the GMM to represent more sub-spaces.

3.5 Conclusion

In this chapter, the drawbacks of the existing speaker diarization systems had been investigated, the meeting characteristics that may cause these problems were examined, and potential solutions for these drawbacks were deduced. The experiments in this chapter focused on 3 parts of the speaker diarization system: SAD, SCD and the construction of the UBM.

For the SAD process, 5 conclusions can be drawn from Experiment 3.1 and Experiment 3.2. First, if the number of components contained in the GMM for speech or non-speech is increased, the corresponding model accuracy will increase; on the other hand, the model accuracy of its counterpart will decrease. Second, a minimum total error rate is achieved when the speech GMM has 7 components and the non-speech GMM has 1 component, based on the entire development-set. Third, when the NLR value is high, more components should be incorporated for better performance. Fourth, if the audio material used to train the speech / non-speech GMM and the test audio material used to test the performance of the GMMs are from the same meeting, the performance of the SAD process increases significantly. Fifth, 0.4 seconds is a suitable choice for the segment length in SAD.

Taking advantage of the above conclusions, a new algorithm will be to improve the system performance in Section 6.3. The new algorithm first detects

speech and non-speech using the existing SAD algorithm, and then it re-trains the speech and non-speech GMM by adding the new detected information and increasing the number of components in non-speech GMM if the noise length ratio is high.

The aspects of all meeting characteristics that affect the performance of SAD are the ASNR and the NLR. Because the new algorithm can adjust the GMM component number according to the speech and non-speech detected, it will improve the system performance, especially when the NLR value of a target meeting is high. On the other hand, consistent with the experiment results, the error rate of the system will decrease with the ASNR value.

Based on Experiment 3.3, we can derive some conclusions for the SCD process. First, the FDR, the error rate of the FDC and the average distance from errors to the FDC are all capable of determining whether a pair of short segments is from different speakers or the same speaker. Second, some features are far from the rest of the features of the same speaker. Considering the results from Experiment 3.4, this is caused by the phonetic variability in the acoustic feature space. Because there is no gap between the minimum average distance of the same speaker and the maximum average distance of different speakers, it is unclear whether features in a given short segment will traverse several sub-spaces or how many sub-spaces they span. Third, 0.5 seconds is a reasonable choice of segment length in SCD.

The measurements applied in Experiment 3.3 are evaluated based on a short segment length of 0.5 seconds, so the new algorithm based on the experiment should obtain a better performance when there are many short speaker turns of less than 1 second in a target meeting.

Since the measurements applied in Experiment 3.3 are evaluated based on short segment length of 0.5 second, the new algorithm based on the experiment

should obtain a better performance when plenty of short speaker turns exist in a target meeting. Being referred to as short speaker turns, their length should be less than 1 second.

From Experiment 3.4, we can conclude that 1) in the acoustic feature space, the inter-speaker variability is intertwined with the phonetic variability; as a result, features from different speakers split the feature space into many small sub-spaces; 2) the number of sub-spaces tends to increase with the length of the speech in a target meeting; and 3) a reduction in the number of speakers in a meeting will hinder this trend.

A GMM for a particular speaker should contain fewer components to diminish the influence of the intra-speaker variability, which is the phonetic variability within a speaker. On the other hand, the UBM needs to represent as much inter-speaker variation as possible to represent more sub-spaces in the feature space. In Chapter 5, a new algorithm will be derived for both speaker modelling and UBM modelling. The number of components in the GMM will be controlled so that fewer components are allowed in a speaker model, while more components are allowed in the UBM.

An increase in the speaker number or the speech length in a meeting will lead to more sub-spaces that are isolated from the features of the same speaker. Therefore, more components must be included in the UBM. After adopting the new speaker modelling and the UBM modelling algorithm, the system performance will improve, especially when the speech length is long and the speaker number is high.

Chapter 4

Fisher Linear Discriminant Based Speaker Change Detection

In the previous chapter, Fisher Linear Discrimination Analysis (FDA) was used to detect the overlap between short segments. Three different measurements, the FDR, the error rate of the FDC, and the average distance from errors to the FDC, were derived to represent the difference in overlap between segments of different speakers and segments of the same speaker. In this chapter, these measurements will be combined to develop a new algorithm for the SCD task. In Section 4.1, a description of the new algorithm will be given. In Section 4.2, all of the parameters of the new algorithm will be adjusted by the development set. In Section 4.3, the results of the new algorithm will be compared to those of the algorithm used in the baseline system.

4.1 Description of the FDA-based SCD algorithm

In Section 3.3, we saw that the FDR, error rate of the FDC, and average distance from errors to the FDC can be used to determine whether two short segments are from the same speaker, although the latter two measures would produce results

with errors. In this section, these three measurements are combined to obtain an optimum solution that might perform better than any single measurement; therefore, a new SCD algorithm is created. The new SCD algorithm checks for the existence of speaker change in a given meeting at each feature vector. First, for each point, the new measurements are computed based on two short segments of the same length, before and after a selected point. According to the analysis of the previous chapter (Section 3.2.1), the length of the segments is set to 0.5 s. For the features in the first 0.5 s and the last 0.5 s of the meeting, this computation is ignored because no two complete segments can be obtained before or after these points. Subsequently, the peak points of the new measurements are selected, and if the adjacent peaks are close to each other (less than 0.1 second); the peak with the smaller value is removed. Around a real change point, false change points are always detected because when computing the new value for a point near the change point, the segment before or after the selected point contains speech from more than one speaker, which could affect the value of the measurement. Therefore, manually removing peaks that are close to each other will reduce the number of false changes detected by the algorithm. However, the time restriction of removing extra peaks must be shortened to avoid missing frequent speaker changes. Finally, the remaining peaks with values higher than a threshold are confirmed as the speaker change points.

When comparing the changes detected by the algorithm (detected changes) and the real changes, the detected changes are mapped to the real changes in a one-to-one relationship. If the detected changes are within a 0.1 second interval of a real change point, they are mapped to that real change point. The detected points that cannot be mapped to any real change points are called “false changes”, and the real change points that are not the images of any detected points are called “missed changes”. Two types of error rate are adopted to mea-

sure the performance of an SCD algorithm: the missed change rate, which is defined as the ratio of the number of missed changes to the number of real change points, and the false change rate, which is defined as the ratio of the number of false changes to the number of real changes. There are two reasons why the first type of error has a greater influence on the speaker diarization system as a whole. The first reason is that the detected sections between the change points will be clustered according to their speaker in the next step, so there is no chance that the missed change points will be detected later. The second reason is that the speaker models will be trained by these sections and the features from other speakers will decrease the accuracy of the models. Although the second type of error can be corrected later in the system, if two SCD algorithms have similar missed change rates, the one with the lower false change rate is preferred. A lower false change rate means that longer sections are obtained between change points, and therefore, more training material can be used to build the speaker models. The point where an overlap begins or ends will be processed as a real change.

To combine the FDR, the error rate of the FDC, and the average distance from errors to the FDC into a new measurement, a parameter must be introduced to balance their levels of influence on the new measurement. The error rate of the FDC has two characteristics: (1) it has a higher value when two segments are from the same speaker, which is in contrast to the others, and (2) it is always smaller than 1; therefore, the FDC error rate can be used as the denominator in the new measurement. If the FDC error rate is equal to zero, it is assigned a very small value (0.001) to avoid division by zero. Using the parameter α in the numerator to adjust the balance of the other two measurements, the new measurement is given by Formula 4.1:

$$\frac{(FDR + \alpha * (average\ distance\ from\ errors\ to\ FDC))}{FDC\ error\ rate}.$$

In the new algorithm, a threshold is set in the final step to separate the change points from the other peaks. Therefore, it needs to be set to a value between the values of the new measurement between segments of the same speaker and those of different speakers. The scale of the gap varies for the FDR and the average distance from the errors to the FDC.

The value of the threshold should vary according to α, since different values of α adjust the combination and thus change the scale of the gap. In the next section, experiments will be conducted to test different α values and their corresponding thresholds to find the best combination.

4.2 Parameter adjusting

In this section, experiments will be set up to determine the value of α that optimises the performance of the new SCD algorithm and the corresponding threshold that minimises the missing error rate. The meeting data applied in these experiments are the same fifteen meetings that were used throughout the last chapter. To exclude the influence of the non-speech segments in the SCD task, all of the non-speech segments are removed from the meetings according to the transcription. The 19 MFCCs and the energy feature are extracted as the feature vectors in the experiment. The performance of different values of the parameters will be measured by the missed change rate.

Figure 4.1 shows the missed change rate along with different values of α. The missed change rate is averaged over all fifteen meetings and is obtained by choosing the optimum threshold value for each corresponding α. It can be observed from Figure 4.1 that the missed error rate reaches its minimum when

$\alpha = 500$ and $\alpha = 550$. In the figure, the performance of α is shown in the range between 5 and 850. Since the value of the FDR is about ten times more than the average distance from errors to the FDC (measured by their mean and median values), when $\alpha < 10$, the FDR has more influence on the new measurement, and when $\alpha > 100$, the average distance from errors to the FDC has a much greater impact on the new measurement. When $\alpha < 5$, the missed change rate increases rapidly. When $\alpha > 850$, the missed change rate stabilises around 0.0370. Though the average distance from errors is more capable of detecting change points than the FDR, based on this experiment, an appropriate combination of all three FDA-based measurements is more suitable for the SCD.

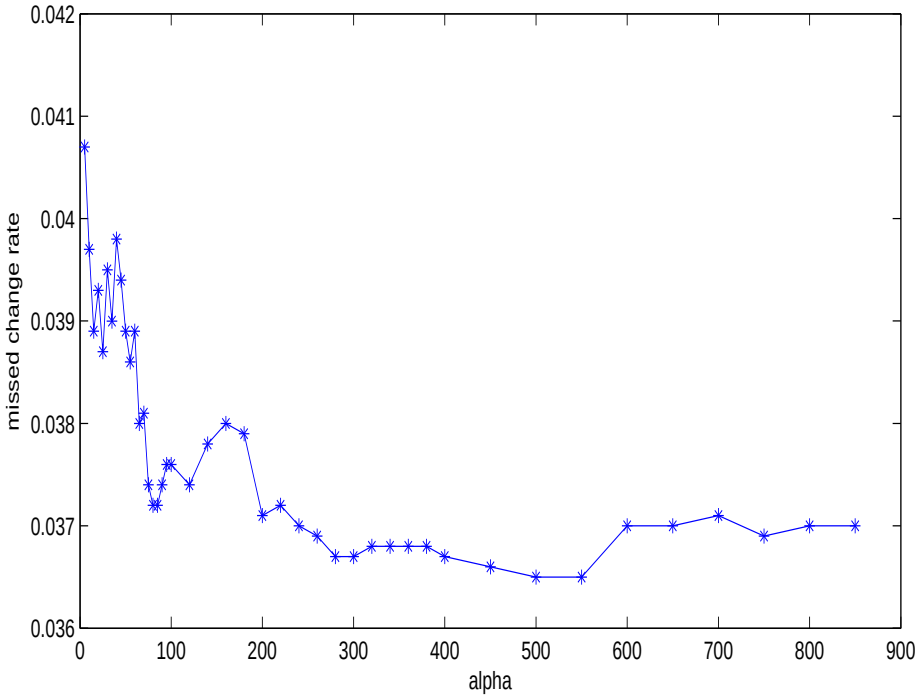


Figure 4.1: The variation of the missed change rage as α increases, using the new measurement.

When the value of α is assigned as 500, the variation of the missed change rate with the threshold value is illustrated in Figure 4.2. From Figure 4.2, it can

be seen that the missed change rate increases when the threshold value increases, and its minimal value is 0.0365. As long as the threshold is between 0 and 150000, the value is unchanged. Since the false change rate increases with the threshold, a higher threshold is preferred. However, the threshold also requires a certain degree of tolerance of fluctuations in the unknown data (other meetings). Therefore, in the new algorithm, the value of α is set to 500, and the value threshold is assigned to be 120000.

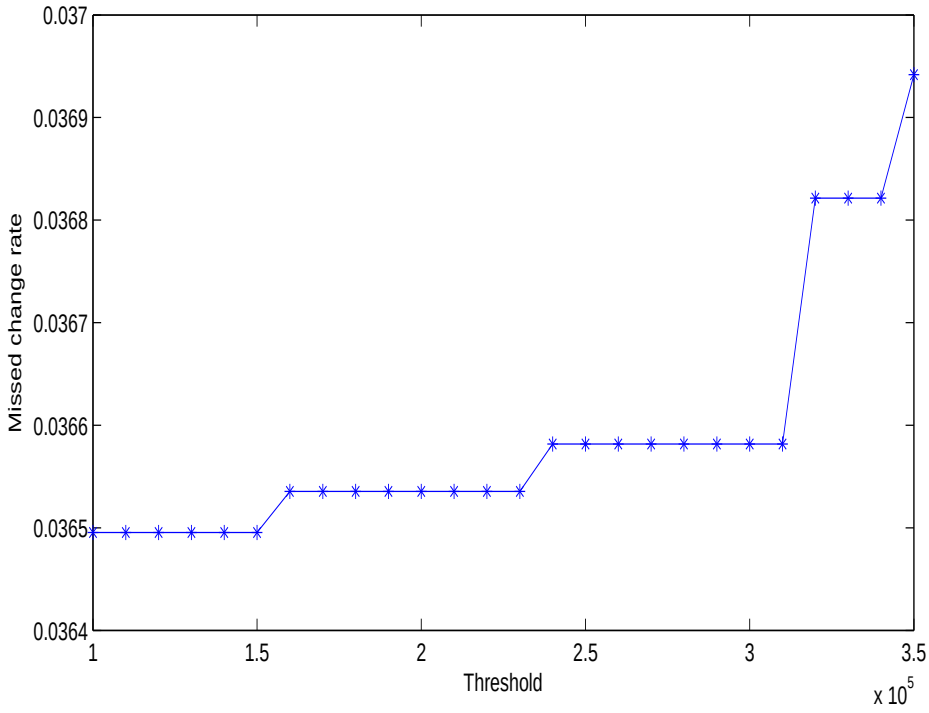


Figure 4.2: The variation in the missed change rate as the threshold increases, using the new measurement.

4.3 Comparing the new SCD algorithm with the KL2-based SCD algorithm

In the baseline system, a similar SCD algorithm is applied using the KL2 Divergence to determine whether two segments are from the same speaker. The variation of the missed change rate as a function of the threshold is shown in Figure 4.3 for the baseline SCD algorithm. The experiment set-up is the same as those described in the previous section. The range of the threshold that minimises the missed change rate is below 40. By comparing Figure 4.2 and 4.3, it can be seen that the missed change rate increases more rapidly when the KL2 Divergence-based SCD algorithm is applied. Therefore, the new algorithm is less affected by the choice of the threshold. In the baseline system, the threshold value for the KL2 Divergence is set to 30.

The mean missed change rate averaged over the fifteen meetings is shown in Figure 4.4(a). The mean false change rate is given in Figure 4.4(b), and the standard deviation of the missed change rate is illustrated in 4.4(c). We can conclude that the new algorithm obtains lower error rates for both types of errors. At the same time, the smaller standard deviation value demonstrates that the new algorithm is less affected by the variability of the data.

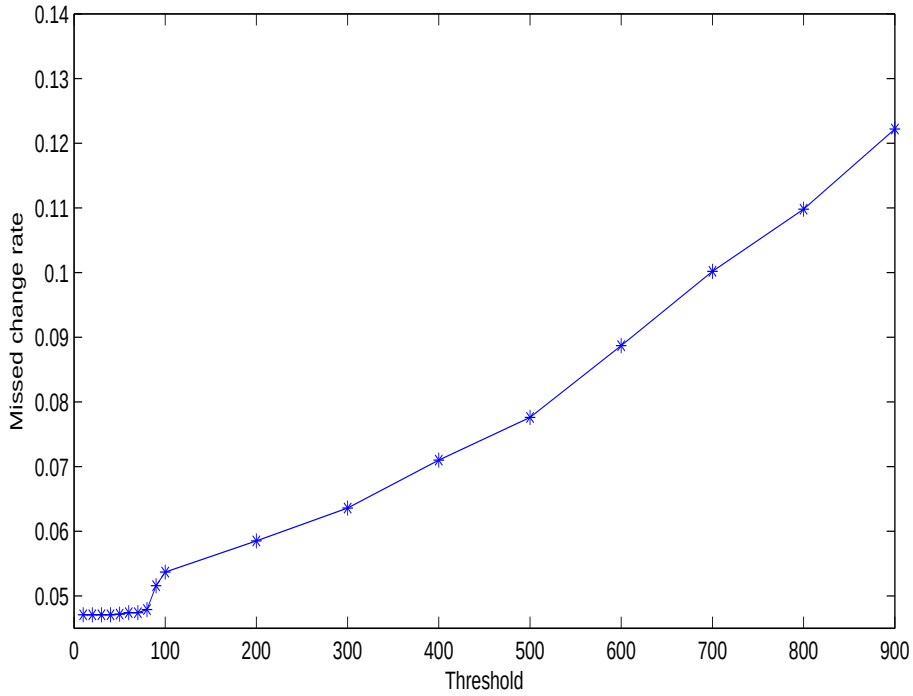
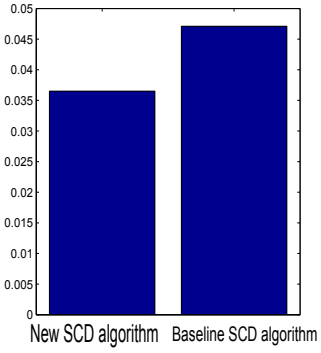
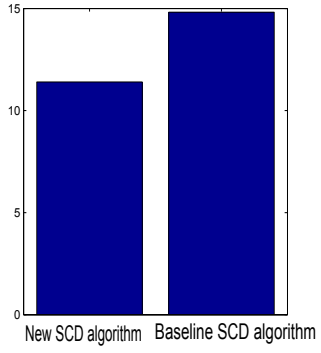


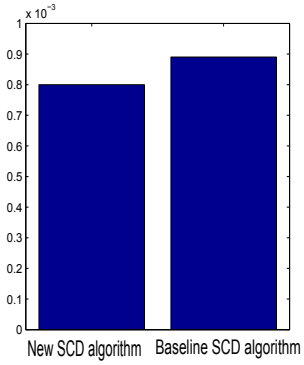
Figure 4.3: The variation in the missed change rate as the threshold increases using the KL2 Divergence.



(a) The mean value of the missed change rate.



(b) The mean value of the false change rate.



(c) The standard deviation of the missed change rate.

Figure 4.4: Comparison of the new SCD algorithm and the KL2 Divergence based SCD algorithm.

Chapter 5

Model Complexity Determination

In the Chapter 3, the data analysis results showed that fewer components should be included in the GMM for speaker models so as to reduce intra-speaker variance, while more components should be preserved in the UBM so as to represent inter-speaker variance.

In this chapter, a method for calculating the new speaker model complexity is proposed. From data analysis, it has been observed that both the number of components used in the model and the location of their mean values are essential for the success of the system. So the novel method described in this chapter will not only select the appropriate component number, but also arrange these components in their correct position.

In section 5.1, an overview of the model complexity selection criterion will be given. Then a new criterion, named Equal Weight Penalty Criterion will be developed in section 5.2. This criterion can remove extra components in the GMM by using a removal scheme, which is controlled by a parameter δ. The intra-speaker variance can be reduced by setting δ low for speaker modelling, and in the UBM more components will be preserved by increasing the value of δ. Furthermore, a new EM training algorithm derived by Figueiredo and Jain (2002) will be integrated into the new criterion, so as to eliminate extra com-

ponents in the GMM automatically based on parameter dimension (number of parameters). In section 5.3, a weight and mean adaptation UBM that can remove the uncovered components automatically will be explained.

5.1 Model complexity determination

Gaussian Mixture Model (GMM) is a flexible and powerful probabilistic modelling tool. It has been introduced in section 2.2 and it is described by Equation 2.3 and 2.4. The model effectiveness is determined by the number of components in the GMM (model complexity).

Assume the true value of model complexity M is within the range ($M_{min} \leq M \leq M_{max}$). In the Bayesian framework, a way of selecting the model complexity is to choose the one with the highest posterior probability. By Bayes theorem, the posterior probability of model complexity M_l given dataset X is defined by Equation 5.1:

$$p(M_l|X) = \frac{p(X|M_l)p(M_l)}{\sum_{M_r=M_{min}}^{M_{max}} p(X|M_r)p(M_r)} \tag{5.1}$$

where $p(X|M_l)$ is the conditional probability of X given the model complexity M_l and $p(M_l)$ is its prior probability. Thus the optimum model complexity $\hat{M}$ satisfies Equation 5.2:

$$\hat{M} = \arg\max_{M} \left[\log p(X|M) + \log p(M)\right] \quad (M_{min} \leq M \leq M_{max}) \tag{5.2}$$

The right hand side of Equation 5.2 can be treated as a model selection criterion. Its first term concerns how the model with complexity M fits X and the second term focuses on the model with complexity M. It may not be restricted to the prior probability of the complexity, it can be the smoothness of the model, its

parameter distribution, and so on. The second term can be generalized as a penalty term; and then a generalized model complexity selection criterion has the form of Equation 5.3:

$$\hat{M} = \arg\min_{M} IC(\hat{\lambda_M}, M) \tag{5.3}$$

where $IC(\hat{\lambda_M}, M)$ is defined by Equation 5.4:

$$IC(\hat{\lambda_M}, M) = -\log p(X|\hat{\lambda_M}) + Pe(M, \hat{\lambda_M}) \tag{5.4}$$

where $\hat{\lambda_M}$ is the ML estimate (has been introduced in section 2.2.3) of GMM parameters λ_M when M components are included. $Pe(M, \hat{\lambda_M})$ is the penalty term. Since the data's likelihood will not decrease when M increases, $Pe(M)$ takes the opposite sign to the second term in Equation 5.2 in order to penalize higher values of M.

Five main types of such criteria have been used for selecting model complexity (McLachlan and Peel, 2000):

1. Bias correction based criteria using $Pe(M)$ to eliminate the Kullback Leibler (KL) Divergence between the true distribution and the estimated approximation based on the samples. Bootstrap-Based Information criterion (McLachlan, 1987) and Cross-Validation-Based information criterion (Smyth, 2000) belong to this type.
2. Laplace Approximation (Schwarz, 1978) based information criteria have been derived within a Bayesian framework for model selection, but it can be applied also in a non-Bayesian framework. It approximates the Equation 5.2 to select $\hat{M}$ with the highest posterior probability. Examples of this kind of criteria include BIC (Campbell et al., 1997) (Dasgupta and Raftery, 1998) (Fraley and Raftery, 1998), Laplace-Empirical Criterion (Roberts et al., 1998)

and Laplace-Metropodis Criterion (Meinicke and Ritter, 2001).

3. Coding theory based criterion select $\hat{M}$ by minimizing the code length necessary to describe the parameter λ_M and to represent the data given the parameter $\hat{\lambda_M}$. MDL criterion (Rissanen, 1989) (Cover and Hall, 1991), Minimum Message Length criterion (Oliver et al., 1996) (Wallace and Dowe, 1999) (Wallace and Freeman, 1987), Akaike's Information Criterion (Whindham and Cutler, 1992), and Information Complexity Criteria (Bozdogan, 1993) all exploit coding theory.

4. Classification based Information Criteria takes the classification likelihood of the data into account when determining model complexity (Banfield and Raftery, 1997) (Cheung, 2005). Classification likelihood is applied in the EM framework as complete-data likelihood for model fitting. It uses $Pe(M)$ to penalize the model whose components are not well-apart. Classification Likelihood Criterion (CLC) (Biernacki and Govaert, 1997), Normalized Entropy Criterion (Biernacki and Govaert, 1999) (Celeux and Soromenho, 1996), and Integrated Classification Likelihood (Biernacki et al., 2000) are computed using complete-data information.

5. The Fully Bayesian approach (Neal, 1992) (Rasmussen, 2000) (Richardson and Green, 1997) has been proposed for model selection. The Reversible jump Markov Chain Monte Carlo method is applied for sampling to check model posterior probability (Bensmail et al., 1997) (Mengersen and Robert, 1996) (Roeder and Wasserman, 1997). It is computationally demanding (McLachlan and Peel, 2000), so the Variational Bayes (Richardson and Green, 1997) (Ghahramani and Beal, 2000) has been developed to determine the model complexity under a Bayesian framework. It belongs to the mean field methods (Jaakkola, 2000). The factored posterior distribution of the parameters are updated depending on each other to approximate the

true joint distribution of the parameters $p(\lambda_M)$. This algorithm will remove the components whose posterior probability are close to zero (Attias, 2001) (Corduneanu and Bishop, 2001) (Ueda and Ghahramani, 2002). The updating of model parameters also depends on the EM (Neal and Hinton, 1998), and it can be applied on-line (Sato, 2001).

5.2 Derivation of the new criterion

In the beginning of this chapter, the demands for the new derived model complexity determination criterion were listed. To reduce intra-speaker variance or maintain the inter-speaker variance, how training data affects the modelling procedure needs to be reviewed. In this process the latent variable that links an observation with a particular model is important.

The CLC is a model complexity selection criterion based on these latent variables. By analysing the CLC criterion, the relation between component mixing parameters and the latent variables will be illustrated. Thus, in this section, CLC is introduced first (section 5.2.1), followed by the derivation of the new criterion (section 5.2.2). In section 5.2.3, the model selection criterion developed by Figueiredo and Jain (2002) will be introduced and how the new criterion is integrated into EM algorithm will be described in section 5.2.4.

5.2.1 CLC

A GMM with complexity M has M Gaussian components and its parameters are described by $\lambda_M = \{\mu_i, \Sigma_i, w_i\}$ where $i = 1, \cdots, M$. Assuming that the dataset $X = \{x_1, \cdots, x_N\}$ are features that are independently and identically distributed

(iid) according to the model, then their generation mechanism is described by:

$$p(x|\lambda_M) = \sum_{j=1}^{M} w_j g_j(x|\mu_j, \Sigma_j) \tag{5.5}$$

and the likelihood of dataset X follows Equation 5.6

$$L(X|\lambda_M) = \sum_{i=1}^{N} \log \sum_{j=1}^{M} w_j g_j(x_i|\mu_j, \Sigma_j). \tag{5.6}$$

Let $Z = \{z^1, \cdots, z^N\}$ be the latent variables that show the component from which the observations originate. In contrast to w, which is the probability of x_i generated from each component in the GMM, z is an indicator parameter that relates x_i to the component containing the highest probability of x_i occurrence.

$$\begin{aligned}
&z^i = \{z_1^i, \cdots, z_M^i\}^T, \\
&z_j^i = 1 \quad x_i \; is \; from \; component \; j, \\
&z_j^i = 0 \quad otherwise, \\
&\sum_{j=1}^{M} z_j^i = 1 \quad (1 \leq i \leq N, \; 1 \leq j \leq M).
\end{aligned}$$

The probability that x_i is generated by a particular component can be calculated by Equation 2.4.

The observations X is referred to as incomplete data, and $\{X, Z\}$ is called complete data. Assume x_i is randomly generated from one of the components, and the Z are iid given model parameters. Further assume that Z is a multinomial distribution, so that the marginal joint density of Z is given by Equation 5.7

$$p(Z|\lambda_M) = \prod_{i=1}^{N} \prod_{j=1}^{M} (w_j)^{z_j^i} \tag{5.7}$$

Suppose X are conditionally independent given Z, the conditional density of X given Z is described by Equation 5.8

$$p(X|Z,\lambda_M) = \prod_{i=1}^{N}\prod_{j=1}^{M} g(x_i|\mu_j,\Sigma_j)^{z_j^i} \tag{5.8}$$

Consequently the joint density of the complete data is given by Equation 5.9

$$p(X,Z|\lambda_M) = \prod_{i=1}^{N}\prod_{j=1}^{M} (w_j g(x_i|\mu_j,\Sigma_j))^{z_j^i} \tag{5.9}$$

Therefore the complete data log likelihood is given by Equation 5.10:

$$L_c(X,Z|\lambda_M) = \sum_{i=1}^{N}\sum_{j=1}^{M} z_j^i(\log w_j + \log g(x_i|\mu_j,\Sigma_j)) \tag{5.10}$$

The $L_c(X,Z|\lambda_M)$ is also referred to as classification log likelihood. How the classification information is contained can be shown by the link between $L_c(X,Z|\lambda_M)$ and $L(X|\lambda_M)$ described by Equation 5.11.

$$\begin{aligned} EC_M(X|\lambda_M) &= L_c(X,Z|\lambda_M) - L(X|\lambda_M) \\ &= -\sum_{i=1}^{N}\sum_{j=1}^{M} z_j^i \log \tau_j^i \end{aligned} \tag{5.11}$$

where τ is described by Equation 5.12.

$$\begin{aligned} \tau_j^i &= Pr(z_j^i = 1|x_i,\lambda_M) \\ &= \frac{w_j g(x_i|\mu_j,\Sigma_j)}{\sum_{j=1}^{M} w_j g(x_i|\mu_j,\Sigma_j)} \end{aligned} \tag{5.12}$$

τ_j^i is the posterior probability of the jth component given x_i. It is also equal to $Pr(z_j^i = 1|x_i,\lambda_M)$, which is the conditional probability of x_i from the component j given x_i and λ_M. EC_M is the entropy of Z.

The entropy EC_M is a measure of the ability of the M component mixture model to partition dataset X. If X is well separated by the M components, $EC_M \approx 0$. However, if the mixture components are mingled together, EC_M has a large value. Therefore, EC_M and τ provide data classification information.

CLC is a model complexity selection criterion that uses EC_M as the penalty term Pe. Thus the criterion selects the model that maximizes the complete data likelihood, and as a result it prefers the model that spreads apart the data. But the CLC criterion does not consider the influence of parameter dimension on the model's generality. Moreover, the prior distribution of other parameters is also neglected. Therefore, a new criterion, Equal Weight Penalty Criterion (EWPC) will be developed to overcome these drawbacks and make the model selection fits the UBM better.

5.2.2 Equal Weight Penalty Criterion (EWPC)

When there are extra components included in a mixture model, they may have little data to support the existence of the components (Ueda and Nakano, 1998) or share close position with other components (Hofmann and Buhmann, 1997). In the first case, these components have a low mixing parameter $w_j \approx 0$; in the second case, they have similar weight parameters in the mixture (Ueda et al., 2000). The first kind of extra components can be removed by the criteria that penalizes the model parameter dimension (the number of parameters in the model) or removes these components with $w_j \approx 0$. In the new criterion a penalty term based on the KL divergence of the prior and posterior distribution of w is adopted to overcome the second situation.

The conjugate prior of the multinomial distribution is the prior distribution of w, $p_0(w)$. w follows a Dirichlet distribution $Dir(\delta_1, \cdots, \delta_M)$ (Bernardo and Smith, 1994), where parameter δ controls the shape of the distribution. $p_0(w)$ is

set according to Equation 5.13:

$$p_0(w|\delta) = Dir(w|\delta) \propto w_1^{\delta-1} \cdots w_j^{\delta-1} \cdots w_M^{\delta-1} \tag{5.13}$$

The change of $p_0(w)$ with different values of δ in a one dimensional case is shown in Figure 5.1. It can be seen that when $\delta < 1$ the distribution has a concave shape

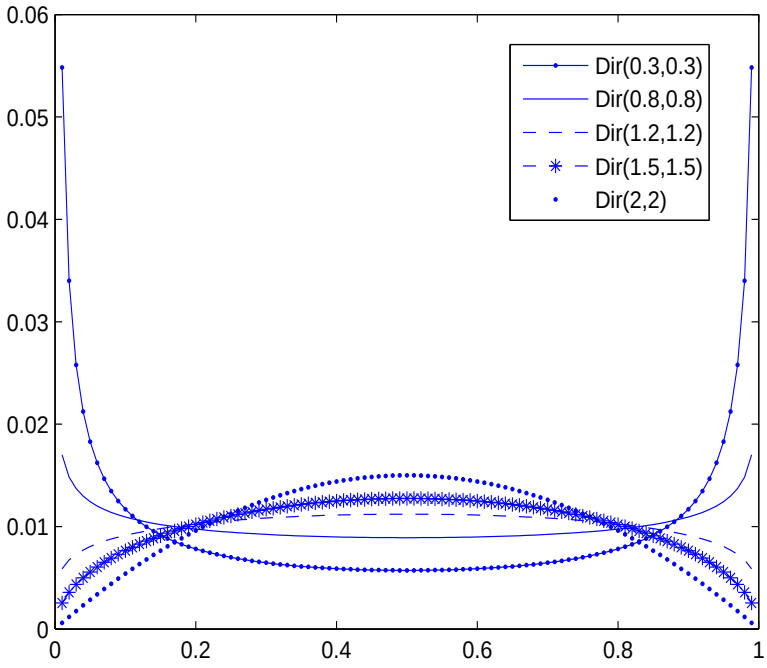


Figure 5.1: Dirichlet prior with different negative parameter.

and that w has a high probability when it is near 0 or 1. When δ approaches 1 the distribution becomes flatter; and when $\delta > 1$ the distribution has a convex shape reaching its highest value at $w = 0.5$. The Dirichlet distribution of w with less than 1 makes the existence of the components unstable and they must 'compete to survive'. By controlling the value of δ, different prior distributions for $p_0(w)$ can be obtained. Assuming a concave distribution for $p_0(w)$ with $\delta < 1$, the prior favours w with value near 0 or 1. The KL divergence measures the differ-

ence between the prior distribution and the posterior distribution. If the posterior distribution has a flat distribution, $D_{KL}(p_{w|z}, p_0)$ (defined later in Equation 5.17) becomes large and the model will be penalized more. The competition between components is fierce and among two components that share the same data space; only one will win and the other will be removed. If a flatter Dirichlet distribution is applied to $p_0(w)$, more components are allowed in the mixture model.

The new criterion measures the KL divergence between $p(w|Z)$ and $p_0(w)$ with respect to $p(w|Z)$, and is labelled as $D_{KL}(p_{w|z}, p_0)$. Setting $\delta' = \delta - 1$, then the prior distribution of w follows

$$\begin{aligned} p_0(w) &= \prod_{j=1}^{M} w_j^{\delta'} / A_1 \\ A_1 &= \frac{\Gamma(\delta)^M}{\Gamma(M\delta)} \end{aligned} \tag{5.14}$$

where Γ is the Gamma function. The distribution of $p(Z|w)$ follows:

$$p(Z|w) \propto w_1^{\sum_{i=1}^{N} z_1^i} \cdots w_j^{\sum_{i=1}^{N} z_j^i} \cdots w_M^{\sum_{i=1}^{N} z_M^i} \tag{5.15}$$

Since $p(w|Z) \propto p_0(w)p(Z|w)$,

$$\begin{aligned} p(w|Z) &= \prod_{j=1}^{M} w_j^{\sum_{i=1}^{N} z_j^i + \delta'} / A_2 \\ A_2 &= \frac{\Gamma(\sum_{i=1}^{N} z_1^i + \delta) \cdots \Gamma(\sum_{i=1}^{N} z_M^i + \delta)}{\Gamma(N + M\delta)} \end{aligned} \tag{5.16}$$

Submitting Equation 5.14 and Equation 5.16 into $D_{KL}(p_{w|z}, p_0)$,

$$
\begin{aligned}
D_{KL}(p_{w|z}, p_0) &= E_{p(w|Z)}\{\log(p_{w|Z}) - \log(p_0(w))\} \\
&= \sum_{j=1}^{M}(\sum_{i=1}^{N} z_j^i \log(\hat{w_j})) + \log A \\
A = A_1/A_2 &= \frac{\Gamma(\delta)^M \Gamma(N + M\delta)}{\Gamma(\sum_{i=1}^{N} z_1^i + \delta) \cdots \Gamma(\sum_{i=1}^{N} z_M^i + \delta)\Gamma(M\delta)}
\end{aligned} \tag{5.17}
$$

where $E_{p(w|Z)}\{\cdots\}$ is the expected value of $\{\cdots\}$ with respect to the probability density function $p(w|Z)$. Using the absolute value of Equation 5.17 as Pe, the extra components will be removed from the model. However, the influence of the data size and the number of parameters also need to be taken into consideration. BIC criterion (defined in Equation 2.8) selects the appropriate model complexity depending on both data size and parameter dimension (the number of parameters used in the model). Applying BIC to approximate $L(X|\lambda_M)$, the new criterion becomes:

$$
\begin{aligned}
EWPC_M &= -\log p(X|\hat{\lambda_M}) + Pe(M, \hat{\lambda_M}) \\
&= -\log p(X|\hat{\lambda_M}) + |\frac{1}{2} M \log N + D_{KL}(p_{w|z}, p_0)| \\
&= -\log p(X|\hat{\lambda_M}) + \frac{1}{2} M \log N + |\sum_{j=1}^{M}(\sum_{i=1}^{N} z_j^i \log(\hat{w_j})) + \log A| \\
&= -\log p(X|\hat{\lambda_M}) + \frac{1}{2} M \log N + |\sum_{j=1}^{M}(N\hat{w_j}) \log \hat{w_j} + \log A| \\
\hat{M} &= \arg\min_M EWPC_M
\end{aligned} \tag{5.18}
$$

Two examples are used to show the performance of the new criterion. In the first example, 1000 samples are generated from a four component bivariate GMM. They are referred to as dataset1 below, and the samples and their generation model is illustrated in the sub-figure (a) of Figure 5.2. All the components

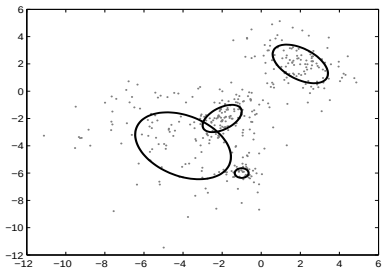
have different means and are located close to each other. One of them has a low variance and a low mixing proportion. GMMs are trained based on dataset1 using the EM algorithm. The range of the model complexity is $1 \leq M \leq 10$, and the EWPC is applied to select the optimum component number that minimizes $EWPC_M$. A random initialization of GMM with 10 components is shown in Figure 5.2 (b). Figure 5.2 (c)-(f) shows the GMM selected by the EWPC with different settings for parameter δ.

It can be seen from Figure 5.2 that when $\delta = 0.3$ and $\delta = 0.5$, the EWPC selects the correct model for the dataset. As the value of δ increases, the criterion allows more components to be contained in the model. When $\delta = 0.1$, the smallest component fails the competition and the larger component occupies its space. If the dataset is well separated in the space, the EWPC selects the true generated model no matter what the value of δ. This will be illustrated in the next example.

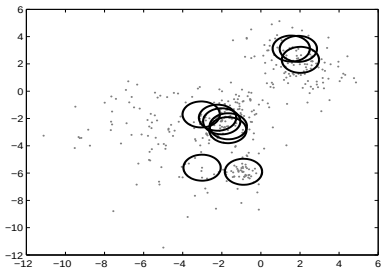
In the second example, 1000 samples are generated from a different four-component bivariate GMM. This time they are well separated from each other. They are referred to as dataset2 below, and the samples and their generated model are illustrated in the Figure 5.3 (a). It can been in the figure that the same model is selected by different settings of δ. Therefore, the EWPC will control the component number by the parameter setting only when the data distribution is ambiguous or overlapped.

When applying the new criterion to select the model complexity, $EWPC_M$ (represented by Equation 5.18) needs to be calculated for a range of M, from M_{min} to M_{max}. It is time consuming because $\hat{\lambda_M}$ needs to be estimated for each M. Although to obtain $\hat{\lambda_{M-1}}$ by EM, the model can be initialized by removing the component with least likelihood in $\hat{\lambda_M}$, but the algorithm is still inefficient.

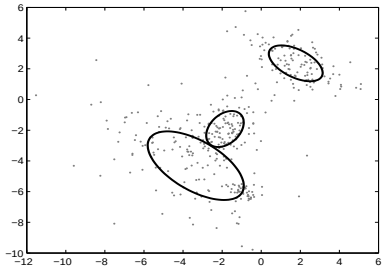
Another problem that will be encountered by the new criterion is due to EM.



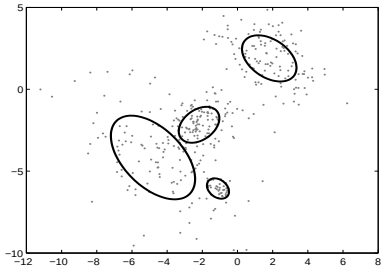
(a)The experimental dataset and the true mixture model. The data is denoted by gray points and the model is represented by the ellipse.



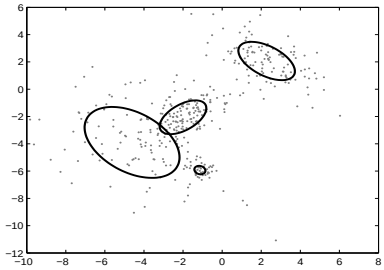
(b)Random initialization with 10 mixtures.



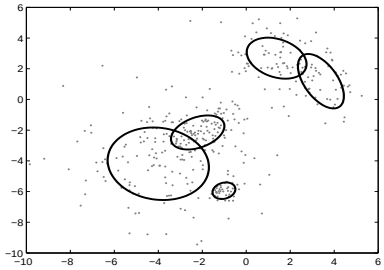
(c)The model selected by the EWPC, set $\delta = 0.1$. It has 3 mixtures.



(d)The model selected by the EWPC, set $\delta = 0.3$. It has 4 mixtures.

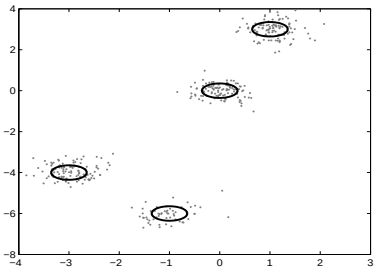


(e)The model selected by the EWPC, set $\delta = 0.5$. It has 4 mixtures.

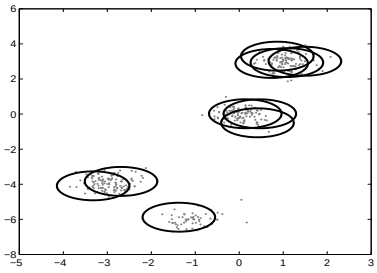


(f)The model selected by the EWPC, set $\delta = 0.8$. It has 5 mixtures.

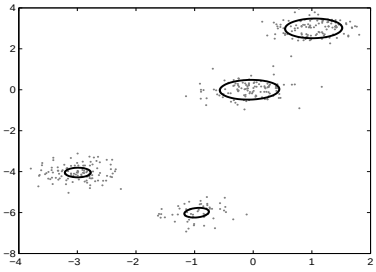
Figure 5.2: Fitting a GMM to dataset1 according to EWPC



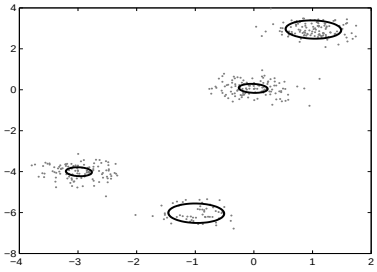
(a)The experimental dataset and the true mixture model. The data is denoted by gray points and the model is represented by the ellipse.



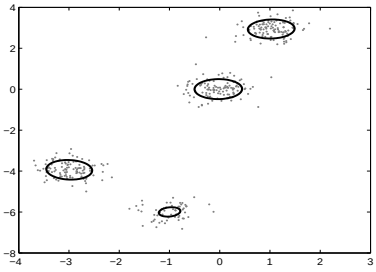
(b)Random initialization with 10 mixtures.



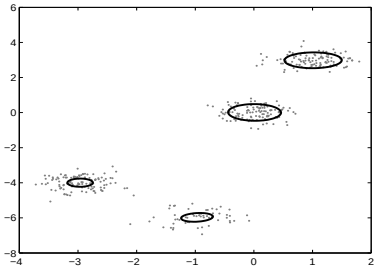
(c)The model selected by the new criterion, set $\delta = 0.1$. It has 4 mixtures



(d)The model selected by the new criterion, set $\delta = 0.3$. It has 4 mixtures



(e)The model selected by the new criterion, set $\delta = 0.5$. It has 4 mixtures



(f)The model selected by the new criterion, set $\delta = 0.8$. It has 4 mixtures

Figure 5.3: Fitting a GMM to dataset2 according to EWPC

EM has two main drawbacks. First, it is sensitive to the initialization; second, it may converge to the boundary of the parameter space (Kloppenburg and Tavan, 1997) (Meinicke and Ritter, 2001). Thus to select the initialization model complexity, M_{max} is difficult to compute. Using high M_{max} results in a heavy computational burden, and will increase the risk of components converging to the space boundary (Rose, 1998). On the other hand, using low M_{max} the model cannot well fit the features.

To overcome these problems, the criterion developed in (Figueiredo and Jain, 2002) will be integrated into the EWPC. It is based on Laplace's Method of Approximation, which will be introduced in the next section.

5.2.3 Laplace's Method of Approximation

The marginal distribution of dataset X can be described by Equation 5.19, given model complexity M:

$$\begin{aligned} p(X|M) &= \int p(X, \lambda_M) d\lambda_M \\ &= \int \exp\{\log p(X, \lambda_M)\} d\lambda_M \end{aligned} \tag{5.19}$$

Using second-order Taylor series to approximate $p(X, \lambda_M)$ at $\lambda_M = \tilde{\lambda_M}$,

$$\log p(X, \lambda_M) \approx \log p(X, \tilde{\lambda_M}) - \frac{1}{2}(\lambda_M - \tilde{\lambda_M})^T H(\tilde{\lambda_M})(\lambda_M - \tilde{\lambda_M}) \tag{5.20}$$

where $\tilde{\lambda_M}$ denotes the posterior mode of λ_M satisfying $\partial \log p(X, \tilde{\lambda_M})/\partial \lambda_M = 0$. $H(\tilde{\lambda_M})$ is the negative Hessian matrix of $\log p(X, \lambda_M)$ evaluated at $\lambda_M = \tilde{\lambda_M}$.

Substituting the expansion described by Equation 5.20 into Equation 5.19,

$$p(X|M) = \exp\{\log p(X, \tilde{\lambda_M})\} \int \exp\{-\frac{1}{2}(\lambda_M - \tilde{\lambda_M})^T H(\tilde{\lambda_M})(\lambda_M - \tilde{\lambda_M})\} d\lambda_M$$
$$= p(X, \tilde{\lambda_M})(2\pi)^{\frac{1}{2}\bar{D}} |H(\tilde{\lambda_M})|^{-\frac{1}{2}} \tag{5.21}$$

Therefore, from Equation 5.21, the marginal log likelihood can be approximated as

$$\log p(X|M) \approx \log p(X|\tilde{\lambda_M}) + \log p(\tilde{\lambda_M}) - \frac{1}{2} log|H(\tilde{\lambda_M})| + \frac{1}{2}\bar{D} log(2\pi) \tag{5.22}$$

Usually, the ML estimate $\hat{\lambda_M}$, is used instead of the posterior mode $\tilde{\lambda_M}$. Since the negative Hessian matrix is the negative of the square matrix of second-order partial derivatives of all parameters, $H(\hat{\lambda_M})$ is equal to the observed information matrix $I(\hat{\lambda_M}|X)$, which is the negative of the second derivative of the logarithm of the likelihood function based on observations in dataset X. Then Equation 5.22 can be approximated by

$$\log p(X|M) \approx \log P(X|\hat{\lambda_M}) + \log p(\hat{\lambda_M}) - \frac{1}{2} \log |I(\hat{\lambda_M}|X)| + \frac{1}{2}\bar{D} log(2\pi) \tag{5.23}$$

The BIC criterion is derived by replacing $\log |I(\hat{\lambda_M}|X)|$ as $M \log N$, as described by Equation 2.8. $M \log N$ is the number of parameters in the GMM. $\frac{1}{2}\bar{D} log(2\pi)$ is a constant term and when the size of X increases, this term will become considerably small compared with other terms. So it is treated as an o(1) term (a term that converges to 0 when data size is large) and ignored in BIC.

Figueiredo and Jain (2002) integrated the model selection criterion in the likelihood function, so the model complexity can be optimized gradually using

EM. It approximates $|I(\hat{\lambda_M}|X)|$ using the complete data information matrix with a block diagonal structure (Titterington et al., 1991) (McLachlan and Peel, 1997):

$$I_c = N\ blockdiag\{w_1 I^{(1)}(\mu_1, \Sigma_1), \cdots, w_M I^{(1)}(\mu_M, \Sigma_M), \Lambda\} \tag{5.24}$$

where $|\Lambda| = (w_1 w_2 \cdots w_M)^{-1}$.

$blockdiag$ refers to a block diagonal matrix, which is a square diagonal matrix in which the diagonal elements are square matrices of any size, and the off-diagonal elements are 0.

Square matrices $w_1 I^{(1)}(\mu_1, \Sigma_1), \cdots, w_M I^{(1)}(\mu_M, \Sigma_M)$ and Λ are on the diagonal of $blockdiag\{w_1 I^{(1)}(\mu_1, \Sigma_1), \cdots, w_M I^{(1)}(\mu_M, \Sigma_M), V\}$ and the blocks off the diagonal are zero matrices. $I^{(1)}(\mu_i, \Sigma_i)$ is the observed information matrix with respect to component i's parameters μ_i and Σ_i given a single observation. Therefore the value of $|I_c|$ is defined as Equation 5.25:

$$\log |I(\hat{\lambda_M}|X)| = \sum_{j=1}^{M} \Omega(\mu, \Sigma) \log N w_j - \sum_{j=1}^{M} \log w_j + \sum_{j=1}^{M} \log |I^{(1)}(\mu_j, \Sigma_j)| \tag{5.25}$$

where $\Omega(\mu, \Sigma)$ represents the number of parameters in a Gaussian component. Assume parameters of different components are independent, and the mixing parameters are independent of the Gaussian parameters, the standard non-informative Jeffrey's prior of the parameters is adopted as Equation 5.26:

$$\begin{aligned} p(\hat{\lambda_M}) = p(w, \mu, \Sigma) &= p(w_1) \cdots p(w_M) \prod_{j=1}^{M} p(\mu_j, \Sigma_j) \\ &\propto (w_1 w_2 \cdots w_M)^{-1/2} \prod_{j=1}^{M} (|I^{(1)}(\mu_1, \Sigma_1)|)^{1/2} \end{aligned} \tag{5.26}$$

Therefore,

$$\log p(\hat{\lambda_M}) = \log p(w, \mu, \Sigma) = -1/2 \sum_{j=1}^{M} \log w_j + 1/2 \sum_{j=1}^{M} \log I^{(1)}(\mu_j, \Sigma_j) \tag{5.27}$$

Substituting Equation 5.25 and 5.27 into 5.23, then

$$\log p(X|M) \approx \log L(\hat{\lambda_M}) - \frac{1}{2}\Omega(\mu, \Sigma) \sum_{j=1}^{M} \log(N w_j) + \frac{1}{2}\bar{D} log(2\pi) \tag{5.28}$$

Neglecting the last term (because it is an o(1) term), Equation 5.28 becomes:

$$\log p(X|M) \approx \log L(\hat{\lambda_M}) - \frac{1}{2}\Omega(\mu, \Sigma) \sum_{j=1}^{M} \log w_j - \frac{1}{2}\Omega(\mu, \Sigma) \log N \tag{5.29}$$

Then the model complexity selection criterion is described as:

$$\hat{M} = \arg\min_{M} \log p(X|M) \tag{5.30}$$

This criterion has an intuitively appealing interpretation. For each component, the expected number of data points generated from it is Nw_j. According to BIC, the model complexity is penalized by $\Omega(\mu, \Sigma) \log(Nw_j)$. Thus the criterion check for each component is whether there is sufficient evidence for its existence according to BIC. This criterion can be integrated into the EM algorithm, which selects the model complexity automatically during model training. In the next section, the EM algorithm will be introduced briefly.

5.2.4 EM algorithm for GMM parameter estimation

The EM algorithm (Dempster et al., 1977) is a general method of obtaining the maximum-likelihood estimation of the parameters of an underlying distri-

bution from a given incomplete data set. There are two main types of incomplete dataset. The first occurs when the data loses parts of values due to problem restriction or observation process. The second assumes the existence of additional hidden parameters to simplify the optimizing of the likelihood function that is analytically intractable. The EM algorithm can also be applied to find a MAP estimate of the parameters, where MAP is a mode of the parameter's posterior distribution. (McLachlan and Peel, 1997).

For parameter estimation in the GMM, the indicator parameter Z is applied as the latent set of parameters (McLanchlan and Basford, 1988) (McLachlan and Peel, 1997). Z has been introduced in Section 5.2.1 and it shows from which component of the GMM an observation originates. Since the values of Z are unknown in the parameter estimation process, the EM algorithm replaces them by their expected value conditioned by observations of X; and then obtains the parameters $\lambda_M = \{\mu_i, \Sigma_i, w_i\}$ by maximizing the $L_c(X, Z|\lambda_M)$. This procedure includes an E-step and an M-step and these two steps will be run iteratively until the stop criterion is met.

The EM algorithm can be described as follows:

- initialize: Initialize the $\lambda_M = \{\mu_i, \Sigma_i, w_i\}$ as $\lambda_M^1 = \{\mu_i^1, \Sigma_i^1, w_i^1\}$.
- In the E-step: At the tth iteration, assume τ_j^i (described by Equation 5.12) denotes the expected value of z_j^i given the value obtained at the last iteration $t-1$, then the complete data log likelihood conditioned on λ^{t-1} can be presented as Equation 5.31:

$$Q(\lambda_M, \lambda_M^{t-1}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \tau_j^i (\log w_i + \log g(x_i|\mu_j, \Sigma_j)) \tag{5.31}$$

- In the M Step: Take the derivative of Equation 5.31 with respect to w_j, μ_j, and Σ_j respectively, and the optimum values that maximize Equation 5.31

will be obtained. It follows that

$$\lambda_M^t = \arg\max_{\lambda_M} Q(\lambda_M, \lambda_M^{t-1}) \tag{5.32}$$

where

$$\begin{aligned} w_j^t &= \sum_{i=1}^{N} \tau_j^i / N \\ \mu_j^t &= \sum_{i=1}^{N} \tau_j^i x_i / \sum_{i=1}^{N} \tau_j^i \\ \Sigma_j^t &= \sum_{i=1}^{N} \tau_j^i (x_i - \mu_j)(x_i - \mu_j)^T / \sum_{i=1}^{N} \tau_j^i \end{aligned} \tag{5.33}$$

for $j = 1$ to M.

- Stop criterion: The E-step and M-stem will be operated iteratively until the log likelihood of the observations $L(X|\lambda_M)$ increases no further.

5.2.5 Integrating the model complexity selection in the EM

Figueiredo and Jain had integrated an EM algorithm into their criterion, which will find the MAP estimate of parameters, and at the same time removes extra components in the GMM (Figueiredo and Jain, 2002).

Integrating the second term of Equation 5.29, $-\frac{1}{2}\Omega(\mu, \Sigma)\log(w_j)$, into the $Q(\lambda_M, \lambda_M^{t-1})$ defined by Equation 5.31 in order to maximize it with respect to w_i results in Equation 5.34:

$$\frac{\partial \left[\sum_{i=1}^{N} \tau_j^i \log w_i - \frac{1}{2}\Omega(\mu, \Sigma)\log(w_j)\right]}{\partial w_j} = 0 \tag{5.34}$$

where $\sum_{j=1}^{M} w_j = 1$. Therefore in iteration t, w_j will be updated by

$$w_j^t = \frac{\max\{0, \sum_{i=1}^{N} \tau_j^i - \frac{1}{2}\Omega(\mu, \Sigma)\}}{\sum_{t=1}^{M} \max\{0, \sum_{i=1}^{N} \tau_t^i - \frac{1}{2}\Omega(\mu, \Sigma)\}} \tag{5.35}$$

instead of what has been described in Equation 5.33. The component with a weight less than $\frac{1}{2}\Omega(\mu, \Sigma)$ will be removed automatically from the model. The term $\frac{1}{2}\Omega(\mu, \Sigma)$ is only a part of the criterion described in Equation 5.29, so the selection of the model complexity still needs to go through every possible M. However, allowing the weight of parts of the components to reduce to zero and removing them automatically in EM training will greatly accelerate the UBM training.

The performance of the EM depends heavily on initialization. Since EM is a localized algorithm, if its initial values fail to cover some of the data space that space may never be covered by the model. To initialize the model with enough components to cover all of the data space is a way to solve the problem, but it will cause a singularity of the covariance matrix. When w_j approaches 0, the corresponding covariance matrix may become arbitrarily close to singular. If the number of components assumed is much larger than what is optimal, this tends to happen frequently. However, by removing the jth component once w_j is less than $\Omega(\mu, \Sigma)$, this will be avoided.

To integrate this model selection criterion into EWPC, Equation 5.29 is applied to approximate the term $\log L(\hat{\lambda_M}) - \frac{1}{2}\Omega(M) \log N$ in Equation 5.18. Then EWPC becomes

$$EWPC_M = -\log L(\hat{\lambda_M}) + \frac{1}{2}\Omega(\mu, \Sigma) \sum_{j=1}^{M} \log N w_j + |\sum_{j=1}^{M} (N\hat{w}_j) \log \hat{w}_j + \log A| \tag{5.36}$$

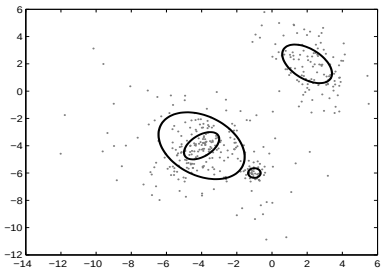
When training the speaker model or the UBM with the EM, the model selec-

tion procedure is first run automatically in EM by updating w_j using Equation 5.35. Then EWPC of the model is calculated as described in Equation 5.36. The components with the smallest likelihood will be removed from the model and the EM training is run again. Finally, the model whose EWPC is smallest will be picked as the optimal model. The performance of the proposed EWPC-based model complexity selection mechanism is illustrated in Figure 5.4. In this example, 1000 samples are generated from a four-component bivariate GMM with plenty of overlap. They are referred to as dataset3 below, and the samples and their generation model is illustrated in the Figure 5.4 (a). In Figure 5.4 (b), the random initialization with ten-components is shown. It automatically shrinks to a six-component model in Figure 5.4 (c). The optimal model selected by this criterion is shown in Figure 5.4 (d). EWPC can also select the correct model when $0.3 \leq \delta \leq 0.6$, as illustrated in Figure 5.4 (e). However, in Figure 5.4 (f), (set $\delta \leq 0.2$) EWPC prefers the model with fewer component.

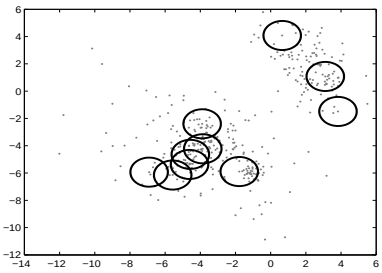
5.3 Efficient sample size UBM adaptation

In a speaker diarization system, where segments are clustered according to the speakers, a model is built for each segment by adapting them from the UBM. It has been shown in section 3.4 that short segments cannot cover all the subspace of a particular speaker. Although the UBM can be used to help understand the subspace structure, it may reduce the inter-speaker variance if there is not enough data for adaptation. Therefore, in this section, an adaptation method is proposed to remove automatically the components that the data in the segments does not support well.

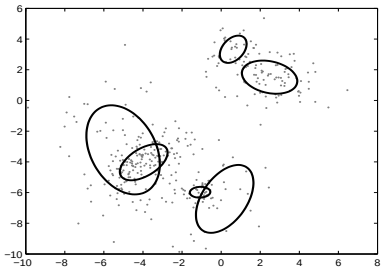
Mean-only adaptation is applied to the task and the resulting segment model has the same model complexity as the UBM. However, if little data in the seg-



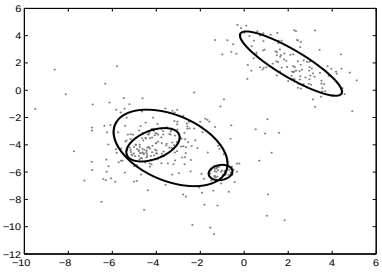
(a)The experimental dataset and the true mixture model. The data is denoted by gray points and the model is represented by the ellipse.



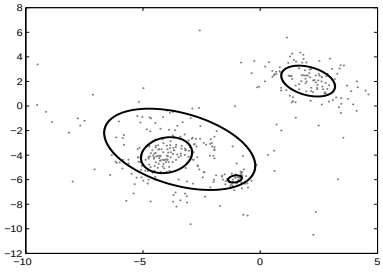
(b)Random initialization with 10 mixtures.



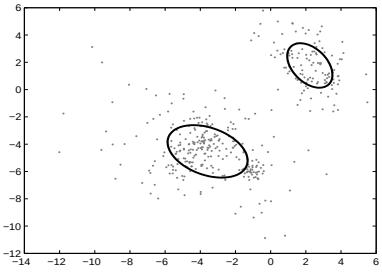
(c)The evolution of the criterion of Figueiredo. After the first iteration with 10 component initialization



(d)The model selected by the criterion of Figueiredo.



(e)The model selected by EWPC, set $\delta = 0.6$



(f)The model selected by the new criterion, set $\delta = 0.2$

Figure 5.4: Fitting GMM to the dataset3 based on the criterion of Figueiredo and Jain (2002) and EWPC.

ment is assigned to a component of the UBM, in the adapted model, the mean value of the component will be dominated by the one from the UBM. It may cause the dissimilarity between the models of two segments to be reduced. In this section, a new UBM adaptation method is described. Both the mean adaptation and the weight adaptation is applied so that the component in the UBM with little data assigned to it will disappear in the segment models. The new adaptation method is based on the criterion of Figueiredo and Jain (2002), which has been described in the last section.

The weight adaptation for a segment from a UBM follows the same formula as Equation 5.35. It is described by Equation 5.37

$$\tilde{w_i} = \frac{\max\{0, \sum_{i=1}^{N} \tau_j^i - \Omega(\mu, \Sigma)\}}{\sum_{t=1}^{M} \max\{0, \sum_{i=1}^{N} \tau_t^i - \Omega(\mu, \Sigma)/2\}} \tag{5.37}$$

where τ_j^i is the posterior of component j given the data x_i from the segment to be adapted. If less than $\Omega(\mu, \Sigma)/2$ data is assigned to the component, it will not appear in the segment's model. This weight adaptation can be explained if one of the components becomes too weak, meaning that it is not supported by enough data, it will be removed. $\Omega(M)/2$ is the threshold to judge if the effective sample size generated from the component is enough (Figueiredo and Jain, 2002).

Then the component's mean will be adapted only for these components whose weight parameter is not zero.

$$\tilde{\mu_i} = \frac{\rho_i \mu_i^{ubm} + \sum_{j=1}^{N} \tau_{ji} x_j}{\rho_i + \sum_{j=1}^{N} \tau_{ji}} \quad if\ w(j) \neq 0 \tag{5.38}$$

where ρ_i is used to control how the UBM's mean affects the adapted mean.

Chapter 6

Experiment and Discussions

In Chapter 3, the characteristics of meetings and those aspects that affect the speaker diarization system were identified. Potential solutions to the system's shortcomings were suggested and developed into new algorithms in Chapter 4 and Chapter 5. In this chapter, experiments will be conducted to evaluate the effectiveness of these novel strategies when they are adopted in the speaker diarization of single channel recorded meetings. In Section 6.1, the meetings used to evaluate the performance of the new algorithms will be introduced. In Section 6.2, the difference between the baseline system and the new system consisting of the new strategies will be presented. In Sections 6.3 through 6.6, the performance of all of the novel strategies derived in this thesis will then be evaluated against the baseline system. In Section 6.7, the overall results will be discussed, and finally, possible conclusions will be drawn.

6.1 Meeting corpus selection

In Chapter 3, fifteen meetings from the AMI corpus were studied to examine the shortcomings of the baseline system in terms of various meeting characteristics. These meetings were also used to select the optimum value for the parameters of

the speaker diarization system, such as the segment lengths of SAD and SCD (in Chapter 3) and the combination parameter α and threshold of SCD (in Chapter 4). In this chapter, we use the same meetings to select other parameters appearing in both the baseline system and the new system. They will be referred to as the "development set" in the rest of the thesis. Because the characteristics detection and parameter selection of all meetings are based on the development set, they are presumed to obtain better results when the new speaker diarization system is applied.

To test the stability of the new strategies and the chosen parameters, we should use other meetings to test the system to check the consistency of the results. Those meetings, which are collectively termed the "evaluation set", should be different from the development set. Therefore, another 30 meetings from the AMI corpus are selected as the evaluation set. They are of different types and are recorded in the Edinburgh Room and the IDIAP Room. None of these meetings is affected by irresolvable recording problems, which in most cases were due to equipment failure. Only single-channel recordings obtained by mixing lapel recordings are used in the evaluation set, based on the same reasoning of development set selection described in Section 3.1. Because the AMI corpus includes very limited types of meetings and the range of speaker numbers in these meetings is narrow (from 3 to 5), meetings from another corpus should be selected to increase the diversity of the evaluation set. Therefore, meetings from the ISL Meeting Corpus Part I (ISL-MC1) are included in the evaluation set.

The ISL Meeting Corpus Part I (ISL-MC1) is the first published subset of the ISL Meeting Corpus (112 meetings). It contains 18 meetings collected at the Interactive Systems Laboratories (ISL) at Carnegie Mellon University. All meetings were recorded in an open-plan office, with background noises similar to a quiet cubicle office environment. Each participant wore a lapel microphone,

and the meetings were recorded directly onto a hard disk at 16 kHz as WAV files. For each meeting, all channels were provided in separate files, as well as a single WAV file containing a mix of all channels tracks. Three meetings, m053, m054 and m057, could not be used in our experiment because of recording problems (some speakers are off microphone). One of the meetings (m039) is recorded in two parts, m039a and m039b. Because our research focus is on single-channel recordings, 16 mixed-channel meeting audios are used in our experiments. The durations of the ISL-MC1 meetings range from 8 to 64 minutes, and the average is 34 minutes. Four types of meetings are included in the ISLMC1: project meetings, discussion, chatting and game playing. Among them, project meetings are natural meetings that occur in the real world, whereas discussion, chatting and game playing are artificial meetings that were designed for the purpose of data collection. The number of speakers appearing in the meetings of the ISL-MC1 ranges from 3 to 9, a wider range than in the meetings of the AMI corpus. A detailed description of ISL-MC1 can be found in (Burger et al., 2002).

The DER, which is the main metric for measuring the performance of the speaker diarization systems, was introduced in Section 2.7. The DER first finds an optimal one to one mapping between the speakers detected by the speaker diarization system and the real speakers. This mapping should minimise the total fraction of time that is attributed to an incorrect source. As introduced in Section 2.8, incorrect attributions occur in three different cases. In this chapter, when speech is rated as non-speech, the resulting error rate is denoted E_{MISS}. The error rate caused by rating non-speech as speech is denoted E_{FA}. When speech is attributed to the wrong speaker, the error rate is denoted E_{spkr}. Equations 2.17, 2.18 and 2.16 described the method to calculate E_{MISS}, E_{FA} and E_{spkr}. DER is used to represent the total error rate, which is the sum of E_{MISS}, E_{FA} and E_{spkr}. When multiple speakers talk at the same time, the speech can be assigned to any

of them without increasing the DER.

The real speakers and the time stamps of their dialogue are provided in the reference document of the corresponding corpus, which is called the transcription. In the AMI corpus, forced alignments are used as the transcription, and they can be downloaded at http://corpus.amiproject.org/download. In the ICL corpus, the meetings are transcribed by hand. Compared to forced alignments, hand transcription extends the durations when multiple speakers speak at the same time and is unreliable for detecting short silence segments or the boundary between speech and non-speech. Table 6.1 summarises the data used in the experiment in this chapter. For a complete list of the individual files, refer to Appendix A.

	Development set	Evaluation set	
Corpus	AMI	AMI	ICL
Number of meetings	15	30	16
Range of speaker numbers	3-4	3-4	3-9
Number of room types	2	2	1
Number of meeting types	3	3	4

Table 6.1: Meetings used in experiments in this chapter

6.2 Differences between the baseline system and the new system

The baseline system described in Section 2.8 contains four phases: Speaker Activity Detection (SAD), Speaker Change Detection (SCD), clustering, and post processing. In the new system, new algorithms are proposed for all phases to improve system performance. In Chapter 3, 12 MFCCs + energy are used as acoustic feature vectors to detect speech/non-speech characteristics in the SAD phase, and 19 MFCCs + energy are used in the other parts. In this chapter, 19 MFCCs + energy feature vectors are used throughout all phases to maintain con-

sistency.

In the SAD phase, the baseline system applies a model-based speech detection method to remove the non-speech segments in the audio. All meetings in the development set are used to train the speech and non-speech models. A single Gaussian model is used as the non-speech model, and an eight-component GMM is used as the speech model. Meetings are segmented into small segments with lengths of 0.4 seconds; these segments are then clustered as speech or non-speech based on the GMM models. The number of components used in the GMMs and the segment length are determined by the analysis in Section 3.2.1.

By Experiment 3.1, we have seen that more components should be incorporated in the GMM when the NLR value is high. Furthermore, if the speech and non-speech segments from a meeting are used to train the GMM, better performance will be obtained (Experiment 3.2). Based on these conclusions, a new SAD algorithm is proposed. The new algorithm has two steps: the first step is the same as that used in the baseline system; in the second step, the detected speech and non-speech segments are used to adjust the GMMs. If the NLR is higher than a certain threshold, the number of components used in the non-speech GMM will be increased. Then, the new GMMs will be used to detect speech and non-speech segments. The performance of the new SAD algorithm will be discussed in 6.3.

In the SCD phase, the KL2-based speaker segmentation strategy is used in the baseline system as described in Section 2.8. The new SCD algorithm, which is based on FDA analysis, will be applied in the new system. Both algorithms were described in Chapter 4, as well as all values of the parameters. The threshold value of the new algorithm was determined in Section 4.2, and the threshold value of the baseline KL2-based algorithm was determined in Section 4.3. Their performance will be compared in Section 6.4.

In the clustering phase, the detected speech sections between speaker change

points produced by the SCD steps are then used to train the speaker models. The Gaussian model is used to initialise potential speaker models, such that each potential speaker model is trained by a speech section. These potential speaker models will then be clustered based on their similarity. In the baseline system, ΔBIC (defined in Equation 2.9) is used as the measurement of similarity. The pair of potential speaker models with the lowest ΔBIC values are merged into one, and a new GMM are trained on all the sections assigned to them. In the new GMM, the number of components is the sum of the model complexities of the two GMMs being merged. The merging process terminates when the remaining potential speaker number is below a certain threshold. Then, every speech section detected between speaker change points will be re-assigned to the remaining potential speaker model with the highest probability.

The post-processing phase includes three steps: UBM building, model adaptation and speaker clustering. In the baseline system, a GMM with 128 components is trained by all the speech in the meeting as the UBM. Then mean-only adaptation is used to derive the speaker models of all remaining potential speakers from the UBM. The CLR is used as the similarity measure between the UBM-adapted speaker models, and the pair of potential speaker models with the largest CLR value are merged. The whole process is terminated when the CLR between all the pairs of potential speakers is below a certain threshold. Again, all speech sections lying between detected speaker change points are re-assigned to the remaining potential speakers. Finally, the non-speech segments detected in the SAD, the speech sections and their corresponding speakers are output by the system as final results.

In the new system, a new model complexity decision criterion, EWPC, is applied to determine the model complexity of the potential speaker models and the UBM. The EWPC is described in Chapter 5; it controls the model complexity us-

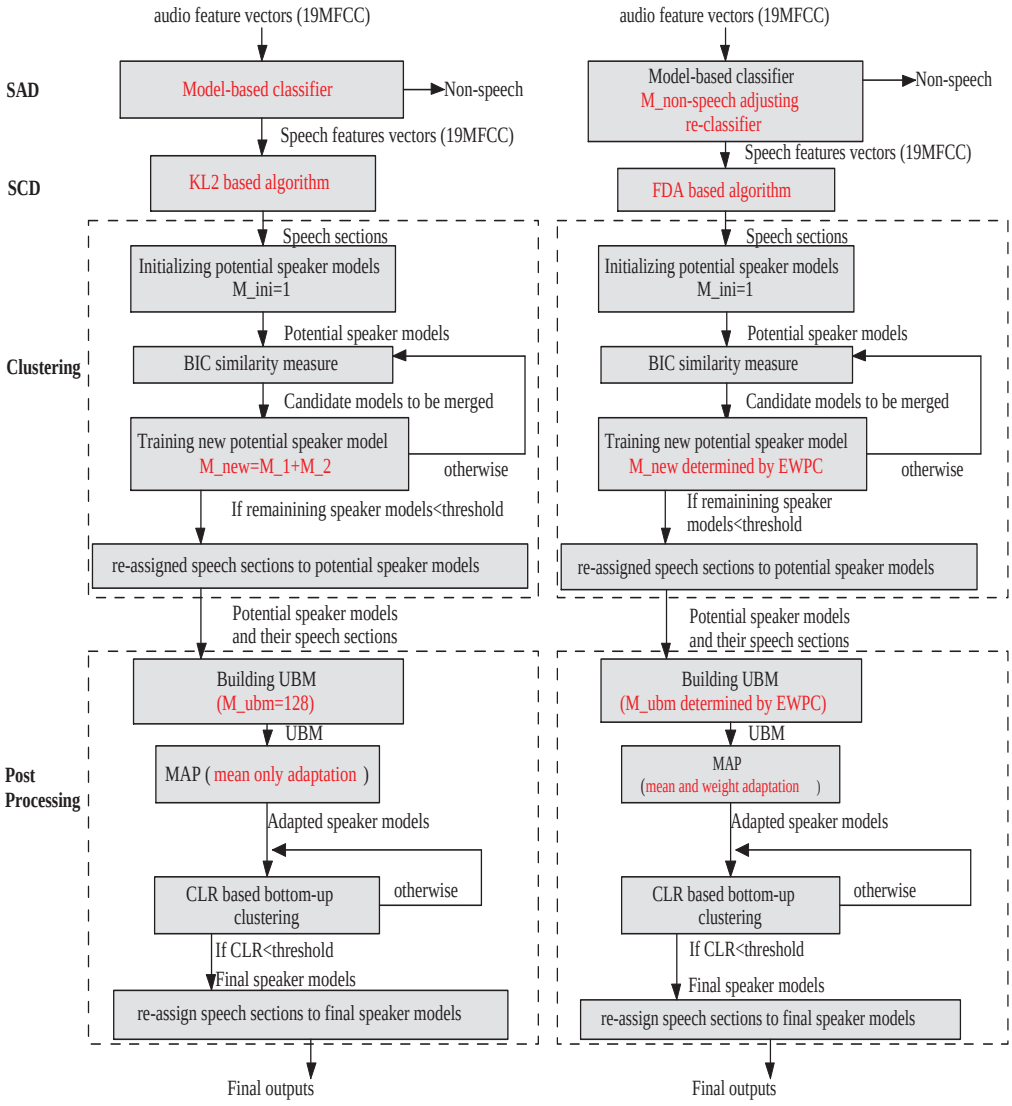
ing KL-divergence in such a way that fewer components are allowed in a potential speaker model, whereas more components are allowed in the UBM. The CLR similarity measure is used to select the candidate pair of potential speakers to be merged instead of ΔBIC. In addition, a mean and weight adaptation method (also described in Chapter 5) is applied to derive potential speaker models from the UBM. The performance of the new model complexity decision scheme and the new adaptation method will be given in Section 6.5.

In Figure 6.1, both the baseline system (Figure 6.1 (a)) and the new system (Figure 6.1 (b)) are described, with their differences highlighted in red. Other than the new algorithms labelled in red in Figure 6.1, a new termination scheme, which is based on the Normalized Cuts (NC), will be introduced in Section 6.6. The structure of the new system integrating the new termination scheme will also be illustrated in Section 6.6.

6.3 The performance of the new SAD algorithm

In this section, the new SAD algorithm will be described in detail, and its performance will be compared to the SAD process in the baseline system. The aim of the new algorithm is to integrate the information detected in the first round of speech/non-speech classification into a second round to improve the classification models.

First, we determine the model complexity of the non-speech GMM in the second round depending on the NLR detected in the first round. Based on the data analysis in Section 3.5, when the NLR is lower than 40%, the one-component GMM is sufficient to model the non-speech segments. When the NLR becomes higher, more components (2-3) should be included in the GMM. In Experiment 3.1, every recording extracted and used to determine the appropriate model com-



(a) The illustration of the baseline system. All of the parts that are different from their counterparts in the new system are highlighted in red. M represents the model complexity of the GMM. The suffix after the underline shows the types of the GMM, where ini denotes the initialisation models for each small section, and ubm refers to the UBM. M_new represents the model complexity of the new potential speaker model obtained by merging two models, whose model complexities are M_1 and M_2.

(b) The illustration of the new system. All of the parts that are different from their counterparts in the new system are highlighted in red. M represents the model complexity of the GMM. The suffix after the underline indicates the types of GMM, where non-speech refers to the $non-speech$ GMM.

Figure 6.1: The basline system, new system and their difference.

plexity for the non-speech GMM had the same length, which suggests that the NLR is proportional to the length of the noise. Therefore, the optimum model complexity value could be affected by either the NLR or the length of the noise. In the experiment described in this section, the meeting length varies from meeting to meeting, so the NLR and the length of the noise are no longer correlated. As a result, we must decide how to adjust the model complexity, whether according to the NLR or to the length of the noise. Because it can be observed in Figure 3.3 that the MISS error rate and the FA error rate change in opposite directions, it is better to adjust the model complexity based on the NLR, which represents the relationship between the non-speech length and the speech length. We here introduce the parameter β to control for the number of components used in the non-speech GMM: the model complexity is equal to the rounding value of $\beta * NLR$.

Second, we adjust the speech and non-speech models using the newly detected information. If the model complexity of the non-speech model does not need to be increased, it will be adapted towards the detected non-speech in the first round. Mean-only adaptation is applied to derive both the speech and the non-speech models in the second round. The mean values of the new models are updated following Equation 6.1.

$$\mu_i^{new} = \frac{\rho\mu_i^{old} + \sum_{j=1}^{N} \tau_{ji} x_j}{\rho + \sum_{j=1}^{N} \tau_{ji}} \tag{6.1}$$

where μ_i^{new} is the mean value of the ith component of the model after mean only adaptation, with μ_i^{old} as its counterpart before the adaptation. x_j is the jth feature vector of the detected speech (or non-speech); $\tau_{ji} = p(x_j|\mu_i^{old})$ is the post probability of x_j given the model before the adaptation; and ρ is the parameter to balance the influence of the training material and the newly detected information. Because we want the classification models to be adapted towards

the target meeting without losing the ability to cover a variety of sound types, the value of ρ is set to 0.5. If the non-speech model complexity must be increased in the second round, it will first be re-trained using the original training material with the redefined model complexity and then adapted towards the detected non-speech using the same adaptation strategy described in Equation 6.1.

The 15 meetings of the development set are used to determine the value of the parameter β for the new SAD algorithm. The variation of the speech/non-speech detection error rate as a function of β is illustrated in Figure 6.2. In Section 3.5, more components must be included in the non-speech GMM when the NLR is higher than 40%, which suggests that the value of β should be 5. However, the optimum value of β that minimises the sum of the E_{MISS} and E_{FA} is much larger than 5 when the value of β is determined by whole meetings. This may occur for two reasons: first, the meetings are much longer than the 10 minute audio segments used in Section 3.5, and therefore contain more non-speech; second, sometimes the non-speech segments detected in the first round are shorter than the actual non-speech segments in the meetings. It can be observed in Figure 6.2 that the sum of the E_{MISS} and E_{FA} achieves its minimum value when $\beta = 12$. Choosing 12 as the optimum value of β, the model complexity of the non-speech GMM is equal to 2 when 17% non-speech has been detected in the first round, and 3 when 25.5% non-speech has been detected. We set the minimum value of the non-speech model complexity to 1 and the maximum value to 5 in the new SAD algorithm.

To examine the performance of the new SAD algorithm, we substitute the new SAD algorithm for its counterpart in the baseline system and compare this new system to the baseline system. The baseline system is denoted as Sys_0 and the system using the new SAD algorithm is denoted as Sys_sad. The performance of both the baseline system and the system with the new SAD algorithm

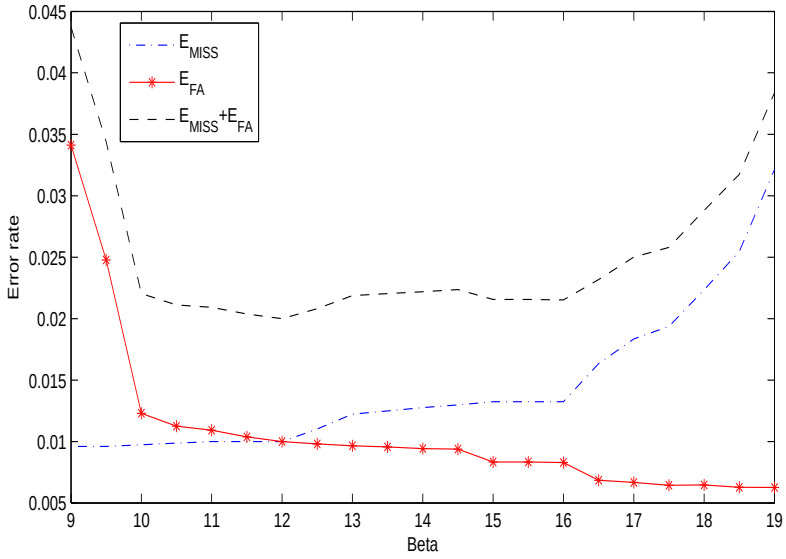


Figure 6.2: How E_{MISS} and E_{FA} vary with β

is displayed in Figure 6.3. E_{MISS}, E_{FA}, and E_{spkr}, the components of the DER, are shown in the three sub-figures of Figure 6.3, respectively. Figure 6.3(a) illustrates the mean value and the standard deviation of E_{MISS}. Using the new SAD algorithm leads to a decrease in the mean value of E_{MISS} in the evaluation set and an increase in the development set. Because the speech and non-speech of the development set have been included in the training material, to adjust the corresponding GMMs towards the detection information in the development set may not be as beneficial as in the evaluation set. Because the meetings from the ISL corpus in the evaluation set are more likely to have a different sound environment from those in the development set, the new algorithm shows the greatest improvement on them. Moreover, the new SAD algorithm allows more components to be included in the non-speech GMM of some meetings, which may increase the value of E_{MISS}, as shown in Figures 3.3 and 6.2.

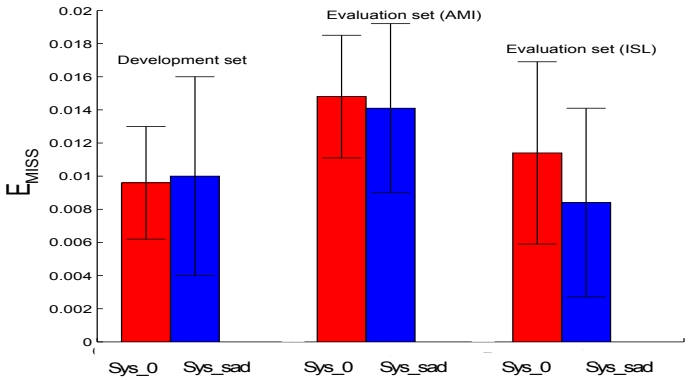
Figure 6.3(b) illustrates the mean value and the standard deviation of the E_{FA}. For the development set and the evaluation set from the AMI corpus, the

mean value of E_{FA} decreases and the standard deviation narrows. For the evaluation set from the ISL corpus, the decrease is not as clear. The meetings from the ISL obtain higher E_{FA} values, as detected both by the Sys_0 and Sys_sad. This may be because the transcription of the ISL corpus is manually produced and is inaccurate. Figure 6.3(c) illustrates the mean value and the standard deviation of the E_{spkr} . The mean value of the Espkr decreases slightly in the system with the new SAD algorithm. However, taking the standard deviation into consideration, the decrease in the E_{spkr} value is not fully supported. This is not a surprise because these errors are not directly caused by the speech /non-speech detection. However, the E_{spkr} value may be affected by the SAD step because if non-speech components are not completely removed from the meetings, they may contaminate the speaker models. The mean value of the E_{MISS}, E_{FA}, and E_{spkr} are shown in Table 6.2 for the development set and the evaluation set.

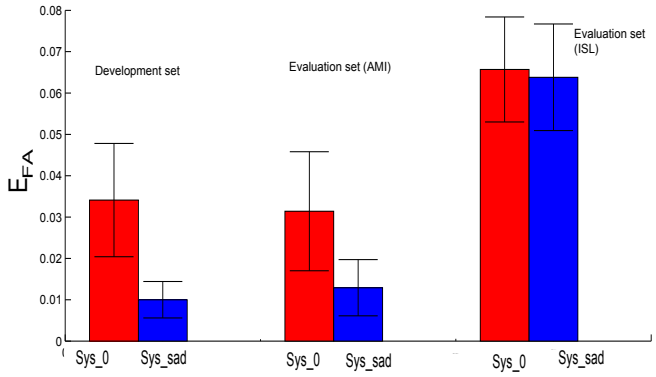
Meeting	SAD	E_{miss}(%)	E_{fa}(%)	E_{spkr}(%)	DER(%)
Development set	Sys_0	0.96%	3.41%	14.24%	18.61%
Development set	Sys_sad	1.00%	1.00%	13.82%	15.82%
Evaluation set (AMI)	Sys_0	1.48%	3.14%	14.27%	18.89%
Evaluation set (AMI)	Sys_sad	1.41%	1.29%	13.79%	16.49%
Evaluation set (ISL)	Sys_0	1.14%	6.57%	13.64%	21.35%
Evaluation set (ISL)	Sys_sad	0.84%	6.38%	13.22%	20.44%

Table 6.2: Performance of the baseline SAD algorithm and the new SAD algorithm

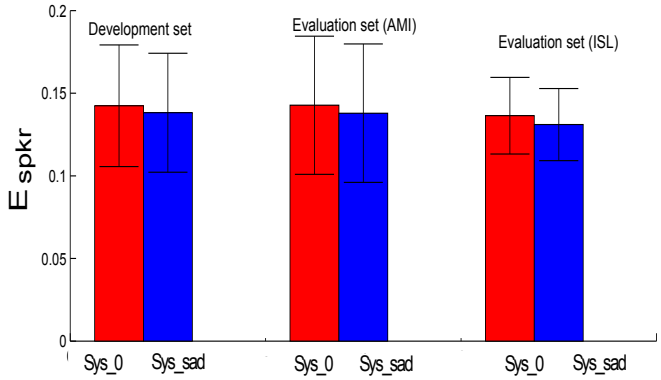
In Chapter 3, we concluded that increasing the model complexity of the non-speech GMMs when the NLR is high will decrease the E_{FA}, based on Experiment 3.1. Indeed, the mean value of the E_{FA} does show a decrease when the new SAD algorithm is applied. We display the E_{MISS} and E_{FA} of speaker diarization systems with different SAD algorithms in Figure 6.4, to show the influence of the NLR on the system performance. The E_{MISS} of the development set is shown in Figure 6.4(a), and the E_{FA} of the development set is shown in Figure



(a)E_{MISS} of the meetings. Bar: mean of the E_{MISS}; error bars: standard deviation of the E_{MISS}. "Sys_0": the baseline system; "Sys_sad": the system with new SAD algorithm



(b)E_{FA} of the meetings. Bar: mean of the E_{FA}; error bars: standard deviation of the E_{FA};



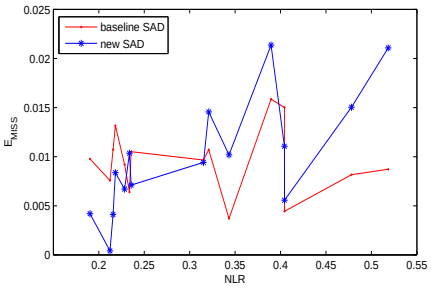
(c)E_{spkr} of the meetings. Bar: mean of the E_{spkr}; error bars: standard deviation of the E_{spkr};

Figure 6.3: The performance of the baseline SAD and the new system SAD.

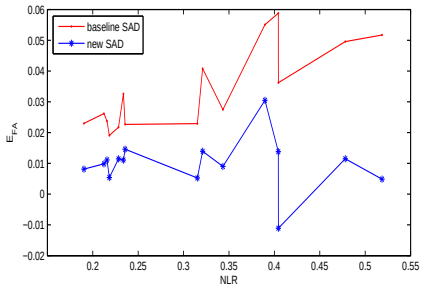
6.4(b). The new SAD algorithm obtains lower E_{FA} than the baseline SAD algorithm, and the difference between the performance of the baseline SAD and the new SAD increases when the NLR increases. Correspondingly, the E_{MISS} of some meetings increases, as predicted in Experiment 3.1. In Figure 6.4(c) and (d), similar results are observed, except that the increase of the E_{MISS} and the decrease of the E_{FA} are not always consistent with the NLR. This may be because the meetings with high NLR detected in the first round may not be the meetings with actual high NLR values. In Figure 6.4(e) and (f), no correlation could be observed between the error rate and the NLR. Again, this may be caused by the inaccurate transcription of the ISL meetings. Alternatively, as observed from Figure 6.4(e) and (f), the NLR of meetings in the ISL corpus is much lower than that of the meetings in the AMI corpus; therefore, the model complexities of the non-speech GMM of most meetings are unchanged.

6.4 The performance of the new SCD

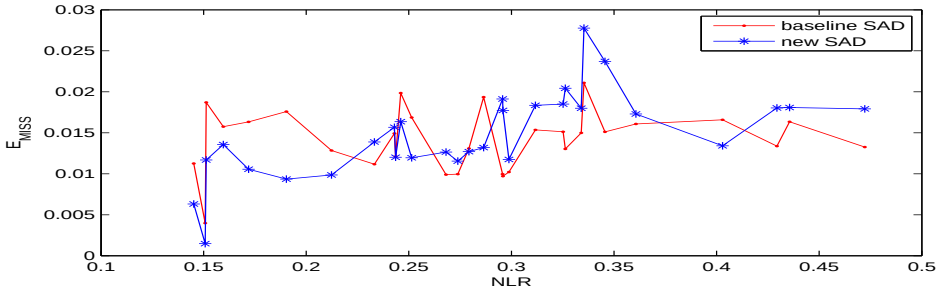
The new SCD algorithm has been described in detail in Chapter 4, and in this section, we examine its performance when integrated into the speaker diarization system. Two speaker diarization systems are used to compare the performance of the two SCD algorithms: (1) the baseline system with the new SAD analysed in Section 6.3 and the old SCD and (2) the baseline system with the new SAD and the new SCD. The first one is denoted as "Sys_sad" and the second one is denoted as "Sys_scd". By performing this comparison, the influence of different SAD algorithms can be removed. The performance is shown in Figure 6.5 and the specific values of the mean and standard deviation are displayed in Table 6.3. In Figure 6.5, it can be observed that the new SCD algorithm obtains the lower mean value of the DER in all sets of meetings. However, the standard deviation



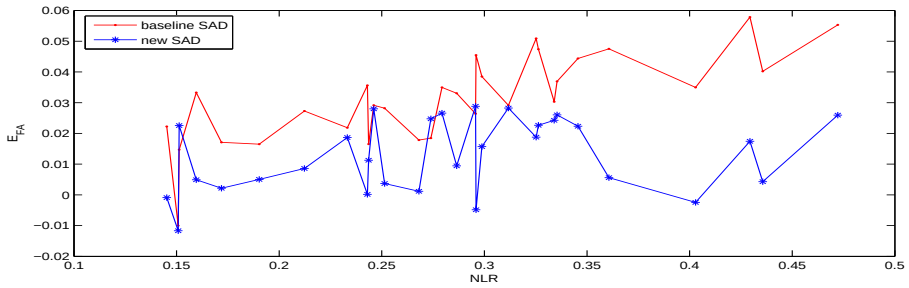
(a)E_{MISS} of the development set.



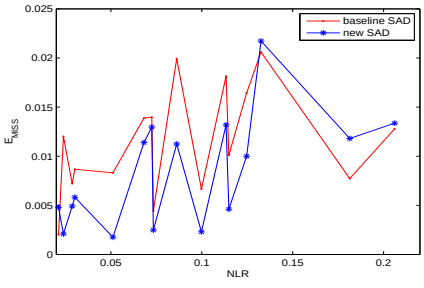
(b)E_{FA} of the development set.



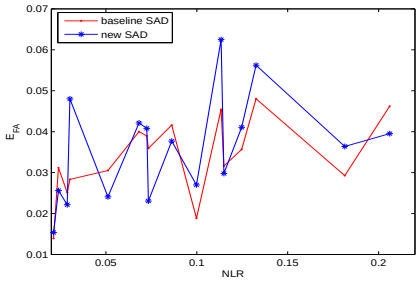
(c)E_{MISS} of the evaluation set from AMI corpus.



(d)E_{FA} of the evaluation set from AMI corpus.



(e)E_{MISS} of the evaluation set from ISL corpus.



(f)E_{FA} of the evaluation set from ISL corpus.

Figure 6.4: How E_{MISS} and E_{FA} changes with the NLR.

shows no difference. We observed in Section 4.3 that the new SCD algorithm misses fewer speaker changes and detects fewer false change points than the SCD algorithm. However, better performance in the SCD step cannot ensure an improvement of the entire system because the results of the SCD only serve as initialisation material for the potential speaker model training. An inefficient training method may affect the performance of the entire system. Although in the conclusion of Chapter 3 the new SCD algorithm is suggested to improve the system performance when there are more short turns, no evidence for this can be found in the experiment. The reason could be either that the performance of the new SCD algorithm is not connected to the number of short turns in a meeting or that the result of the SCD steps has a limited influence on the whole system.

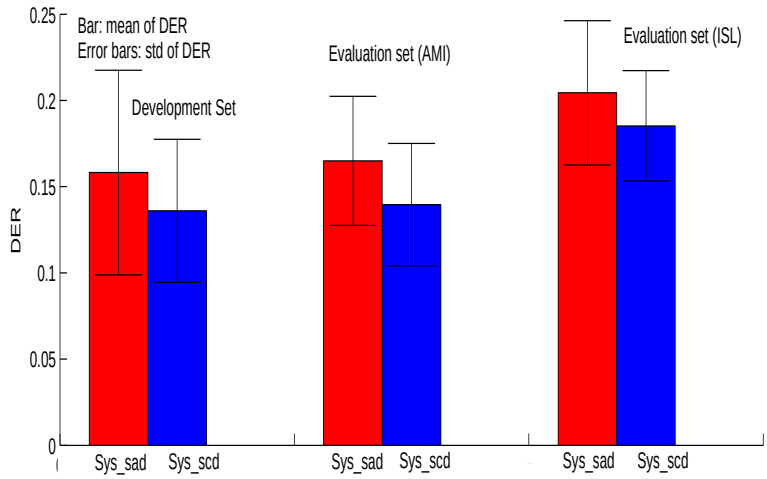


Figure 6.5: Performance of the baseline SCD algorithm and the new SCD algorithm

Meeting	SCD	E_{miss}(%)	E_{fa}(%)	E_{spkr}(%)	DER(%)
Development set	Sys_sad	1.00%	1.00%	13.82%	15.82%
Development set	Sys_scd	1.00%	1.00%	11.60%	13.60%
Evaluation set (AMI)	Sys_sad	1.41%	1.29%	13.79%	16.49%
Evaluation set (AMI)	Sys_scd	1.41%	1.29%	11.26%	13.96%
Evaluation set (ISL)	Sys_sad	0.84%	6.38%	13.22%	20.44%
Evaluation set (ISL)	Sys_scd	0.84%	6.38%	11.30%	18.52%

Table 6.3: Performance of the baseline SCD algorithm and the new SCD algorithm

6.5 The performance of the new model complexity selection algorithm and the mean adaptation method

In Chapter 5, three new algorithms were proposed to improve the model training in speaker diarization systems. First, a new criterion was developed to determine the model complexity in Section 5.2.2. Second, a new EM algorithm, with the model complexity selection scheme integrated into it, was introduced in Section 5.2.5. Third, a weight and mean adaptation method was described in Section 5.3. In this section, the performance of a combination of these three new algorithms will be analysed.

In the clustering step of the speaker diarization systems, the speech sections lying between speaker change points are merged according to the similarity of the potential speaker models trained by these sections. In the post-processing step, the UBM is trained, and these potential speaker models are adapted from the UBM. Therefore, it is essential to train efficient models with appropriate model complexity to ensure the success of the speaker diarization systems. If too many components are used to model small training sets, the model will suffer from over-fitting. However, with too few components to represent data characteristics, the model will fail to discriminate. In Section 5.2.2, a new model complexity de-

termination criterion, EWPC, was developed to determine the model complexity for both the potential speaker models and the UBM. The EWPC determines the model complexity for a GMM by selecting the one with highest penalised likelihood. The penalty term contains two parts, one part based on the parameter dimension and data size and the other based on the KL divergence between the prior and posterior distributions of the mixing parameter. The second part can be controlled by δ, which is the distribution parameter of the prior distribution of the mixing parameter w. The higher the value of δ, the more components will be included in the model. Therefore, by setting δ low ($\delta = 0.2$) for the potential speaker models and setting δ high ($\delta = 0.8$) for the UBM, fewer components could be included in the GMM for the speaker models to reduce the intra-speaker variance, whereas more components could preserved in the UBM to represent the inter-speaker variance.

A standard EM algorithm is usually used to train the GMMs. It cannot guarantee to achieve a local maximum, and it is sensitive to the initialisation of the parameters. To overcome the problem that the EM algorithm is sensitive to the initialisation parameter, the complexity-integrated EM algorithm proposed by Figueiredo (2002) that was introduced in Section 5.2.5 will be used to train the model. By integrating a model complexity penalty term into the EM algorithm, it initialises the EM training with a large number of components and then removes them if there is insufficient evidence to support their existence during the training.

In the post processing, the UBM is adopted to derive the speaker models. When the training data for a single speaker is insufficient, the speaker models derived from the UBM will capture more speaker characteristics and have a better presented structure. However, if the training data belonging to a speaker are too short, the resulting speaker model characteristics will be dominated by the

UBM and make it hard to discriminate it from other speakers. In Section 5.3, a new UBM adaptation algorithm was proposed, which adapted both the weight and the mean of the UBM. The components in the UBM that are not supported by the speaker data are removed, and their weights are re-assigned among the remaining components.

In the new speaker diarization system, after the SAD step and SCD step, potential speaker models will be trained using the detected speech sections between the speaker change points. The average speaker turn length is approximately 1.5 seconds, as displayed in Figure 3.5. In addition, the parameters of the SCD algorithm are adjusted to minimise the missed speaker changes, which will cause more false speaker changes to be detected and cause the average detected turn length to be less than the real average turn length. Therefore, most of the detected speech sections lying between speaker change points are short (shorter than two seconds), so a single full-covariance Gaussian model is trained for these sections. If long sections are detected (longer than 5 seconds), the model complexity M_{ini} is determined by Equation 6.2:

$$M_{ini} = round(N_s/100) \tag{6.2}$$

where N_s is the number of feature vectors in a speech section. A speech section with 100 feature vectors is equivalent to two seconds long. A Full-covariance Gaussian is used in the GMMs. When the potential speaker models have been initialised for all speech sections, the similarity between all pairs of models will be measured. CLR (defined by Equation 6.3) is used to measure the similarity.

$$CLR(X_1, X_2) = \frac{1}{n_1} \log \frac{p(X_1|\lambda_2)}{p(X_1|\lambda_1)} + \frac{1}{n_2} \log \frac{p(X_2|\lambda_1)}{p(X_2|\lambda_2)} \tag{6.3}$$

where X_1 and X_2 are speech sections assigned to a pair of potential speaker

models λ_1 and λ_2, respectively, and n_1 and n_2 are the number of feature vectors included in the sections. The pair of potential speaker models with the largest CLR are merged. The speech sections assigned to the original two potential speakers are assigned to the new potential speaker model, and the model will be retrained by these speech sections. Using the combination the two original potential speaker models as the initialisation, with half the weight value of the components, the new model is retrained using the new EM algorithm. The new EM algorithm automatically removes the extra components in the GMM. The remaining model complexity of the GMM is reduced by one, and the training process is repeated until the remaining model complexity is less than any of the original two potential speaker models. Then, the GMM with the model complexity that minimises the EWPC is chosen as the new potential speaker model (with $\delta = 0.2$). The merging process terminates when the number of remaining potential speakers is less than a given threshold.

Because, in post processing, the potential speaker models will be updated by adaptation from the UBM, it is better to begin when the data assigned to every cluster are long enough to support the adaptation. During the adaptation, each component in the UBM will be adjusted towards a particular potential speaker. If there are not enough data assigned to a cluster, its adapted model will be dominated by the characteristics of the UBM instead of its own characteristics. As a result, the post-processing should begin when all of the clusters have enough data.

Meetings have different numbers of speakers, and their utterances are of different lengths. Moreover, some meetings have one or several dominant speakers so that the other speakers occupy only a small proportion of the overall audio stream. Hence, it is difficult to decide when there are enough data in a cluster to start the adaptation. However, the number of potential speakers remaining in

the process can be used to start the post-processing. Because in the AMI corpus the range of speaker numbers is 3-4 and in the ISL corpus the range of speaker numbers is 3-8, we choose 20 as the threshold for the remaining number of potential speakers to enter the post-processing step. This value is larger than the actual number of speakers in the meetings and at the same time leaves room for adjusting the speaker model. In post-processing, the UBM is trained by a method similar to the way the potential speaker models are trained, except that the UBM is initialised by a combination of all of the potential speaker models, with the weight value averaged over all of them, and the lowest model complexity of the UBM is the upper limit of the model complexity of all individual potential speaker models. Then, the entire potential speaker models need to be re-trained by adapting the UBM towards the speech sections assigned to each potential speaker, using the weight and mean adaptation method. Then, the similarity between each pair of new potential speaker models, which is adapted from the UBM, will be measured by a slightly changed version of the CLR, defined in Equation 6.4:

$$CLR(X_1, X_2) = \frac{1}{n_1} \log \frac{p(X_2|\lambda_1)}{p(X_1|\lambda_{ubm})} + \frac{1}{n_2} \log \frac{p(X_1|\lambda_2)}{p(X_2|\lambda_{ubm})} \qquad (6.4)$$

where λ_{ubm} represents the UBM model. The pair of the potential speakers with the largest CLR will be merged, and the new potential speaker model will be re-adapted from the UBM, using our new weight and mean adaptation. The merging process will terminate when the CLR measurements between all pairs of the remaining potential speakers are below a certain threshold.

In the baseline system, a simple model complexity selection scheme and a common EM algorithm are applied. All of the potential speaker models are initialised as a single Gaussian model. The model complexity of the new potential

speaker model is the sum of the original two models, and the UBM has 128 full covariance components. To compare the three new algorithms proposed in Chapter 5, we compare the new system illustrated in Figure 6.1 with a revised baseline system, whose original SAD and SCD steps are replaced by the new SAD and SCD algorithms explained in Section 6.3 and 6.4, respectively. In the experiment in this section, the revised version of the baseline system will be denoted "Sys_0".

The three new algorithms proposed in Chapter 5 are related to each other. The new EM algorithm integrates the parameter dimension and data size part of the penalty term into the EWPC and can automatically remove extra components from the GMM. This will accelerate the model complexity selection process, which will pick the GMM with the lowest EWPC value from among all possible model complexity values. The component-removing scheme is integrated into the weight and mean adaptation strategy in a similar way as it was integrated into the EM algorithm. Therefore, instead of evaluating the individual performance of the three new algorithms, we check the performance of their combination. As a result, the entire new system, which will be referred to as "Sys_{new}" in the rest of the section, will be applied in the performance analysis in the section.

In addition to our new algorithms, many other model training algorithms and model complexity selection criteria have been implemented to improve the performance of the speaker diarization system. In (Anguera et al., 2006a), the number of components used in the potential speaker models is correlated with the quantity of training data. The CCR will be used to decide the initial number of components used for each potential speaker model. The number of components used in a cluster is defined by Equation 6.5:

$$M_{ini} = round(\frac{N_j}{CCR}) \tag{6.5}$$

where N_j is the number of features. CCR=7 is the value recommended in (Miro,

2006). This algorithm is called the CCR model selection criterion in the rest of this section.

An incremental method to train the GMM is described in the HTK toolkit (Young et al., 2005). In this method, a Gaussian model is constructed for the whole training data set. The Gaussian model is then split into two, and the GMMs are trained. The splitting process continues until the given model complexity has been achieved. In this way, the position of each component will be better modelled. This algorithm is called "incremental training" in the rest of this section.

Cross validation EM (CVEM) is an algorithm to adjust the positions of a fixed number of components (Anguera et al., 2007). During the EM training, the feature vectors are split into P parallel partitions, and a GMM is trained on each partition of the data. In the E-step, the expected conditional probability of the hidden variables of all of the GMMs will be calculated based on their corresponding partitions. Then, in the M-step, for each GMM, the data and hidden parameter of other partitions will be used to cross maximise the GMMs. The parameter P of the CVEM is recommended to be set to 35 in (Miro, 2006). This algorithm is called "CVEM training" in the rest of this section.

To compare the performance of these model complexity selection algorithm and model training algorithms to our new algorithm, two more speaker diarization systems are built. All of them using the new SAD step and SCD as S_0 and S_{new}. CCR model selection criterion is used in both systems to determine the model complexity. Incremental training is applied to one of the systems for model training of both potential speaker models and the UBM. CVEM training is used in another system, only for the UBM training since the short speech sections are not suitable to be split into many partitions. The system with CCR model selection criterion and incremental training will be referred to as "Sys_1"

in the rest of the section, and the system with CCR model selection criterion and CVEM training will be referred to as "Sys_2". The other part of the Sys_1 and Sys_2 are the same as the baseline system.

The performance of all Sys_0, Sys_1, Sys_2 and Sys_{new}, is displayed in Figure 6.6. It can be observed that for the development set, both the mean value of DER and the standard deviation are lower in Sys_{new} than in Sys_0, Sys_1, Sys_2. For the two evaluation sets, the mean DER of Sys_{new} decreases, but the standard deviation is slightly higher. This may be due to the fact that the new algorithms are more sensitive to the threshold of the CLR. In all other systems, the model complexity of the adapted potential speaker models is fixed to 128, which is also the fixed model complexity of the UBM. In the Sys_{new} system, however, the model complexity of the UBM changes from meeting to meeting, and the model complexity of the adapted potential speaker models is also non-fixed. The higher flexibility of the new algorithm makes it more likely to be affected by the value of the parameter. In the other two systems, Sys_1 and Sys_2, the mean value of the DER shows no obvious reduction. This may be due to the complexity selection criterion of the CCR model. Determining the model complexity based on the quantity of training data may cause excessive components to be included in the GMMs, especially when the speech sections assigned to it are long. The standard deviation of the Sys_2's DER is wide. This may be because the CVEM model training algorithm is sensitive to the parameter P, which is the number of split partitions used to train the UBM. The specific mean values of the DER of these speaker diarization systems are shown in Table 6.4.

In Chapter 3, it was suggested that the new model complexity selection criterion copes better when the speech length in a meeting is longer or the number of speakers is higher. Therefore, we show how the speech length and the number of speakers affect the DER of the Sys_0 and Sys_{new} in Figure 6.7 and 6.8.

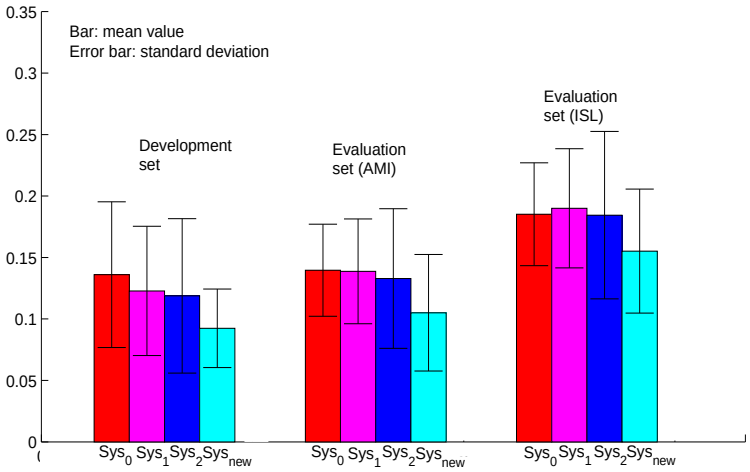


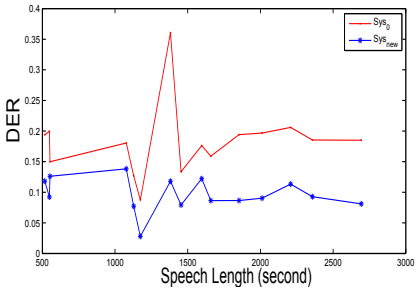
Figure 6.6: Performance of the speaker diarization systems Sys_0, Sys_1, Sys_2, and Sys_{new}.

Meeting	System	E_{miss}(%)	E_{fa}(%)	E_{spkr}(%)	DER(%)
Development set	Sys_0	1.00%	1.00%	11.60%	13.60%
Development set	Sys_1	1.00%	1.00%	10.28%	12.28%
Development set	Sys_2	1.00%	1.00%	9.88%	11.88%
Development set	Sys_{new}	1.00%	1.00%	7.24%	9.24%
Evaluation set (AMI)	Sys_0	1.41%	1.29%	11.26%	13.96%
Evaluation set (AMI)	Sys_1	1.41%	1.29%	11.17%	13.87%
Evaluation set (AMI)	Sys_2	1.41%	1.29%	10.59%	13.29%
Evaluation set (AMI)	Sys_{new}	1.41%	1.29%	7.80%	10.50%
Evaluation set (ISL)	Sys_0	0.84%	6.38%	11.30%	18.52%
Evaluation set (ISL)	Sys_1	0.84%	6.38%	11.58%	19.00%
Evaluation set (ISL)	Sys_2	0.84%	6.38%	11.22%	18.44%
Evaluation set (ISL)	Sys_{new}	0.84%	6.38%	8.30%	15.48%

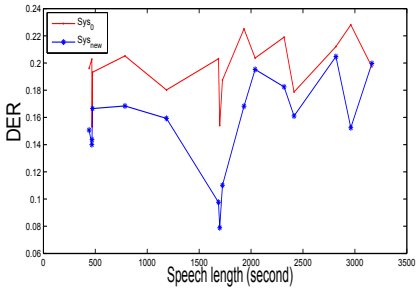
Table 6.4: Performance of the speaker diarization system Sys_0, Sys_1, Sys_2, and Sys_{new}

Figure 6.7 (a), (b) and (c) shows the DER of the meetings in all data sets as functions of the speech. For almost all of the meetings, Sys_{new} obtains the best performance. However, in contrast to the assumption in Chapter 3, there is no evidence that the new system has a greater advantage when dealing with long meetings with long speech lengths. In Chapter 3, it is observed that feature vectors from different speakers split the feature space into many small sub-spaces when the speech length is higher; therefore, the assumption that using more components in the UBM to model the inter-speaker variability will improve the system performance was made. However, a UBM that is more capable of modelling the inter-speaker variability does not necessarily lead to more accurate potential speaker models. Moreover, the sensitivity to the CLR threshold may also have an effect on the outcomes, since the trend that the difference between the Sys_0 and Sys_{new} increases can be observed in the development set.

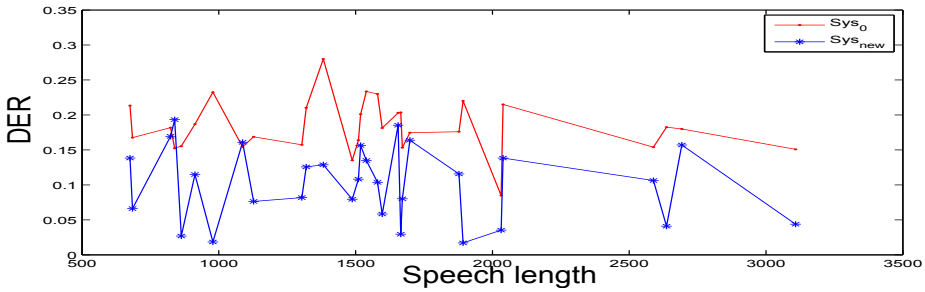
In Figure 6.8(a), the DER values of the meetings from the AMI corpus are displayed against the number of speakers. In either case, the DER decreases when Sys_{new} is used. The decrease is clearer when the number of speakers is 3. What is worth noting is that when Sys_0 is used, the system performs worse when the speaker number is 3, and when Sys_{new} is used, the system performs worse when the speaker number is 4. Among all of the meetings, only five of them have three speakers, and all of the others contain four speakers. It is hard to tell whether the difference in the DER is due to the individual cases. Figure 6.8(b) displays the DER of the meetings from the ISL corpus. It can be observed that Sys_{new} performs better when the speaker number is higher.



(a)DER of the development set.

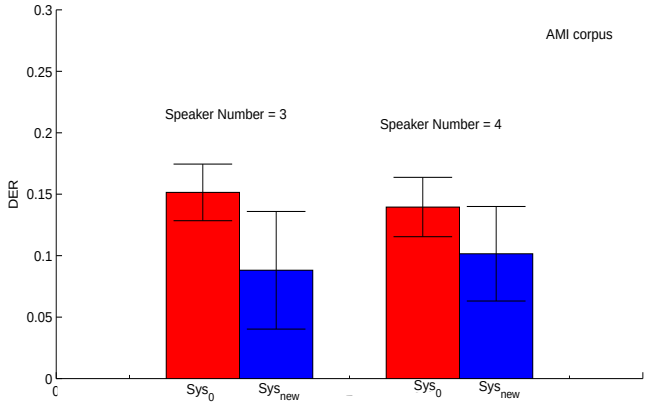


(b)DER of the evaluation set from ISL corpus.

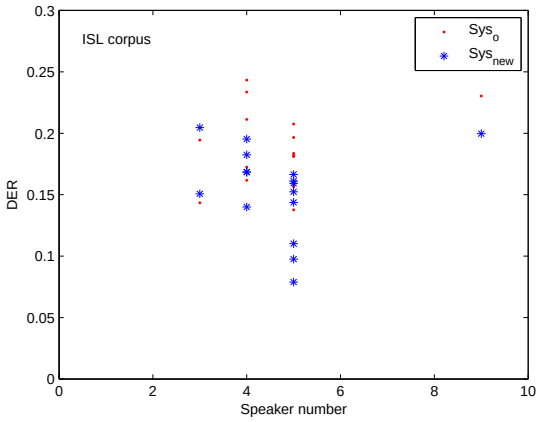


(c)DER of the evaluation set from AMI corpus.

Figure 6.7: How DER changes with the Speech length.



(a)DER of the AMI corpus.



(b)DER of the ISL corpus.

Figure 6.8: Variation of DER with the speaker number.

6.6 Normalized Cuts applied to clustering

In Figure 6.8, it can be observed that the standard deviation of the DER from the evaluation set is high. This may be because the performance of the new system depends strongly on the value of the CLR threshold to determine when to terminate the merging process of the potential speakers. Moreover, the CLR threshold based termination strategy is a local solution rather than a global solution. In the new system, the timing to end the process is based on the similarity between the closest pair of potential speaker models, regardless of the overall similarity between all of the potential speaker models. The choice of a wrong time to terminate the merging will not only cause speech sections to be wrongly assigned but also lead to errors in the estimation of the speaker number. Therefore, in this section I will develop a new method to terminate the potential speaker merging without a threshold, taking the similarity among all potential speaker models into consideration.

The mean values of the components in the potential speaker GMMs that were adapted from the UBM are thought to be a reliable representation of speaker characteristics (Tsai et al., 2005) (Tsai et al., 2007). The potential speaker models created by mean-only adaptation have the same number of components and during the adaptation process, all of the components in the GMM are forced to follow the order of the UBM. Therefore, by conjoining all the mean vectors of the components in the GMM one by one, a large feature vector is formed for each cluster (dimension of the acoustic features * number of components in the GMM). The normalized inner product can be used to measure the similarity between these super-vectors. The normalized inner product of two vectors v_i and

v_j will be defined as in Equation 6.6:

$$S(v_i, v_j) = \frac{< v_i, v_j >}{||v_j||||v_j||} \tag{6.6}$$

Using the inner-product of the super-vectors to measure the similarity between potential speaker models, a merging process termination scheme based on the ratio of the intra-speaker variability and inter-speaker variability can be developed using Normalized Cuts (NC). NC ((Shi and Malik, 2000)) was first proposed for two-class graph partitions, which measure the normalized dissimilarity between two disjoint sets. Assuming that two data sets A and B satisfy the conditions $A \bigcup B = V$ and $A \bigcap B = \oslash$, their dissimilarity can be measured by $Ncut(A, B)$:

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \tag{6.7}$$

where $cut(A, B)$ is the total dissimilarity from A to B and $assoc(A, V)$ is the total connection from A to V. Assume that dw_{ij} denotes the dissimilarity between v_i and v_j; then $cut(A, B)$ and $assoc(A, V)$) are given by Equations 6.8 and 6.9:

$$cut(A, B) = \sum_{i \in A, j \in B} dw_{ij} \tag{6.8}$$

$$assoc(A, V) = \sum_{i \in A, j \in V} dw_{ij} \tag{6.9}$$

where a lower value of dw_{ij} indicates a greater distance between i and j. Because $cut(A, B) = assoc(A, V) - assoc(A, A)$, $Ncut$ is directly proportional to the total inter-class dissimilarity and inversely proportional to the total intra-class

dissimilarity. Extending the definition of the NC to the multi-class situation:

$$Ncut_k = \frac{cut(A_1, V - A_1)}{assoc(A_1, V)} + \frac{cut(A_2, V - A_2)}{assoc(A_2, V)} + \cdots \frac{cut(A_k, V - A_k)}{assoc(A_k, V)} \tag{6.10}$$

where $A_1 \cdots A_k$ are disjoint sets and $A_1 \bigcup A_2 \bigcup \cdots \bigcup A_k = V$ and k is the number of the remaining potential speakers.

$Ncut_k$ can be used to select the appropriate number of speakers. Because there are always fewer than ten people attending a meeting, the merging process will run without stopping to check until ten clusters are left. After each merging step, the $Ncut_k$ value should be calculated, until there is only one remaining cluster. The partition whose $Ncut_k$ achieves the minimum value will be selected by the system as the final result. Because the dissimilarity measure dw_{ij} is proportional to the distance between i j, the inverse value of $S(v_i, v_j)$ will be used in Equation 6.8 and 6.9 as a dissimilarity measure. Because the potential speaker models adapted from the UBM by weight and mean adaptation have different numbers of components, the super-vectors of these models have different dimensions and $S(v_i, v_j)$ is not computable. Therefore, mean only adaptation will be applied to adapt the potential speaker models from the UBM so that the super-vectors of all potential speaker models have the same model complexity and the inner product of these super-vectors is computable.

Using the NC to terminate the potential speaker merging process, a detailed structure of the new system, which is labelled as "Sys_{new2}", is illustrated in Figure 6.10. The parts of Sys_{new2} that are different from Sys_{new} are labelled in red. The performance of Sys_{new2} will be illustrated in Figure 6.9, compared to the new system, Sys_{new}. It can be observed from Figure 6.9 that the standard deviation of the Sys_{new2} narrows on the evaluation set from the AMI corpus compared to Sys_{new}, without a dramatic change in the mean value of the DER of Sys_{new}.

This may be because when replacing the weight and mean adaptation with the mean only adaptation for the potential speaker adaptation from the UBM, the accuracy of the potential speaker models decreases. However, using the NC-based merging termination scheme, which has no threshold value to adjust to construct global optimum solution, will improve the steadiness of the speaker diarization system. For the evaluation set from the ISL, the mean value of the DER decreases, and the standard deviation narrows. Because the meetings from the ISL have a wider range of speaker numbers, terminating the merging process at the right time is more essential to the system performance for these meetings. The specific mean values of the DER for Sys_{new} and Sys_{new2} are listed in Table 6.5.

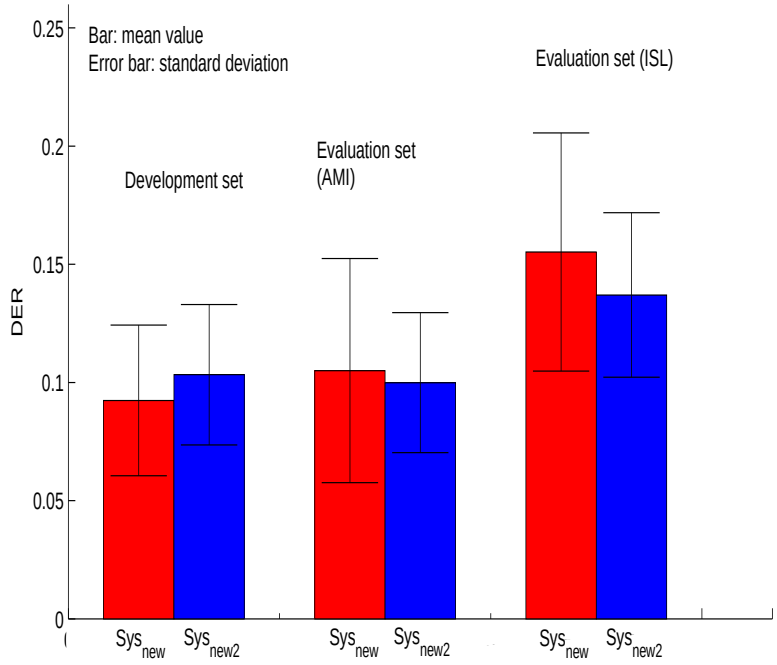


Figure 6.9: The performance of Sys_{new} compared to Sys_{new2}.

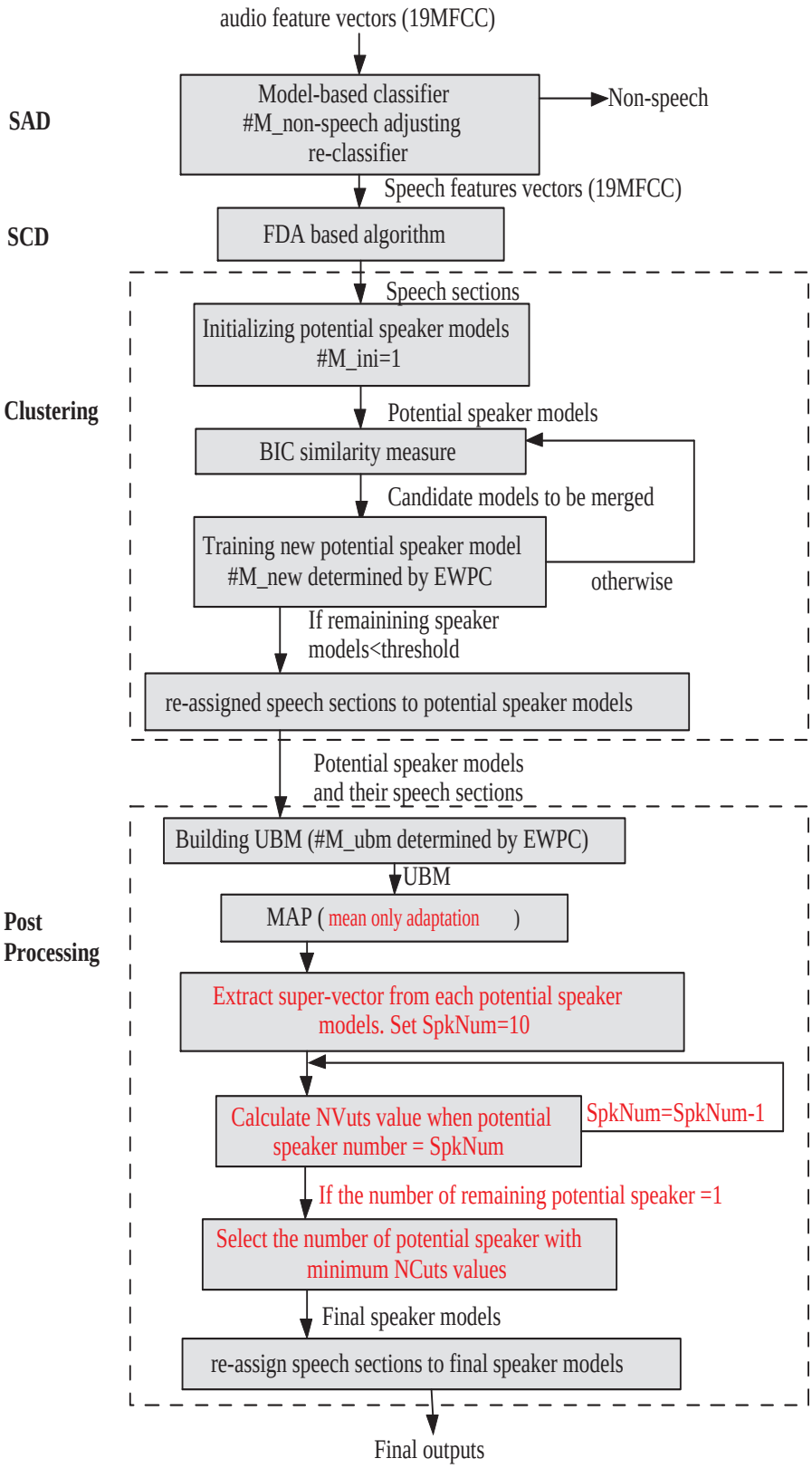


Figure 6.10: The structure of New System Sys_{new2}.

Meeting	System	E_{miss}(%)	E_{fa}(%)	E_{spkr}(%)	DER(%)
Development set	Sys_{new}	1.00%	1.00%	7.24%	9.24%
Development set	Sys_{new2}	1.00%	1.00%	8.33%	10.33%
Evaluation set (AMI)	Sys_{new}	1.41%	1.29%	7.80%	10.50%
Evaluation set (AMI)	Sys_{new2}	1.41%	1.29%	7.29%	9.99%
Evaluation set (ISL)	Sys_{new}	0.84%	6.38%	8.30%	15.48%
Evaluation set (ISL)	Sys_{new2}	0.84%	6.38%	6.48%	13.70%

Table 6.5: Performance of the NC-based merging termination scheme

6.7 Overall Experiments and Analysis of Results

In this chapter, the new algorithms derived in earlier chapters have been integrated into the new system and their performance presented and discussed. Here I summarise these new algorithms as follows:

1. a new model-based SAD algorithm that contains two rounds and the speech and non-speech models in the second round will be adjusted according to the detected information from the first round;
2. a new SCD algorithm that is based on the FDA analysis;
3. a new model complexity selection criterion, the EWPC, that allocates lower model complexity for potential speaker models to reduce the influence of intra-speaker variability, while allocating higher model complexity for the UBM to capture more inter-speaker variability;
4. a new EM algorithm that integrates the data size penalty term of the EWPC to accelerate the removal of extra components in the GMM automatically in the training process;
5. a weight and mean adaptation algorithm to adapt models from the UBM for potential speakers;
6. a new NC-based merging process termination scheme to decide when the remaining potential speakers will be output by the system as final results.

Among all of the new algorithms, the performance of (1), (2) and (6) is compared, respectively, with the new algorithms. The algorithms (3)-(5) are combined in application and their performance is analysed together. The performance levels of all new algorithms are displayed in a row in Figure 6.11 and compared with the baseline system. In Figure 6.11, the baseline system is denoted Sys_{old}. The system that is similar to the baseline system, except for its application of the new SAD algorithm instead of its counterpart in the baseline system, is denoted Sys_{sad}. The system that uses the new SAD algorithm in the SAD step, the new SCD algorithm in the SCD step, and keeps the other algorithms the same as those of the baseline system is called Sys_{scd} in Figure 6.11. The new systems described in Figure 6.10 and in Figure 6.1 are referred to as Sys_{new} and Sys_{new2}, respectively. In Figure 6.11, it can be observed that each new algorithm improves the performance of the speaker diarization systems by decreasing the mean of the DER, except when using the NC-based merging termination scheme on the development set. In addition, the standard deviation of the systems' performance is wide, except when integrating the NC-based merging termination scheme in all datasets.

It has been stated in Section 6.3 that using the new SAD algorithm will decrease the value of E_{FA}, especially when the NLR of the meetings is high. There are some exceptions because the disproportionateness of the NLR and the non-speech length in the meetings may reduce the efficiency of the new algorithm, or the NLR of the detected non-speech in the first round may not be consistent with the NLR of the whole meeting. As explained in Section 6.4, because the standard deviation is wide, it is difficult to measure the efficiency of the new SCD algorithm. However, because the SCD step is an early step in the whole system, the advantages of the new SCD algorithm may accumulate, and better performance is observed when it is combined with the new algorithms (3)-(5). Integrating the

new algorithms (1)-(5) into the baseline speaker diarization system, we obtain the new system described in Section 6.2. The combination of the new algorithms (3)-(5) provides the largest contribution to the performance of the system, compared to the prior system in Figure 6.11. However, the new system is sensitive to the CLR threshold because the success of the new system heavily depends on the model accuracy, as discussed in Section 6.5. In Chapter 3, better performance is expected when using new model complexity selection criterion, especially when the speech length is long and the speaker number is high. According to the experimental results, no evidence supports the assertion that the new system has a greater advantage when dealing with meetings with long speech lengths. For the evaluation set from the ISL corpus, the new algorithms (3)-(5) improve the system performance when the speaker number becomes higher. However, the same results cannot be found for the evaluation set from the AMI.

In Section 6.6, a new NC-based potential speaker merging termination scheme is developed. This new scheme is without threshold and makes the decision based on the global information. The new speaker diarization system (illustrated in Figure 6.10) that includes the NC-based termination scheme is steadier. In Table 6.6, the mean values of the DER of all systems and the system improvements are listed. "Improvement vs prior" measures the decrease in the DER of each system compared to its prior system, in the order shown in Figure 6.11. 'Improvement vs baseline' measures the decrease in the DER of each system compared to the baseline system. The DER of all meetings obtained using Sys_{old}, Sys_{sad}, Sys_{scd}, Sys_{new} and Sys_{new2} are specified in Appendix A.

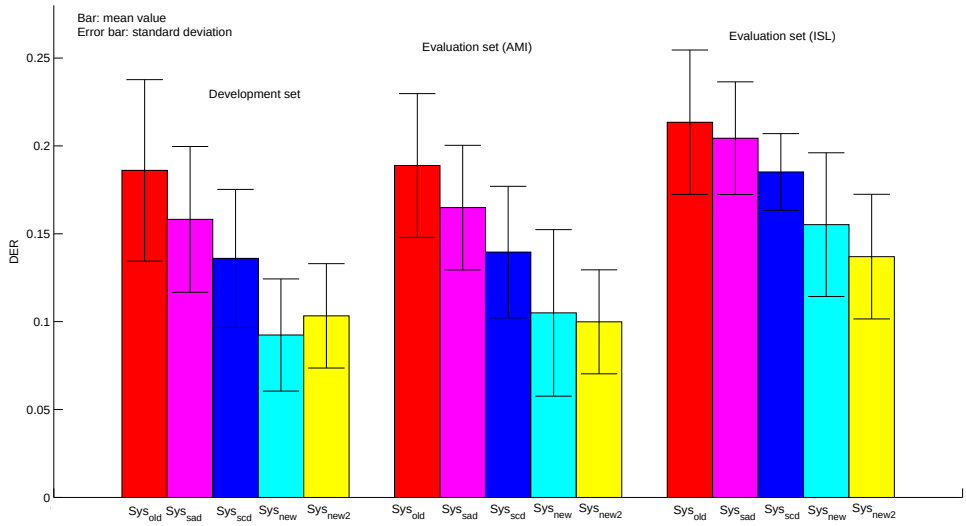


Figure 6.11: The performance of all systems

Meeting	System	DER(%)	Improvement vs prior	Improvement vs baseline
Development set	Sys_{old}	18.61%	0%	0%
Development set	Sys_{sad}	15.82%	2.79%	2.79%
Development set	Sys_{scd}	13.60%	2.22%	5.01%
Development set	Sys_{new}	9.24%	4.36%	9.37%
Development set	Sys_{new2}	10.33%	-0.17%	9.20%
Evaluation set (AMI)	Sys_{old}	18.89%	0%	0%
Evaluation set (AMI)	Sys_{sad}	16.49%	2.40%	2.40%
Evaluation set (AMI)	Sys_{scd}	13.96%	2.53%	4.93%
Evaluation set (AMI)	Sys_{new}	10.50%	3.46%	8.39%
Evaluation set (AMI)	Sys_{new2}	9.99%	0.51%	8.90%
Evaluation set (ISL)	Sys_{old}	21.35%	0%	0%
Evaluation set (ISL)	Sys_{sad}	20.44%	0.91%	0.91%
Evaluation set (ISL)	Sys_{scd}	18.52%	1.92%	2.83%
Evaluation set (ISL)	Sys_{new}	15.48%	3.04%	5.87%
Evaluation set (ISL)	Sys_{new2}	13.70%	1.18%	7.01%

Table 6.6: Summary of average DER for all new algorithms

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, I have investigated the shortcomings of the existing speaker diarization systems and examined the meeting characteristics that may cause these problems by focusing on the SAD, SCD and the construction of the UBM steps of the speaker diarization system (Chapter 3). Based on the problems discovered in Chapter 3, four new technologies for speaker diarization processing, including an SAD algorithm, a change point detector, a model complexity criterion and a weight and mean model adaptation technique, were investigated in this thesis. Those technologies significantly improve the performance of the speaker diarization system, especially when combined. In addition, the new EM algorithm proposed in (Figueiredo and Jain, 2002) was introduced to accelerate the training of the model, and the NC (Shi and Malik, 2000) was introduced to determine when to terminate the potential speaker merging process. Although these algorithms were not developed in this thesis, this is the first time that they have been applied to speaker diarization. The performance of these new algorithms was examined and compared to the baseline system in Chapter 6. The detailed conclusions of each step and performance of new systems are summarised as

follows:

SAD: It was discovered that more components should be incorporated for better performance when the NLR value is higher. Moreover, the performance of the SAD process improves if the audio material used to train the speech/non-speech GMM and the test audio material used to test the performance of the GMMs are from the same meeting. Based on these observations, a new SAD algorithm was proposed in Section 6.3 Compared to the SAD algorithm in the baseline system, the new algorithm reduces both the E_{MISS} and the E_{FA} values, especially when the E_{FA} of the NLR is high. When the new SAD algorithm was employed to replace its counterpart in the baseline system, the mean value of E_{MISS} was increased from 0.96% to 1.00% percentage points, and the mean value of E_{FA} was reduced from 3.41% to 1.00% in the development set. The same trend was observed in the evaluation set from the AMI corpus, where the mean value of E_{MISS} decreased from 1.48% to 1.41% and the mean value of E_{FA} decreased from 3.14% to 1.29%. For the evaluation set from the ISL, the mean value of E_{MISS} decreased from 1.14% to 6.57%, and the mean value of the E_{FA} decreased from 6.57% to 6.38%. The mean value of the DER decreased 18.61% to 15.82%, from 18.89% to 16.49% and 21.35% to 20.44%, respectively, for the three datasets.

SCD: FDA-based measurements were introduced to examine the overlap between the data distributions of a pair of short segments. It was discovered that the FDR, the error rate of the FDC, and the average distance from errors to the FDC are all capable of determining whether a pair of short segments is from different speakers or the same speaker. In Chapter 4, a new speaker change detection algorithm was developed based on the combination of various measurements of the FDA. Compared to the SCD algorithm in the baseline system, the new algorithm minimises the missing change error rate, while at the same time reducing

the false change error rate and narrowing the standard deviation of the two types of errors. In Section 6.4, the speaker diarization system with the new SCD algorithm is compared to the speaker diarization system with the baseline SCD algorithm. When using the new SCD algorithm, a decrease of the mean of the DER is observed. In Section 3.4, I concluded that in the acoustic feature space, inter-speaker variability is intertwined with phonetic variability; as a result, features from different speakers split the feature space into many small sub-spaces. In the development set, when both the new SAD algorithm and the new SCD algorithm were employed in the baseline system, a reduction from 15.82% to 13.60% in the mean of the DER was observed compared to the system with only the new SAD algorithm. The decrease in the mean of the DER was 16.49% to 13.96% for the evaluation set from the AMI corpus and from 20.44% to 18.52% for the evaluation set from the ISL corpus.

Model Training: Depending on the analysis in Chapter 3, in the acoustic feature space, the inter-speaker variability is intertwined with the phonetic variability; as a result, features from different speakers split the feature space into many small sub-spaces. The number of sub-spaces tends to increase with the length of the speech and the number of speakers in a target meeting. A new model complexity criterion was proposed in Chapter 5. By setting the parameter δ to different values, the new criterion could reduce the model complexity to reduce intra-speaker variability and allow more model complexity in the UBM to capture more inter-speaker variability. Combining the new criterion with EM algorithm developed by (Figueiredo and Jain, 2002) and a new weight and mean adaptation algorithm, the new diarization system significantly decreased the mean of the DER compared to the baseline system. However, the standard deviation of the DER is still wide. No clear evidence supports the hypothesis that the new criterion works better when the speech length is longer. For the evaluation set

from the ISL corpus, the DER of the new system decreases when the speaker number becomes higher.

In the new system, when the EWPC criterion, the new EM algorithm (Figueiredo and Jain, 2002), and the weight and mean adaptation were all employed, the mean of the DER decreased from 13.60% to 9.24% in the development set, from 13.96% to 10.50% in the evaluation set (AMI), and 18.52% to 15.48% in the evaluation set (ISL), compared to the system with only the new SAD and SCD algorithms.

Termination Scheme: In Section 6.6, a new NC-based potential speaker merging termination scheme was developed to improve the steadiness of the new speaker diarization system. This new scheme is threshold free and makes the decision based on the global information. The new speaker diarization system containing the NC-based termination scheme narrowed the standard deviation of the DER, compared to the system with a local merging termination solution. When the NC is applied as the termination strategy for the potential speaker merging process and the stacks of the mean values of the potential speaker models are used as super-vectors, the standard deviation of the DER decreases. Although the mean of the DER increased from 9.24% to 10.33% for the development set, it decreased from 10.50% to 9.99% and from 15.48% to 13.70% for the evaluation sets (AMI) and (ISL).

The performance of the new systems: In contrast to the baseline system, the new systems with or without the new termination scheme had better performance. For the development set, the new system without the new termination scheme decreased the mean value of the DER from 18.61% to 9.24%, making an improvement of 9.37 percentage points; the new system with the new termination scheme decreased the mean value of the DER from 18.61% to 10.33%, making an improvement of 9.20 percentage points. For the evaluation set (AMI),

the new system without the new termination scheme decreased the mean value of the DER from 18.89% to 10.50%, making an improvement of 8.39 percentage points; the new system with the new termination scheme reduced the mean value of the DER from 18.89% to 9.99%, making a 8.90 percentage point improvement. For the evaluation set (ISL), the new system without the new termination scheme decreased the mean value of the DER from 21.35% to 15.48%, making an improvement of 5.86 percentage points; the new system with the new termination scheme reduced the mean value of the DER from 21.35% to 13.70%, making an improvement of 7.01 percentage points. Among the three datasets, the lowest mean value of the DER appears when using the new system without the new termination scheme. The new system with the new termination scheme, on the other hand, is steadier because the standard deviation of the two evaluation datasets is narrower. Therefore, we conclude that both systems have their own strengths.

7.2 Future work

An interesting area of recent work for speaker recognition is the use of latent factor analysis to compensate for speaker variability (Tsai et al., 2007). These methods adopt a GMM super-vector consisting of the stacked means of the GMM that is mean-only adapted from the UBM. Because this super-vector is of a high dimension (several hundreds or thousands dimension), SVM is seen to be a competent clustering strategy based on super-vectors. SVM is a popular classification strategy that clusters by projecting the data into a high dimension latent space. The kernels of the projected data are calculated, and the SVM algorithm clusters the data based on the kernel matrix directly.

SVM has been used in both the speaker recognition task and the speaker

verification in recent years. Because a collection of the mean values of speaker models adapted from the UBM can be used as a super-vector, and they are more discriminable between different speakers, adopting it for speaker diarization will reduce the influence of noise and speaker overlaps. SVM cluster data only depend on some vectors being at the class boundary (support vectors). Thus, using SVM avoids the need to detect the complicated intrinsic structure of the speaker data. However, SVM is always executed in a supervised way, whereas speaker diarization is an unsupervised task. Therefore, some modifications must be made if adopting SVM to speaker diarization.

The NC has been used as a cluster number selection criterion in this thesis. In graph theory, the optimum data partition can be obtained by minimising the NC. To solve a standard eigensystem, the second smallest eigenvector carries a clustering solution for a bi-cut. The other eigenvectors also carry different levels of dissimilarities in a graph. Combining these eigen-vectors, the global solution for clustering will be achieved, and the number of clusters may also be detected. Introducing the NC theory into the speaker diarization process to determine the speaker number appears to be an interesting future direction.

If, in speaker diarization, the speaker model can be sufficiently trained and the influence of the noise and speaker overlaps can be clearly removed, as is the case in speaker recognition research, the recognition rate will achieve a high value. However, the speaker diarization process has time constraints on many steps that make it difficult to identify speaker utterances of less than one second. Therefore, even if the speaker models were sufficiently trained and the number of speakers correctly detected, the performance of the speaker diarization will still be restricted by false alarm errors and missed short speaker turns.

Appendix A

Meeting characteristics and new system performance

Table A.1 and A.2 and A.3 shows all meetings used in experiments in Chapter 6, in terms of their type, number of speakers, the length of the speech in the meetings, NLR, and whether it is used in the development set (D) or the evaluation set (E).

Table A.4 lists the abbreviations of the experimental systems and their description.

Tables A.5, A.6 and A.7, show the results of the seven strategies performed on each of the meeting for the development and evaluation set.

Name	Room and Type	Number of Speakers	Speech Length (second)	NLR	Development or Evaluation
EN2002a	EN	4	1659.1	0.2338	D
EN2006a	EN	3	1852.8	0.4780	D
EN2009c	EN	3	2357.8	0.2183	D
ES2003a	ES	4	548.8	0.5185	D
ES2009a	ES	4	1077.1	0.2356	D
ES2016c	ES	4	1381.5	0.4043	D
IB4001	IB	4	1174.1	0.3433	D
IB4002	IB	4	1128.4	0.4044	D
IB4005	IB	3	1596.1	0.2123	D
IN1001	IN	3	2694.2	0.2284	D
IN1002	IN	4	2011.1	0.1903	D
IN1005	IN	4	2208.0	0.2157	D
IS1001b	IS	4	1454.5	0.3152	D
IS1006a	IS	4	516.8	0.3896	D
IS1009a	IS	4	552.8	0.3210	D

Table A.1: Meetings characteristics of development set

Name	Room and Type	Number of Speakers	Speech Length (second)	NLR	Development or Evaluation
EN2002b	EN	4	1303.3	0.2793	E
EN2002d	EN	4	1671.2	0.2514	E
EN2006b	EN	3	1597.5	0.4722	E
ES2002d	ES	4	1877.9	0.2461	E
ES2003b	ES	4	1539.4	0.2739	E
ES2003d	ES	4	1665.1	0.2959	E
ES2004a	ES	4	675.3	0.3607	E
ES2004d	ES	4	1510.1	0.3252	E
ES2005b	ES	4	1698.5	0.2681	E
ES2007a	ES	4	684.3	0.4356	E
ES2007b	ES	4	1127.3	0.3340	E
ES2007d	ES	4	823.8	0.3456	E
ES2009b	ES	4	1087.1	0.2436	E
ES2016b	ES	4	1381.9	0.4294	E
ES2016d	ES	4	913.1	0.4029	E
IB4004	IB	4	2032.0	0.1508	E
IB4011	IB	4	1892.7	0.2123	E
IN1007	IN	4	2039.2	0.1596	E
IN1008	IN	4	2636.3	0.2332	E
IN1009	IN	4	863.1	0.3117	E
IN1012	IN	4	2588.3	0.1719	E
IN1013	IN	4	2692.3	0.1513	E
IN1016	IN	4	3108.2	0.1452	E
IS1001c	IS	4	978.6	0.3263	E
IS1002c	IS	4	1580.2	0.2430	E
IS1002d	IS	4	838.9	0.3354	E
IS1003c	IS	4	1319.8	0.2865	E
IS1003d	IS	4	1487.5	0.2957	E
IS1006b	IS	4	1518.2	0.2988	E
IS1009b	IS	4	1655.4	0.1903	E

Table A.2: Meetings characteristics of evaluation set from AMI corpus

Name	Room and Type	Number of Speakers	Speech Length (second)	NLR	Development or Evaluation
m035	Game	4	2318.3	0.1814	E
m036	Game	5	1698.2	0.0287	E
m038	Disc	5	472.9	0.0212	E
m039a	Game	4	466.5	0.0998	E
m039b	Game	4	440.1	0.0734	E
m042	Chat	4	785.6	0.0725	E
m043	Proj	5	468.7	0.0511	E
m045	Disc	5	2414.2	0.0239	E
m046	Disc	4	1932.1	0.1247	E
m048	Disc	3	2817.0	0.0862	E
m051	Game	5	1185.7	0.2061	E
m052	Game	5	1686.4	0.1149	E
m055	Disc	9	2960.6	0.1134	E
m061	Disc	5	3163.9	0.0302	E
m063	Proj	5	1724.5	0.0682	E
m064	Disc	4	2039.0	0.1326	E

Table A.3: Meetings characteristics of evaluation set from ISL corpus

System Notation	System Description
Sys_{old}	Baseline system
Sys_{sad}	Baseline system with New SAD algorithm
Sys_{scd}	Baseline system with New SAD algorithm New SCD algorithm
Sys_{new}	Baseline system with New SAD algorithm New SCD algorithm Equal Weight Penalty Criterion a new EM algorithm Weight and mean adaptation
Sys_{new2}	Baseline system with New SAD algorithm New SCD algorithm Equal Weight Penalty Criterion a new EM algorithm Normalized Cuts based termination scheme

Table A.4: Experimental systems abbreviations and description

Name	Sys_{old}	Sys_{sad}	Sys_{scd}	Sys_{new}	Sys_{new2}
EN2002a	0.2074	0.1589	0.1085	0.0863	0.0607
EN2006a	0.1529	0.1941	0.1246	0.0864	0.1178
EN2009c	0.1531	0.1854	0.1689	0.0927	0.1168
ES2003a	0.2447	0.1995	0.1760	0.0921	0.0946
ES2009a	0.1988	0.1804	0.1828	0.1382	0.1067
ES2016c	0.3115	0.3605	0.2097	0.1179	0.0906
IB4001	0.1066	0.0874	0.1137	0.0278	0.0856
IB4002	0.1236	0.1269	0.1006	0.0772	0.1157
IB4005	0.1690	0.1762	0.1176	0.1222	0.1033
IN1001	0.1802	0.1849	0.1752	0.0810	0.1026
IN1002	0.1781	0.1966	0.1382	0.0902	0.1155
IN1005	0.2443	0.2059	0.1504	0.1133	0.1113
IS1001b	0.1750	0.1335	0.1112	0.0793	0.1080
IS1006a	0.1530	0.1939	0.1327	0.1181	0.1201
IS1009a	0.1980	0.1498	0.1141	0.1261	0.1158

Table A.5: The DER of development set

Name	Sys_{old}	Sys_{sad}	Sys_{scd}	Sys_{new}	Sys_{new2}
EN2002b	0.1705	0.1573	0.1384	0.0818	0.0995
EN2002d	0.1696	0.1532	0.1653	0.0802	0.0975
EN2006b	0.2114	0.1814	0.1264	0.0583	0.1007
ES2002d	0.2079	0.1760	0.1747	0.1157	0.0789
ES2003b	0.2602	0.2336	0.2058	0.1349	0.1077
ES2003d	0.2539	0.2036	0.1279	0.0295	0.1134
ES2004a	0.1702	0.2133	0.1137	0.1384	0.1172
ES2004d	0.1410	0.1638	0.1806	0.1082	0.0843
ES2005b	0.1672	0.1746	0.1032	0.1638	0.1085
ES2007a	0.1405	0.1676	0.0996	0.0661	0.1011
ES2007b	0.1910	0.1687	0.1565	0.0763	0.1165
ES2007d	0.2048	0.1819	0.1368	0.1695	0.1104
ES2009b	0.2611	0.1543	0.1411	0.1609	0.1180
ES2016b	0.1882	0.2800	0.2479	0.1288	0.0959
ES2016d	0.0908	0.0824	0.0900	0.1149	0.1204
IB4004	0.2399	0.1869	0.0868	0.0353	0.0540
IB4011	0.1812	0.0848	0.1281	0.0171	0.0883
IN1007	0.1476	0.2200	0.1014	0.1384	0.0721
IN1008	0.1911	0.2150	0.1230	0.0410	0.0909
IN1009	0.1756	0.1826	0.1496	0.0269	0.1128
IN1012	0.1650	0.1554	0.1228	0.1063	0.1165
IN1013	0.1694	0.1540	0.1137	0.1571	0.1104
IN1016	0.2027	0.1800	0.1291	0.0438	0.1180
IS1001c	0.2376	0.1509	0.1914	0.1804	0.0962
IS1002c	0.1936	0.2325	0.1469	0.1474	0.0959
IS1002d	0.2214	0.2299	0.1653	0.1037	0.1204
IS1003c	0.1555	0.1524	0.1460	0.1933	0.0953
IS1003d	0.1899	0.2102	0.1627	0.1257	0.1075
IS1006b	0.1896	0.1351	0.1516	0.1565	0.1224
IS1009b	0.1892	0.2010	0.1226	0.1853	0.1137

Table A.6: The DER of evaluation set from AMI corpus

Name	Sys_{old}	Sys_{sad}	Sys_{scd}	Sys_{new}	Sys_{new2}
m035	0.2169	0.2190	0.2336	0.1823	0.1390
m036	0.1872	0.1541	0.1377	0.0921	0.1481
m038	0.2036	0.1932	0.1811	0.1422	0.1392
m039a	0.2005	0.2028	0.1724	0.1583	0.1192
m039b	0.2083	0.1961	0.1434	0.1266	0.1278
m042	0.2467	0.2052	0.1617	0.0984	0.1345
m043	0.1814	0.1531	0.1579	0.1811	0.1661
m045	0.2400	0.1787	0.1820	0.1645	0.1200
m046	0.2307	0.2251	0.2113	0.1728	0.1525
m048	0.2165	0.2121	0.1944	0.0982	0.1324
m051	0.1841	0.1802	0.1966	0.1318	0.1162
m052	0.2392	0.2031	0.1565	0.1654	0.1314
m055	0.2302	0.2281	0.2075	0.1982	0.1353
m061	0.1973	0.1972	0.2304	0.2566	0.1081
m063	0.2130	0.1875	0.1836	0.1469	0.1408
m064	0.1905	0.2037	0.2433	0.2326	0.1208

Table A.7: The DER of evaluation set from ISL corpus

Glossary of Acronyms

AMI	Augmented Multi-party Interaction
ASNR	Average Speech to Noise Ratio
BIC	Bayesian Information Criterion
CCR	Cluster Complexity Ratio
CLC	Classification Likelihood Criterion
CLR	Cross Log-likelihood Ratio
CVEM	Cross Validation EM
DCT	Discrete Cosine Transform
DER	Diarization Error Rate
EHMM	Evolutive Hidden Markov Model
EM	Expectation-Maximization
EWPC	Equal Weight Penalty Criterion
FA	False Alarm
FDA	Fisher linear Discriminant Analysis
FDR	Fisher linear Descriminant Ratio
FDC	Fisher Descriminant Classifier
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
ISL	Interactive Systems Laboratories
KL2	Kullback Divergency 2
MAP	Maximum A Posterior
MFCC	Mel-Frequency Cepstrum Coefficients
MISS	Missing speech error rate
ML	Maximum Likelihood
MST	Minimum Spanning Tree
NC	Normalized Cuts
NGMM	Non-speech GMM

NIST-RT	National Institute of Standards and Technolog-Rich Transcripti
NLR	Noise Length Ratio
SAD	Speech Activity Detection
SCD	Speaker Change Detection
SGMM	Speech GMM
SVM	Support Vector Machine
UBM	Universal Background Model
VQ	Vector Quantization

Experimental Systems

Abbreviations

System Notation	System Description
Sys_0 or Sys_{old}	Baseline system
Sys_{sad}	Baseline system with New SAD algorithm
Sys_{scd}	Baseline system with New SAD algorithm New SCD algorithm
Sys_1	Baseline system with New SAD algorithm New SCD algorithm Cluster Complexity Ratio criterion Incremental training
Sys_2	Baseline system with New SAD algorithm New SCD algorithm Equal Weight Penalty Criterion Cross-validation EM
Sys_{new}	Baseline system with New SAD algorithm New SCD algorithm Equal Weight Penalty Criterion a new EM algorithm Weight and mean adaptation
Sys_{new2}	Baseline system with New SAD algorithm New SCD algorithm Equal Weight Penalty Criterion a new EM algorithm Normalized Cuts based termination scheme

Glossary of Symbols

f_{mel}	Mel-scale frequency
f_c	Centre frequency
$e(n)$	Energy vector of the nth frame
$o(t)$	The tth discrete signal in a frame
T	Number of feature vectors in a frame
D^{hat}	Dimension of a feature vectors
x	A feature vector
X	A sequence of feature vector
N	Number of feature vectors in a sequence
μ_i	Mean of the ith component in the GMM
Σ_i	Covariance Matrix of the ith component in the GMM
w_i	Weight of the ith component in the GMM
M	Number of components in the GMM / model complexity
λ	Collection of all parameters in the GMM
$g_i(x)$	Probability of the appearance of x given the ith component
$p(x\|\lambda)$	Conditional probability of the appearance of x given parameter λ
$p(X\|\lambda)$	Conditional probability of the appearance of X given parameter λ
λ_{speech}	Collection of all parameters in the speech GMM
$\lambda_{non-speech}$	Collection of all parameters in the non-speech GMM
$\hat{k}$	Selected acoustics cluster for x
$BIC(M)$	BIC score of model whose model complexity is M
ΔBIC	BIC score difference
$L(X\|M)$	log likelihood of X given the model whose model complexity is M
ΔM	Model complexity difference
ϕ	Constant parameter in the BIC
$D_{KL}(P_1\|\|P_2)$	KL divergence between distribution P_1 and P_2
$tr(\Sigma)$	Trace of covariace matrix Σ
K	Number of speakers in a meeting
μ_i^{ubm}	Mean of the component i in the UBM
$\tilde{\mu_i}$	Adapted mean of the component i in the speaker model
ρ	Fixed relevance factor for mean adaptation
τ_{ji}	posterior probability UBM component i given x_j
$CLR(X_1, X_2)$	Cross log-likelihood ratio of X_1 and X_2

E_{MISS}	Missed speech error rate
E_{FA}	False alarm error rate
E_{spkr}	Wrong speaker error rate
ψ	Weight of a hyperplane in high dimension space
b	Bias of the hyperplane in high dimension space
$J_f(\psi)$	FDR when data is projected onto the hyperplane $< \psi^*, x > +b = 0$
ψ^*	Weight of the hyperplane that maximizes the FDR
α	Balance control parameter in the speaker change detection
λ_M	Collection of all GMM parameters when the model complexity is M
$\hat{\lambda_M}$	Maximum likelihood estimate of λ_M
$\hat{M}$	Maximum likelihood estimate of M
$Pe(M, \hat{\lambda_M})$	Penalty term based on parameters λ_M and model complexity M
$IC(\hat{\lambda_M}, M)$	Model complexity selection criterion
Z	Latent indicator variables
$p(Z\|\lambda_M)$	Probability of Z given parameters λ_M
$p(X\|Z, \lambda_M)$	Conditional probability of X given Z and parameters λ_M
$p(X, Z\|\lambda_M)$	Joint probability of X and Z given parameters λ_M
$L_c(X, Z\|\lambda_M)$	Complete joint log-likelihood of X and Z given paramters λ_M
$EC_M(X\|\lambda_M)$	Entropy of Z
τ_j^i	Posterior probability of the jth component given x_i
$p_0(w)$	Prior probability of w
δ	Parameter of the multinomial distribution
Dir	Dirichlet distribution
Γ	$\Gamma(\delta) = \int_0^\infty e^{-t}dt$
$p_{w\|z}$	Posterior distribution of w
$D_{KL}(p_{w\|z}, p_0)$	KL divergence between $p_{w\|z}$ and p_0
$\tilde{\lambda_M}$	Posterior mode of λ_M
$H(\tilde{\lambda_M})$	Hessian matrix with respect to λ_M
$I(\hat{\lambda_M}\|X)$	Observed information matrix with respect to λ_M given X
$blockdiag$	Block diagonal matrix
$I^{(1)}(\mu_i, \Sigma_i)$	Observed information matrix given a single observation
Λ	$\Lambda = (\prod_{j=1}^M w_j)^{-1}$
$\Omega(\mu, \Sigma$	Number of parameters in a Gaussian component
Λ_M^{t-1}	Λ_M computed in the $(t-1)$th iteration
Λ_M^t	Λ_M computed in the tth iteration
$Q(\lambda_M, \lambda_M^{t-1})$	$L_c(X\|Z, \lambda_M^t)$ when Z is computed using Λ_M^{t-1}
$\tilde{i}$	Adapted weight of the component i in the speaker model
β	Parameter that controls model complexity in the non-speech GMM
v_i	The ith vector
$S(v_i, v_j)$	The normalized inner product of two vectors
$Ncut(A, B)$	The normalized dissimilarity between disjoint sets A and B
$cut(A, B)$	The total dissimilarity from A to B
V	$A \bigcup B = V$
$assoc(A, V)$	The total connection from A to V
dw_{ij}	The dissimilarity between V_i and V_j

References

Ajmera, J., Lapidot, I., and McCowan, I. (2002). Unknown multiple speaker clustering using HMM. In *Proc. International Conference on Spoken Language Processing*, Denver, USA, pages 573–576.

Ajmera, J. and Wooters, C. (2003). A robust speaker clustering algorithm. In *Proc. IEEE International Workshop on Automatic Speech Recognition and Understanding*, Virgin Islands, US, pages 411–416.

Akita, Y. and Kawahara, T. (2003). Unsupervised speaker indexing using anchor models and automatic transcription of discussions. In *Proc. International Conference on Eurospeech*, Geneva, Switzerland, pages 25 – 33.

Anguera, X., Shinozaki, T., Wooters, C., and Hernando, J. (2007). Model complexity selection and cross-validation EM training for robust speaker diarization. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Honolulu, USA, pages IV–273–IV–276.

Anguera, X., Wooters, C., and Hernando, J. (2006a). Automatic cluster complexity and quantity selection: toward robust speaker diarization. In *Proc. Workshop on Machine Learning for Multimodal Interaction Workshop*, Washington DC, USA, pages 248–256.

Anguera, X., Wooters, C., and Hernando, J. (2006b). Friends and enemies: a novel initialization for speaker diarization. In *Proc. IEEE International Conference on Spoken Language Processing - Interspeech 2006*, Pittsburgh, USA, pages 1661–1664.

Anguera, X., Wooters, C., and Hernando, J. (2006c). Purity algorithms for speaker diarization of meeting data. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Toulouse, France, pages 993–996.

Anguera, X., Wooters, C., Peskin, B., and Aguilo, M. (2005). Robust speaker segmentation for meetings: the ICSI-SRI spring 2005 diarization system. In *Proc. Workshop on Machine Learning for Multimodal Interaction Workshop*, Edinburgh, UK, pages 402–414.

Atal, B. and Hanauer, S. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50:637–655.

Attias, H. (2001). Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conference of Uncertainty Artificial Intelligence*, volume 13, pages 1649–1681.

Banfield, J. and Raftery, A. (1997). Model-based Gaussian and Non-Gaussian clustering. *Biometrics*, 49:1–10.

Barras, C. and Gauvain, J. (2003a). Feature and score normalization for speaker verification of cellular data. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Process*, Hong Kong, China, pages 49–52.

Barras, C. and Gauvain, J. (2003b). Feature warping for robust speaker verification. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Hong Kong, China, pages 753–756.

Barras, C., Zhu, X., Meignier, S., and Gauvain, J. (2004). Improving speaker diarization. In *Proc. Fall Rich Transcription Workshop (RT-04)*, New York, USA.

Barras, C., Zhu, X., Meignier, S., and Gauvain, J. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1505–1513.

Bensmail, H., Celeux, G., and Rafter, A. (1997). Inference in model-based clutser analysis. *Statistics and Computing*, 7(6):1–10.

Beranek, L. (1949). *Acoustic measurements*. McGraw-Hill, New York.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley and Sons, UK.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated classification likelihood. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Biernacki, C. and Govaert, G. (1997). Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse-Fisher information matrix. *Computing Science and statistics*, 29:451–457.

Biernacki, C. and Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20:267–272.

Boakye, K., Hornero, B., Binyals, O., and Friedland, G. (2008). Overlapped speech detection for improved speaker diarization in multiparty meetings. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages 4353–4356.

Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *The 5th Annual ACM Workshop on COLT*, page 144152.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture model using a new information matrix. *Information and classification*, pages 40–54.

Bridle, J. and Brown, M. (1974). An experimental automatic word-recognition system. Technical report, Joint Speech Research Unit, Ruislip, England.

Bruno, E., Loccoz, N., and Maillet, S. (2008). Design of multimodal dissimilarity spaces for retrieval of video documents. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 30(9):1520–1533.

Brutti, A., Omologo, M., and Svaizer, P. (2008a). Localization of multiple speakers based on a two step acoustic map analysis. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages 4349–4352.

Brutti, A., Omologo, M., Svaizer, P., and Zieger, C. (2007). Classification of acoustic maps to determine speaker position and orientation from a distributed microphone

network. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Honolulu, USA, pages IV–493–IV–496.

Brutti, A., Omologo, M., Svaizer, P., and Zieger, C. (2008b). The signal change-point detection using the high-order statistics of log-likelihood difference functions. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages 4381–4384.

Burger, S., MacLaren, V., and Yu, H. (2002). The ISL meeting corpus: the impact of meeting type on speech style. In *Proceedings of the ICSLP '02*, pages 301–304.

Campbell, J., Fraley, C., Murtagh, F., and Raftery, A. (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*, 18(2):1539–1548.

Campbell, J. P. (1997). Speaker recognition: a tutorial. *IEEE Transactions On Speech and Audio Processing*, 1(9):1437–1462.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, 41(2):181190.

Celeux, G. and Soromenho, G. (1996). An Entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13:195–212.

Chan, W., Lee, T., Zheng, N., and Hua, O. (2006). Use of vocal source features in speaker segmentation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, pages 3435–3438.

Chen, S. and Gopalakrishnam, P. (1998). Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proc. Workshop on 1998 DARPA Broadcast News Transcription and Understanding*, Lansdowne, VA, pages 127–132.

Cheung, Y. (2005). Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection. *IEEE Transaction on Knowledge and Data Engineering*, 17(6):750–761.

Corduneanu, A. and Bishop, C. (2001). Variational Bayesian model selection for mixture distributions. In *Proc. Artificial Intelligence and Statistics*, pages 27–34.

Cortes, L. (2008). Efficient annotation of vesicle dynamics in video microscopy. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 30(11):1998–2010.

Cover, T. and Hall, W. (1991). *Elements of information theory*. John Wiley and Sons, New York.

Dasgupta, A. and Raftery, A. (1998). Detecting features in spatial point patterns with clutter via model-based clustering. *Journal of American Statistical Society (B)*, 93:294–302.

Davis, S. and Mermelstein, P. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 64(4):460–475.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of Royal Statistical Society (B)*, 39:1–38.

Deshayes, J. and Picard, D. (1986). Off-line statistical analysis of change-point models using non-parametric and likelihood methods. In *Lecture Notes in Control and Information Sciences, Volume 77*, pages 103–168.

Dijkstra, E. (1960). Some theorems on spanning subtrees of a graph. *Indagationes Mathematicae*, 28(4):196–199.

Faltlhauser, R. and Ruske, G. (2001). Robust speaker clustering in eigenspace. In *Proc. IEEE International Workshop on Automatic Speech Recognition and Understanding*, Trento, Italy, pages 57–60.

Farell, K., Mammone, R., and Assaleh, K. (1994). Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions On Speech and Audio Processing*, 2(1):194–205.

Figueiredo, M. and Jain, A. (2002). Unsupervised learning of finite mixture models. *IEEE Transaction on Pattern Analysis and Machine Intelligent*, 24(3):381–396.

Fine, S., Navratil, J., and Gopinath, R. (2001). A hybrid GMM/SVM approach to speaker

identification. In *Proc. International Conference on Acoustic Speech and Signal Processing*, Salt Lake City, Utah, pages 417–420.

Fiscus, J., Radde, N., Garofolo, J., Le, A., Ajot, J., and Larun, C. (2005). The rich transcription 2005 spring meeting recognition evaluation. In *Proc. Workshop on Machine Learning for Multimodal Interaction*, Edinburgh, pages 369–389.

Fraley, C. and Raftery, A. (1998). How many clusters and which clustering method. Technical report, Department of statistics, University of Washington, Seattle, WA, USA.

Fredouille, C. and Evans, N. (2008). New implementations of the E-HMM-based system for speaker diarization in meeting rooms. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages 4357–4360.

Friedland, G., Vinyals, O., Huang, Y., and Muller, C. (2009). Prosodic and other long-term features for speaker diarization. *Journal of Computer Science and Technology*.

Fu, R. and Benest, I. (2007a). An improved speaker diarization system. In *Proc. IEEE International Conference on INTERSPEECH 2007*, Antwarp, Belgium, pages 2605–2608.

Fu, R. and Benest, I. (2007b). Improvement in diarization system. In *Proc. IEEE International Conference on Signal Processing and Multimedia Applications*, Barcelona, Spain, pages 918–921.

Gales, M., Kim, D., Woodland, P., Chan, H., Mrva, D., Sinha, R., and Tranter, S. (2006). Progress in the cu-htk transcription system. *IEEE transaction on Audio, Speech and Language Processing*, 14(5):1511–1523.

Gauvain, J., Lamel, L., and Adda, G. (1998). Partitioning and transcription of broadcast news data. In *Proc. IEEE International Conference on Spoken Language Processing*, Sydney, Australia, pages 1335–1338.

Ghahramani, Z. and Beal, M. (2000). *Variational inference for Bayesian mixtures of factor analysis*. MIT Press, Cambridge.

Gish, H., Siu, M.-H., and Rohlicek, R. (1991). Segregation of speakers for speech

recognition and speaker identification. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, pages 873–876.

Gupta, V., Boulianne, G., Kenny, P., Quellet, P., and Dumouche, P. (2008). Speaker diarizatioin of French broadcast news. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages 4365–4368.

Gupta, V., Kenny, P., Quellet, P., Boulianne, G., and Dumouchel, P. (2007). Multiple feature combination to improve speaker diarization of telephone conversations. In *Proc. IEEE International Workshop on Automatic Speech Recognition and Understanding*, Kyoto, Japan, pages 705–710.

Hain, T., Burget, L., Dines, J., Giulia, G., Karafiat, M., Lincoln, M., Vepa, J., and Wan, V. (2007). The AMI system for the transcription of speech in meetings. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Honolulu, Hawai'i, U.S.A, pages 357–360.

Han, K. and Narayanan, S. (2007). A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system. In *Proc. IEEE International Conference on INTERSPEECH*, Antwerp, Belgium, pages 1853–1856.

Han, K. and Narayanan, S. (2008). A novel inter-clustering distance measure combining GLR and ICR for improved agglomerative hierarchical speaker clustering. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages 4373–4376.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):10291054.

Hermansky, H. (1990a). Auditory model for parameterization of speech in real-life environment based on re-integration of temporal derivative of auditory spectrum. Technical report, US WEST Advanced Technologies Research Report, US.

Hermansky, H. (1990b). Perceptual Linear Predictive (PLP) analysis of speech. *Journal of Acoustic Society*, 87(4):1738–1752.

Ho, T. and Basu, M. (2002). Complexity measure of supervised classification problems.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):289–300.

Hofmann, T. and Buhmann, J. (1997). Pairwise data clustering by deterministic annealing. *IEEE Transaction Pattern Analysis and Machine Intelligence*, 19(11):1–14.

Hung, H., Huang, Y., Friedland, G., and Perez, D. (2008). Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages 2197–2200.

Jaakkola, T. (2000). Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, (4):129–160.

Johnson, S. and Woodland, S. (2000). A method for direct audio search with applications to indexing and retrieval. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Istanbul, Turkey, pages 1427–1430.

Kadri1, H., Davy, M., Rabaoui1, A., Lachiri1, Z., and Ellouze1, N. (2008). Robust audio speaker segmentation using one class SVMs. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages 4373–4376.

Kemp, T., Schmidt, M., Westphal, M., and Waibel, A. (2000). Strategies for automatic segmentation of audio data. In *Proc. International Conference on Acoustic Speech and Signal Processing*, Istanbul, Turkey, pages 1423–1426.

Kharroubi, J., Petrovska, D., and Chollet, G. (2001). Combining GMM's with support vector machines classifier. In *Proc. International Conference on Eurospeech 2001*, Aalborg, Denmark, pages 1757–1760.

Kloppenburg, M. and Tavan, P. (1997). Deterministic annealing for density estimation by multivariate Normal Mixtures. *Physical Review*, 55:2089–2092.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. 1st. Ital. Attuari*, 4(10):83–91.

Kruskal, J. (1956). On the shortest spanning tree of a graph and the traveling salesman problem. *American Math Society*, 7:48–50.

Kubala, F., Jin, H., Matsoukas, S., Gnuyen, L., Schwartz, R., and Machoul, J. (1997). The 1996 BBN byblos HUB-4 transcription system. In *Proc. Speech Recognition Workshop*, pages 90–93.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):7986.

Lee, V., Mella, O., and Fohr, D. (2007). Speaker diarization using Normalized Cross Likelihood Ratio. In *Proc. IEEE International Conference on INTERSPEECH*, Antwerp, Belgium, pages 2354–2358.

Leeuwen, D. (2005). The TNO speaker diarization system for NIST RT05s meeting data. In *Proc. Workshop on Machine Learning for Multimodal Interaction*, Edinburgh, pages 428–439.

Li, Q., Zheng, J., Tsai, A., and Zhou, Q. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Transactions On Speech and Audio Processing*, 10(3):103–107.

Liu, D. and Kubala, F. (1999). Fast speaker change detection for broadcast news and indexing. In *Proc. Eurospeech 99*, Budapest, Hungary, pages 1031–1034.

Liu, D. and Kubala, F. (2004). Speaker diarization for broadcast news. In *Proc. Odyssey speaker and language recognition workshop*, Toledo, Spain, pages 337–344.

Liu, Z., Wang, Y., and Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *VLSISignal Process*, 20(1-2):61–79.

Lu, L. and Zhang, H.-J. (2002). Speaker change detection and tracking in real-time news broadcasting analysis. In *Proc. ACM International Conference on Multimedia*, pages 602–610.

Ma, C., Nguyen, P., and Mahajan, M. (2008). Finding speaker identities with a conditional maximum entropy model. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages IV–261–IV–264.

Mahalanobis, P. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):4955.

Martin, A. and Przybocki, M. (2001). Speaker recognition in a multi-speaker environment. In *Proc. European Conference on Speech Communication and Technology*, Aalborg, Denmark, pages 787–790.

Matsui, T. and Furui, S. (2004). Comparison of text independent speaker recognition methods using VQ distorsion and discrete/continuous HMM's. *IEEE Transactions On Speech and Audio Processing*, 2(3):456–459.

McLachlan, G. (1987). On Bootstrapping the likelihood ratio test statistic for the number of components in a Normal mixture. *Journal of Royal Statistical Society (C)*, 36:318–324.

McLachlan, G. and Peel, D. (1997). *The EM algorithm and Extensions*. John Wiley and Sons, New York.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York.

McLanchlan, G. and Basford, K. (1988). *Mixture Models: Inference and Application to Clustering*. Marcel Dekker, New York.

Meignier, S., Bonastre, J., Fredouille, C., and Merlin, T. (2000). Evolutive HMM for multispeaker tracking system. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Istanbul, Turkey, pages 1201–1204.

Meignier, S., Bonastre, J., and Igounet, S. (2001). E-HMM approach for learning and adapting sound models for speaker indexing. In *Proc. Workshop on Odyssey Speaker and Language Recognition*, Crete, Greece, pages 175–180.

Meignier, S., Moraru, D., Fredouille, C., Bonastre, J., and Besacier, L. (2005). Step-by-step and integrated approaches in broadcast news speaker diarization. *Computing Speech and Language*, 20(20):303–330.

Meinicke, P. and Ritter, H. (2001). Resolution-based complexity control for Gaussian Mixture Models. *Neural Computation*, 13(2):453–475.

Mengersen, K. and Robert, C. (1996). Testing for mixtures: a Bayesian entropic approach. In *Proc. 5th International Meeting on Bayesian Statistics*, pages 255–276.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. In *Proc. IEEE International Conference on Pattern Recognition and Artificial Intelligence*, New York, USA, pages 374–388.

Miro, X. (October, 2006). Robust speaker diarization for meetings. Technical report, Speech Processing Group, Department of Signal Theory and Communications, Universitat Polit‘ecnica de Catalunya, Barcelona, Spain.

Moh, Y., Nguyen, P., and Junqua, J. (2003). Toward domain independent speaker clustering. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Hong Kong, China, pages 85–88.

Mood, A., Graybill, F., and Bose, D. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill Companies, The New York, USA.

Moraru, D., Meignier, S., Fredouille, C., Besacier, L., and Bonastre, J. F. (2003). The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Hong Kong, China, pages 6–10.

Neal, R. (1992). Bayesian mixture modeling. In *Proc. 11th International Workshop Maximum Entropy and Bayesian Methods of Statistical Analysis*, pages 197–211.

Neal, R. and Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graph Models*, pages 335–370.

Nishida, M. and Kawahara, T. (2003). Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion. In *Proc. International Conference on Acoustic Speech and Signal Processing*, Hong Kong, China, pages 2106–2110.

Oliver, J., Baxter, R., and Wallace, C. (1996). Unsupervised learning using MML. In *Proc. 13th International Conference of Machine Learning*, pages 364–372.

Pertila, P. and Parviainen, M. (2007). Robust speaker localization in meeting room domain. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Honolulu, USA, pages IV–497–IV–500.

Pfau, T., Ellis, D., and Stolcke, A. (2001). Multispeaker speech activity detection for the ICSI meeting recorder. In *Proc. IEEE International Workshop on Automatic Speech Recognition and Understanding*, Trento, Italy, pages 107–110.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proc. IEEE*, 78(9):1484–1487.

Prim, R. (1957). Shortest connection network and some generalizations. *The Annals of Statistics*, 36:1389–1404.

Quenot, G., Moraru, D., and Besacier, L. (2003). Clips at TRECVID: Shot boundary detection and feature detection. In *TRECVID 2003 Workshop Notebook Papers*, Gaithersburg, USA, pages 124–138.

Rasmussen, C. (2000). The infinite Gaussian Mixture Model. In *Proc. 12th Neural Information Processing Systems*, pages 554–560.

Reynolds, D. (2002). An overview of automatic speaker recognition technology. In *Proc. International Conference on Acoustic Speech and Signal Processing*, Orlando, USA, pages 4072–4075.

Reynolds, D. and Carrasquillo, P. (2004). The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations. In *Proc. Fall 2004 Rich Transcription workshop (RT-04)*, New York, USA, pages 337–347.

Reynolds, D., Quatieri, T., and Dunn, R. (2000). Speaker verification using adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(4):19–41.

Reynolds, D. and Rose, R. (1995). Robust text independent speaker identification using Gaussian Mixture Models. *IEEE Transactions On Speech and Audio Processing*, 3(1):72–83.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with unknown number of components. *Journal of Royal statistics Society (B)*, 59(4):731–792.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. World Scientific, Singapore.

Roberts, S., Husmeier, D., Rezek, I., and Penny, W. (1998). Bayesian approaches to Gaussian Mixture Modelling. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of American Statistical Association*, 92:894–902.

Rose, K. (1998). Deterministic annealing for clustering compression, classification, regression and related optimization problems. *IEEE Transaction on Neural Networks*, 86:2210–2239.

Rumelhart, D., McClelland, J., and Hinton, G. (1986). Parallel distributed processing. Technical report, PDP Research Group, Massachusetts Institute of Technology, University of California, San Diego, USA.

Salton, G. (2000). Text indexing using complex identifiers. In *Proc. the ACM conference on Document processing systems*, Santa Fe, United States, pages 135–144.

Sato, M. (2001). On-line model selection based on variational Bayes. *Neural Computing*, 13(7):1649–1681.

Saunders, J. (1996). Real-time discrimination of broadcast speech/music. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Atlanta, GA, pages 993–996.

Schutze, H. and Manning, C. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, UK.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Shi, J. and Malik, J. (2000). Normalized Cuts and image segmentation. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Siegler, M., Jain, U., Ray, B., and Stern, R. (1997). Automatic segmentation, classification and clustering of broadcast news. In *Proc. Workshop on 1998 DARPA Broadcast News Transcription and Understanding*, Chantilly, VA, pages 97–99.

Sinha, R., Tranter, S., Gales, M., and Woodland, P. (2005). The Cambridge University

March 2005 speaker diarization system. In *Proc. European Conference on Speech Communication Technology*, Lisbon, Portugal, pages 2437–2440.

Smirnov, N. (1948). Tables for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistic*, 1(19):279279.

Smyth, P. (2000). Model selection for probabilistic clustering using Cross-validated likelihood. *Statistics and Computing*, 10(1):63–72.

Titterington, D., Smith, A., and Makov, U. (1991). *Elements of information theory*. John Wiley and Sons, New York.

Tranter, S. (2005). Two-way cluster voting to improve speaker diarization performance. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Philadelphia, USA, pages 753–756.

Tranter, S. (2006). Who really spoke when?-Finding speaker turns and identities in audio. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Toulouse, France, pages 1013–1016.

Tranter, S., Gales, M., Sinha, R., Umesh, S., and Woodland, P. (2004). The development of the Cambridge University rt-04 diarization system. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, New York, USA.

Tranter, S. and Reynolds, D. (2006). An overview of automatic speaker diarization system. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1557–1565.

Tsai, W., Chen, S., and Wang, H. (2007). Automatic speaker clustering using a voice characteristic reference space and maximum purity estimation. *IEEE transaction on Audio, Speech and Language Processing*, 15(4):1511–1523.

Tsai, W., Cheng, S., and Wang, H. (2004). Speaker clustering using a voice characteristic reference space. In *Proc. IEEE International Conference on Spoken Language Processing*, Jeju Island, Korea, pages 2937–2940.

Tsai, W., Cheng, S., and Wang, H. (2005). Clustering speech utterances by speaker using

eigenvoice-motivated vector space model. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Philadelphia, USA, pages 725–728.

Tsai, W.-H. and Cheng, S.-S. andvWang, H.-M. (2006). Speaker clustering of speech utterances using a voice characteristic reference space. In *Proc. International Conference on Speech and Language Processing*, Jeju, Korea, pages 2432–2437.

Ueda, N. and Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, 15(10):1223–1241.

Ueda, N. and Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks*, 11(11):271–282.

Ueda, N., Nakano, Z., Ghahramani, Z., and Hinton, G. (2000). SMEM algorithm for mixture models. *Neural Computing*, 12(9):2109–2128.

Vijayasenan, D., Valente, F., and Bourlard, H. (2007). Agglomerative information bottleneck for speaker diarization of meetings data. Technical report, IDIAP, ASRU.

Vijayasenan, D., Valente, F., and Bourlard, H. (2008). Combination of agglomerative and sequential clustering for speaker diarization. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, Las Vegas, USA, pages 4361–4364.

Wallace, C. and Dowe, D. (1999). Minimum message length and kolmogorov complexity. *Computer Journal*, 42(4):270–283.

Wallace, C. and Freeman, P. (1987). Estimation and inference via compact coding. *Journal of Royal Statistical Society (B)*, 49(3):241–252.

Wan, V. and Renals, S. (2005a). Speaker verification using sequence discriminant support vector machines. *IEEE Transactions On Speech and Audio Processing*, 13(2):203–210.

Wan, V. and Renals, S. (2005b). Speaker verification using sequence discriminant support vector machines. *IEEE Transactions On Speech and Audio Processing*, 13(2):203–210.

Wang, Z., Gemon, D., Luo, J., and Grey, R. (2008). Real-world image annotation and

retrieval: An introduction to the special section. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 30(11):1873–1876.

Whindham, M. and Cutler, A. (1992). Information ratios for validating mixture analysis. *Journal of American Statistical Association*, 87:1189–1192.

Willsky, A. S. and Jones, H. L. (1976). A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, AC-21(1):108–112.

Wooters, C., Fung, J., Peskin, B., and Anguera, X. (2004). Toward robust speaker segmentation: the ICSI-SRI fall 2004 diarization system. In *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, New York, USA.

Xu, D. and Chang, S. (2008). Video event recognition using Kernel methods with multilevel temporal alignment. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 30(11):1985–1997.

Yamaguchi, M., Yamashita, M., and Matsunaga, S. (2006). Spectral cross-correlation features for audio indexing of broadcast news and meetings. In *Proc. International Conference on Speech and Language Processing*, Jeju, Korea, pages 2396–2400.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtcher, V., and Woodland, P. (2005). *Progress in the CU-HTK transcription system*. Cambridge University Engineering Department, Cambridge, UK.

Yu, C., Wu, C., and Huang, C. (2004). A study of using automatic text indexing to analyze web browsing behaviour. *Intelligent Control and Automation*, 5(5):3991–3995.

Zheng, F., Zhang, G., and Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, 16(6):582–589.

Zhou, B. and Hansen, J. (2000). Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. In *Proc. IEEE International Conference on Spoken Language Process*, Beijian, China, pages 714–717.

Zhu, X., Barras, C., Meignier, S., and Gauvian, J. (1998). Combining speaker identi-

fication and bic for speaker diarization. In *Proc. European Conference on Speech Communication Technology*, Sydney, Australia, pages 1335–1338.