Speaker diarization in meetings domain

Nguyen, Trung Hieu

2015

Nguyen, T. H. (2015). Speaker diarization in meetings domain. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/62927>

<https://doi.org/10.32657/10356/62927>

Speaker Diarization in Meetings Domain

A thesis
submitted to the School of Computer Engineering
of the Nanyang Technological University

by

Nguyen Trung Hieu

for partial fulfilment of the requirements
for the degree of Doctor of Philosophy

January 2015

Abstract

The purpose of this study is to develop robust techniques for speaker segmentation and clustering with focus on meetings domain. The techniques examined can however be applied to any other domains such as telephone and broadcast news.

Traditional techniques for speaker diarization developed for telephone conversations or broadcast news are based on a single channel, which is notably different from meetings domain which can have multiple channels. These techniques when adapted to meetings domain however perform poorer than expected since they do not exploit direction of arrival information, which is available in many meeting rooms with the presence of multiple microphones. Moreover, many of these techniques are involved with tunable parameters, which are presumably derived using external data. These parameters need to be individually adjusted for each data set accordingly to obtain reasonable performance. In this thesis, the focus is on robust and accurate speaker diarization techniques in meetings. Our aim is to improve the segmentation and clustering performance in diverse conditions while keeping the number of manually tuned parameters to minimal.

Starting from the widely adopted agglomerative hierarchical clustering framework, a comparative study of various distance metrics is conducted for the purpose of finding the most appropriate metric to use for speaker clustering. In contrast to general practice, it is shown that the popular *Generalized Likelihood Ratio* (GLR) based metrics such as GLR and *Bayesian Criterion Information* (BIC) should not be used as distance metrics since they are not robust to size variations. As a result, a novel metric is proposed which can be seen as an extension of the *Information Change Rate* (ICR) by exploiting the second-order statistics of the likelihood scores. The proposed metric is shown to be much less affected by the length of speech segments and the results on diarization tasks show improvements on diarization error rate (DER) of more than 10% relatively comparing

to GLR. Having addressed the topic of cluster merging by investigating various distance metrics, this work then suggests robust techniques to tackle the issue of determining the number of clusters. In the suggested methods, two novel metrics are presented to measure the partitioning quality in terms of the separation between two distributions: one distribution for distances between segments of the same speakers and one for the distances between segments of different speakers. Such techniques have been evaluated on the RT07s NIST Rich Transcription evaluations for meetings data sets and competitive performance is achieved, without the need to learn the threshold for estimating the number of clusters as in conventional state-of-the-art systems. Finally, multi-stream speaker clustering approach is studied with the emphasis on assessing the relative significance of each individual stream and as a result, an adaptive weighting scheme for each feature stream is suggested. This adaptive weighting scheme is then shown to perform better than fixed weighting scheme, with the additional benefit of no training data is required to determine the weights. The complete systems: one for single channel and one for multiple channels were submitted to the RT09s NIST Rich Transcription evaluations and achieved the first rank in the speaker diarization category.

Acknowledgments

I would like to express my sincere thanks and appreciation to my supervisor, Dr. Chng Eng Siong (NTU), and co-supervisor, Dr. Li Haizhou (I²R). I have benefited tremendously from their invaluable advices. Their assistance allows me to pursue my interests and explore new research areas. I am very grateful for their constant encouragement, their perspectives and especially their patience and tolerance on my stubbornness.

I also want to thank my colleagues Ma Bin and Kong Aik for their understanding and support at work to let me spend time pursuing my own research. Also to my friends Wang Lei and Xiao Xiong whose discussions though not many but are really enjoyable and helpful. To Hanwu and Tin Lay who contributed to the research platform which I am tremendously benefited from.

Lastly, this work could not have been achieved without the love and support of my family, my dad who instills into me the persistence, my mum who cares for me anytime and my fiancée who loves me and brightens up my life.

Contents

Abstract	i
Acknowledgments	iii
List of Figures	vii
List of Tables	ix
List of Abbreviations	xi
Mathematical Notation	xiii
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	3
1.3 Contributions	3
1.4 Overview	5
2 Fundamentals and previous works	6
2.1 Signal processing	7
2.1.1 Wiener filter	7
2.1.2 Acoustic beamforming	8
2.2 Feature extraction	9
2.2.1 Acoustic features	9
2.2.2 Time delay of arrival	12
2.2.3 Feature normalization techniques	14
2.3 Speech activity detection	15
2.3.1 Energy-based speech detection	15
2.3.2 Model based speech detection	16
2.3.3 Hybrid speech detection	16

2.3.4	Multi-channel speech detection	18
2.4	Clustering architecture	19
2.4.1	Speaker modeling	20
2.4.2	Distance measures	25
2.4.3	Speaker segmentation	32
2.4.4	Speaker clustering	36
2.5	Summary	46
3	Speaker diarization systems in meetings	48
3.1	NIST Rich Transcription Evaluation in Meetings	48
3.1.1	Meeting room setup	49
3.1.2	Databases for speaker diarization in meetings	50
3.1.3	Speaker diarization system evaluation	51
3.1.4	Related Projects in Meeting Room	53
3.2	State of the art	53
3.3	Baseline speaker diarization systems	56
3.3.1	Single distance microphone speaker diarization system	56
3.3.2	Multiple distant microphones speaker diarization system	62
3.4	Experiments	66
3.5	Summary	68
4	Analysis of distance measures	69
4.1	Analysis of GLR-based metrics	70
4.1.1	Motivation	70
4.1.2	Generalized Likelihood Ratio	71
4.1.3	Bayesian information criterion	74
4.1.4	Information change rate	74
4.2	Proposed distance measure	75
4.3	Experiments	76
4.3.1	Experiments on distance measures	76
4.3.2	Distance measures and speaker diarization performance	82
4.4	Summary	84

5	Robust techniques for speaker diarization	85
5.1	Stopping criteria	85
5.1.1	Proposed criteria	86
5.1.2	Application of the proposed criteria in speaker diarization system	90
5.2	Spectral subspace	91
5.2.1	Spectral clustering algorithm	92
5.2.2	Spectral clustering algorithm with proposed metrics	93
5.3	Automatic stream weights estimation	95
5.4	Experiments	98
5.4.1	Determine number of clusters in speaker diarization systems . . .	98
5.4.2	Distance metrics in spectral subspace	103
5.4.3	Determine number of clusters in spectral subspace	105
5.4.4	Adaptive stream weights in multichannel speaker diarization system	106
5.5	Summary	107
6	Conclusions and Future Work	108
6.1	Contributions	109
6.1.1	Distance metric	109
6.1.2	Stopping criteria	110
6.1.3	Adaptive feature stream weighting	111
6.2	Future Work	112
	Publications	115
	References	116

List of Figures

2.1	<i>Typical components of a speaker diarization system</i>	7
2.2	<i>Hybrid energy-based and model-based speech activity detector</i>	18
2.3	<i>Hierarchical clustering architecture</i>	19
2.4	<i>Sequential clustering architecture</i>	19
3.1	<i>Typical structure of the state-of-the-art single channel speaker diarization systems</i>	57
3.2	<i>HMM topology for speaker clustering with duration constraint</i>	61
3.3	<i>Typical structure of the state-of-the-art multichannel speaker diarization systems</i>	63
3.4	<i>Speaker diarization performance of SDM system with various GLR stopping thresholds.</i>	67
3.5	<i>Speaker diarization performance of MDM system with various GLR stopping thresholds.</i>	68
4.1	<i>Average inter-speaker distances and intra-speaker distances with different cluster sizes using GLR and BIC metric</i>	79
4.2	<i>Average inter-speaker distances and intra-speaker distances with different cluster sizes using ICR and d_s metric</i>	80
4.3	<i>DET for various distance metrics evaluated on set $\{S_5 \cup S_{10} \cup S_{15}\}$. . .</i>	81
4.4	<i>Performance of baseline speaker diarization system with various clustering distance measures</i>	83
5.1	<i>Meetings with highly varied optimal stopping thresholds</i>	87
5.2	<i>Clusters with different densities</i>	88

5.3	Illustration of T_s and ρ values for different distributions of D_{intra} and D_{inter}	89
5.4	Determine number of clusters with ρ criterion	92
5.5	Illustration of features stream weights for different distributions of s_{intra} and s_{inter} . The red line indicates the projection hyperplane to achieve maximum separation	98
5.6	Speaker diarization performance of SDM system with various BIC stopping thresholds.	99
5.7	Speaker diarization performance of SDM system with various NCLR stopping thresholds.	100
5.8	Speaker diarization performance of SDM system with various ICR stopping thresholds.	100
5.9	Speaker diarization performance of SDM system with various d_s stopping thresholds.	101
5.10	EER(%) of different distance metrics when projected to the spectral subspace with M being the number of neighbors used in spectral clustering algorithm	104
5.11	EER(%) of different distance metrics when projected to the spectral subspace with varying number of eigenvalues	105

List of Tables

3.1	<i>Conference speech databases. N: number of meeting excerpts. T: average length of each meeting excerpt in seconds</i>	50
3.2	<i>Summary of data sources in terms of number of speakers. S: number of speakers. N: number of meeting excerpts</i>	50
3.3	<i>Summary of data sources in terms of recording sites. C: recording sites. N: number of meeting excerpts</i>	50
3.4	<i>Performance of Speech/Non-speech Detector of SDM system</i>	67
3.5	<i>Performance of Speech/Non-speech Detector of MDM system</i>	67
4.1	<i>EER(%) of different speaker distance measures</i>	78
4.2	<i>DER (%) of the baseline speaker diarization system with various clustering distance measures</i>	83
5.1	<i>DER (%) of the SDM speaker diarization system with d_s as distance metric and T_s as stopping criterion</i>	99
5.2	<i>DER (%) of the SDM speaker diarization system with d_s as distance metric and ρ as stopping criterion</i>	99
5.3	<i>DER (%) of the baseline SDM speaker diarization system with various clustering distance measures. The stopping threshold was tuned on RT2006 database. The value in the parenthesis is the optimal threshold obtained from the development data.</i>	102
5.4	<i>DER (%) of the baseline SDM speaker diarization system with various clustering distance measures. The stopping threshold was tuned on RT2007 database. The value in the parenthesis is the optimal threshold obtained from the development data.</i>	102

5.5	<i>DER (%) of the baseline SDM speaker diarization system with d_s as distance metric and oracle stopping criterion</i>	102
5.6	<i>DER (%) of the SDM speaker diarization system with d_s as distance metric and T_s as stopping criterion in spectral subspace</i>	105
5.7	<i>DER (%) of the SDM speaker diarization system with d_s as distance metric and ρ as stopping criterion in spectral subspace</i>	106
5.8	<i>DER (%) of the MDM speaker diarization system with d_s metric and proposed T_s criterion in spectral subspace</i>	107

List of Abbreviation

acp	Average Cluster Purity
ADM	All Distant Microphone
AEP	Asymptotic Equipartition Property
AHC	Agglomerative Hierarchical Clustering
AHS	Arithmetic Harmonic Sphericity
aIB	Agglomerative Information Bottleneck
asp	Average Speaker Purity
ASR	Automatic Speech Recognition
BIC	Bayesian Information Criterion
CCR	Cluster Complexity Ratio
CLR	Cross Likelihood Ratio
CMN	Cepstral Mean Normalization
CSPG	Constant Seconds per Gaussian
DER	Diarization Error Rate
DET	Detection Error Tradeoff
DOA	Direction of Arrival
DPMM	Dirichlet Process Mixture Model
DSD	Divergence Shape Distance
e-HMM	Evolutionary Hidden Markov Model
EER	Equal Error Rate
eHMM	Ergodic Hidden Markov Model
EM	Expectation Maximization
EVSM	Eigen Vector Space Model
FAR	False Alarm Rate
FER	Frame Error Rate
FFT	Fast Fourier Transform
GCC	Generalized Cross Correlation
GCC-PHAT	Generalized Cross Correlation with Phase Transform
GLR	Generalized Likelihood Ratio
GMM	Gaussian Mixture Model
GXPSD	Generalized Cross Power Spectral Density
HDP-HMM	Hierarchical Dirichlet Process Hidden Markov Model
HMM	Hidden Markov Model

IAC	Iterative Agglomerative Clustering
IAHC	Iterative Agglomerative Hierarchical Clustering
ICR	Information Change Rate
IHM	Individual Head Microphone
i.i.d.	Independent and Identically Distributed
ILP	Integer Linear Programming
KL	Kullback-Leibler
KL2	Symmetric Kullback-Leibler
LDA	Linear Discriminant Analysis
LLR	Log-Likelihood Ratio
LFCC	Linear Frequency Cepstrum Coefficient
LPC	Linear Predictive Coding
LPCC	Linear Prediction Cepstral Coefficient
LSP	Line Spectrum Pair
MAP	Maximum <i>A Posteriori</i>
MDM	Multiple Distant Microphones
MDR	Miss Detection Rate
MFCC	Mel Frequency Cepstral Coefficient
ML	Maximum Likelihood
NCLR	Normalized Cross Likelihood Ratio
NLLR	Normalized Log-Likelihood Ratio
PCA	Principal Component Analysis
PLPC	Perceptual Linear Prediction Cepstral
PRC	Precision
RCL	Recall
RMS	Root Mean Square
SAD	Speech Activity Detection
SASTT	Speaker Attributed Speech to Text
SCOT	Smoothed Coherence Transform
SER	Speech Activity Detection Error Rate
SDM	Single Distant Microphone
SID	Speaker Identification
SNR	Signal to Noise Ratio
SOM	Self Organizing Map
STT	Speech to Text
SVM	Support Vector Machine
TDOA	Time Delay of Arrival
UBM	Universal Background Model
VB	Variational Bayesian
VQ	Vector Quantization
ZCR	Zero Crossing Rate

Mathematical Notation

$\mathbf{x}$	Vector $\mathbf{x}$
$\mathbf{A}$	Matrix $\mathbf{A}$
$ \mathbf{A} $	Determinant of matrix $\mathbf{A}$
$\text{tr}(\mathbf{A})$	Trace of matrix $\mathbf{A}$
$(\cdot)'$	Matrix or vector transpose

Chapter 1

Introduction

1.1 Motivation

Meetings domain is a rich source of content for spoken language research and technology. For example, using meeting data one can generate rich transcription (transcription including punctuation, capitalization, and speaker labels), perform transcript indexing and summarization. To develop these applications successfully, the availability of high quality *automatic speech recognition* (ASR) transcripts is therefore required, as such, ASR in meetings is an important and active area of investigation. In addition, because of the presence of multiple participants in these meetings, it is not only important to determine what was said, but who said it. Accurate speaker diarization i.e., determining “Who spoke when?” is therefore also of great importance to spoken language processing in meetings . The goal of speaker diarization is to divide a continuous audio recording into regions and annotate each region with a speaker label. The labels are not necessarily the actual speaker identities as long as the same labels are assigned to the regions uttered by the same speakers. The regions may also overlap as multiple speakers may talk simultaneously. Speaker diarization is essentially the combination of two different processes: segmentation, in which the speaker turns are detected, and unsupervised clustering, in which segments of the same speakers are grouped. The clustering process is considered as unsupervised problem since there is no prior information about the number of speakers, their identities or acoustic conditions [1,2].

Some applications of speaker diarization include:

- **Speaker adaptation** [3]: in ASR system, speaker-dependent models perform better than speaker-independent models when the conditions are matched (e.g. the speaker is in the training set). Speaker diarization could aid in adapting the speaker-independent models to speaker specific data and the recognition accuracy could be improved.
- **Speaker and audio indexing** [4,5]: with the explosive growth of digital archives of multimedia materials, indices are essential part of archives for quick browsing and speaker diarization enables the data to be indexed, browsed or searched by speaker.
- **Rich transcription** [2,6,7]: for traditional automatic speech recognition (ASR), the system produces a stream of symbols (words, phonemes, etc.) given an input audio. This may not be elegant and easy to understand if the audio contains long paragraphs or dialogues. Speaker diarization could help to improve the readability of automatic transcripts by segmenting into speaker turns and identities.

Significant effort has been made in speaker diarization research, especially in *rich transcription* (RT) domain through many international RT benchmarks, which were pioneered and sponsored by the National Institute of Standards and Technology (NIST) [8]. From 2000 to 2009, NIST held eight RT evaluations to encourage research in ASR and speaker diarization. More importantly, NIST has provided common datasets and standard experimental protocols to allow different systems with different approaches to be meaningfully compared and to push the state-of-the-art forward. Accordingly, in this thesis, special emphasis will be placed on systems and techniques which have been proven to work within the context of NIST RT evaluations, which is a reliable indicator for the current state-of-the-art in speaker diarization. In this context, several approaches to speaker diarization have been proposed and evaluated on RT databases, however agglomerative hierarchical clustering is by far the most popular and is at the core of state-of-the-art systems [9–11]. Despite the promising results, many systems in the RT09 benchmark comprise parameters which are critical to the diarization performance, and typically these parameters were learned on development or training data, which may or may not be optimal for the current data. Therefore, the major objectives of this thesis are to

overcome these limitations and to improve the robustness of the system in case of unseen data.

1.2 Objectives

The main objective of this work is to develop robust techniques for speaker diarization with focus on meetings domain. More concrete objectives are:

- Although in the scope of this thesis, speaker diarization systems are implemented to target meetings domain, the algorithms and techniques examined should not be restricted to only meetings and could be applied to any other domains.
- The performance of resulting system should be competitive with focus on robustness. The robustness of the system is implied in terms of: less manually tuned parameters, less dependence on prior knowledge (training data, room configuration, sensor's positions, etc.).
- The work is limited to offline speaker diarization architectures.

1.3 Contributions

The techniques presented in this thesis are to address the robustness issues and to improve the performance of current state-of-the-art speaker diarization systems. The approaches taken are to a great extent relying on unsupervised learning methods to reduce the number of finely tuned parameters, as such robustness could be enhanced in the process. For the starting point, the state-of-the-art systems [11, 12] with *iterative agglomerative hierarchical clustering* (IAHC) framework was investigated, potential weaknesses and refinements were studied. From these studies, the thesis then proposed techniques to address three issues targeting different modules of the systems. The three identified subjects are: (1) *selecting appropriate distance metric for clustering*, (2) *robust techniques to determine number of clusters*, and (3) *adaptive weighting for multiple feature streams fusion*.

Distance metrics [B] As opposed to their popularity in state-of-the-art speaker diarization systems, *Generalized Likelihood Ratio* (GLR) [13, 14] based metrics including GLR and *Bayesian Information Criterion* (BIC) [11, 15] are proved to be linearly proportional to the cluster size, and is thus not robust against data size variation. A novel distance metric is then proposed to minimize such effects by normalizing with the number of data frames. Not only so, this metric is probably the first, in the context of distance measure, to exploit the second-order statistics from the log-likelihood scores to further improve the speaker discriminability. Experiments taken on a large scale text-dependent speaker verification task clearly demonstrate the effectiveness of the proposed technique. When being employed in a speaker diarization system, improvements are also observed in terms of reduction in the overall *diarization error rate* (DER).

Stopping criteria [C,E] The common issue encountered in hierarchical clustering framework is to decide when the merging process should stop. Threshold method and model selection approach are shown to be data dependent, and as such are not robust against data mismatch. This work therefore proposes two novel criteria for measuring the clustering partition quality at each iteration of merging. As shown in the thesis, these criteria are not monotonically increasing or decreasing, instead they tend to peak at the correct number of clusters and thus no prior threshold is required. These criteria are then further investigated in the spectral subspace, which has some nice properties of making the cluster well separated. It is shown that in this subspace, the proposed criteria exhibit sharper peaks at the correct number of clusters and therefore increasing the detection performance. Experiments on Rich Transcription speaker diarization tasks demonstrated the robustness of the proposed technique across multiple data sets.

Adaptive stream weighting [A, D] In the final study, the robustness of multi-channel speaker diarization system is investigated, particularly the issue of fusing multiple feature streams. A novel strategy to adaptively determine the usefulness of each feature stream is proposed. The strategy assigns to each stream a weight, to indicate the contribution of that stream in the fusion and the weights are optimized based on Fisher criterion, which is used to measure the separation between clusters in terms of the ratio of the variance

between clusters to the variance within the clusters. In other words, the algorithm values the features which is more cohesive (small variance within the same classes) and is more discriminative (large variance between different classes). The novelty of this approach is two folds. Firstly the stream weights are selected such that the resulting clusters are better separated. Secondly the weights are evaluated and updated in each iteration of the clustering phase as opposed to a fixed weighting scheme, thus they are better at adapting to the changes in the clusters.

These individual techniques presented in this thesis are parts of the complete IIR-NTU speaker diarization systems [D] which are the best systems [16] in the NIST Rich Transcription 2009 Evaluation workshop [17–19] in the topic of speaker diarization.

1.4 Overview

This thesis is organized as follows:

In chapter 2, the fundamentals and background information on speaker diarization is summarized. Each topic would be covered in details together with the most significant work related to each topic.

Chapter 3 introduces the topic of speaker diarization in the meeting domain and analyzes state-of-the-art speaker diarization systems with discussion on the limiting factors that might affect the performance of such systems. This chapter also presents the baseline systems and their performances. These systems will serve as the reference for subsequent chapters in the thesis.

Chapter 4 addresses an important issue in speaker clustering, which is the distance metric to measure the similarity between speakers or clusters. The chapter first analyzes the robustness of the widely used GLR-based metrics, then proposes a novel distance metric and concludes the chapter with comparative studies of these metrics.

Chapter 5 investigates various robust techniques in speaker diarization systems including: criteria to estimate number of speakers, spectral subspace and automatic feature stream weighting.

Finally, Chapter 6 concludes the thesis with summary of the contributions and discussion about the future direction of research.

Chapter 2

Fundamentals and previous works

This chapter presents the fundamentals of speaker diarization and the most significant works over the recent years on this topic. Figure 2.1 shows the typical components of a speaker diarization system. The *signal processing* module applies standard techniques such as: pre-emphasis, noise reduction and/or beamforming to improve the signal-to-noise ratio (SNR) and to reduce undesired noises. The *feature extraction* module transforms the raw audio signal into feature vectors in which speaker-related characteristics are captured and unintended properties such as noises are suppressed. Subsequently, only useful feature vectors are retained for further processing. These vectors are generally corresponding to speech frames and the selection of these frames is implemented in the *speech activity detection* (SAD) module. Depending on the implementation and type of SAD module, it may be executed before feature extraction, e.g. energy-based SAD, where only the energy level is required to detect speech/non-speech. More sophisticated SAD, however, is typically performed after feature extraction since additional features are required to be extracted for better detection of speech and non-speech. Finally, at the heart of a speaker diarization system is the *clustering architecture* which defines the strategies and approaches to perform speaker clustering from the unlabeled feature vectors. These important components will be covered thoroughly in the upcoming sections.

Figure 2.1: *Typical components of a speaker diarization system*

2.1 Signal processing

2.1.1 Wiener filter

We are given two processes: s_k the signal to be estimated, and y_k the observed noisy signal, which are jointly wide-sense stationary, with known covariance function $R_s(k)$, $R_y(k)$, and $R_{sy}(k)$. A particular case is that of a signal corrupted by additive noise n_k :

$$y_k = s_k + n_k \quad (2.1)$$

The problem is to estimate the signal s_k as a function of y_k . A widely used solution to this problem is *Wiener filtering* [20]. Wiener filtering is used to produce an estimate of desired signal by linear time-invariant filtering an observed noisy signal, assuming known stationary signal and noise spectra, and additive noise. Wiener filtering gives the optimal way of filter out the noisy components, so as to give the best L^2 -norm reconstruction of the original signal. Interested readers may refer to [20] for the solutions.

In the speech community, the Qualcomm-ICSI-OGI front end [21] is commonly used to perform Wiener filtering. Most state-of-the-art speaker diarization systems [10, 11, 15, 22] applied Wiener filtering to all audio channels for speech enhancement before filtered and

summed to produce a beamformed audio channel. In Van Leeuwen and Konecny [23], the filtering, however, was applied after beamforming. The authors observed no difference in performance with the benefit of reduction in computational cost since only one channel was filtered.

2.1.2 Acoustic beamforming

In a typical meeting room environment, there will have two or more distant microphones. Each microphone can have different characteristics, thus the recorded audio quality varies across microphones. For speaker diarization, one may select the best quality microphone, for e.g. highest *signal to noise ratio* (SNR), and work on this selected channel as traditional single channel diarization system. However, a more widely adopted approach is to perform *acoustic beamforming* on multiple audio channels to derive a single enhanced channel and proceed from there. The techniques for acoustic beamforming is a broad field of research on its own. Nevertheless, in the literature of speaker diarization, simple beamforming techniques, particularly delay and sum beamforming and its variations, are the preferred solutions.

Given an array of M microphones, the delay and sum output z at instance k is defined as [24]:

$$z[k] = \frac{1}{M} \sum_{m=1}^M y_m[k + \Delta_m[k]] \quad (2.2)$$

where $y_m[k]$ is the signal for each channel and $\Delta_m[k]$ is the relative delay between each channel and the reference channel at instance k . In practice, many systems adopt the weighted delay and sum beamforming suggested in Anguera et al. [25]:

$$z[k] = \sum_{m=1}^M W_m[k] y_m[k + \Delta_m[k]] \quad (2.3)$$

where $W_m[k]$ is the relative weight of microphone m at instance k with $\sum_{m=1}^M W_m[k] = 1$. The *time delay of arrival* (TDOA) $\Delta_m[k]$ is commonly estimated via cross-correlation methods such as *generalized cross correlation with phase transform* (GCC-PHAT) [26]. For researchers who are not in the field of array processing, there are several freely available implementations of beamforming source code on the web and one such popular

toolkit used by many researchers in the recent NIST 2009 Rich Transcription evaluation benchmark is known as BeamformIt [27]. The toolkit implemented an enhanced delay-and-sum algorithm, together with many multichannel techniques described in [28] which are rather relevant to speaker diarization research. The techniques include: automatic selection of reference microphone channel for GCC-PHAT computation, adaptive weighting of channel based on SNR or cross-correlation metric, and two-pass Viterbi decoding for smoothing spurious TDOA values. These techniques are applied to stabilize the TDOA values before the signals are beamformed.

2.2 Feature extraction

Raw speech signal is normally converted into a sequence of feature vectors carrying characteristic information about the signal; this step is referred to as feature extraction. In the field of speaker diarization, as well as speaker recognition in general, the information that we want to retain is the speaker-related properties. Many types of features have been studied in the literature; commonly used features for speaker diarization could be broadly organized into two categories: (1) *acoustic features*, and (2) *sound-source features*.

2.2.1 Acoustic features

Speech is produced when air is forced from the lungs through the vocal cords and along the vocal tract. Different speech sounds are generated by varying the shape of the *vocal tract* and its *mode of excitation*. The variations occur relatively slowly in the order of 20 ms. Thus, for short frames of 20 – 30 ms, speech signal can be considered to be *quasi-stationary* and the *short-term features* are extracted to model the shape of the vocal tract or the excitation or the combination characteristics of both.

2.2.1.1 Short-term spectral features

Short-term spectral features are based on the *spectral envelope*, the shape of the discrete Fourier transform (DFT) magnitude spectrum, of a short time *windowed frame* of speech (typically 20 – 30 ms). The effectiveness of these features are based on the observations that:

- The phonetic segments in speech appears as energy fluctuation over time in different frequency bands. This is useful for representing the phonetic contents in speech recognition.
- The spectral envelope contains information about the resonance properties of the vocal tract which depends on both phonetics and speakers. This is the most informative part of the spectrum in speaker recognition.

Popular spectral features are Mel Frequency Cepstral Coefficients (MFCC) [29], Linear Prediction Cepstral Coefficients (LPCC) [30] and Perceptual Linear Prediction Cepstral (PLPC) Coefficients [31]. These features differ mainly in the *analysis of time-frequency* and in the techniques for *frequency smoothing*.

In frequency analysis, an important notion is *critical-band* [32] which refers to the capability of human auditory system to localize and process information within the frequency range. The detection of signals within this frequency range is not sensitive to and less affected by interference signals outside of this critical bandwidth. This bandwidth is non-linearly dependent on frequency. A number of functions approximating these critical bandwidths were proposed, among which, two popular functions are *Bark scale* [33] and *Mel scale* [34]. MFCC filter bank follows Mel frequency scale, whereas in PLPC, the spectrum is filtered by a trapezoidal-shaped filter bank with Bark frequency scale.

Another differential factor among various spectral features is the frequency smoothing techniques being used. The frequency smoothing techniques are generally applied to enhance and preserve the *formant* information, which are the pattern of resonances and can be observed from the spectral envelope. To capture this pattern, the spectral envelope in MFCC features is derived from the FFT *power spectrum*, while in LPC and PLPC, the spectrum is approximated by a linear predictor with *all-pole model*.

For MFCC, LPCC and PLPC feature extraction, in the final step the spectral representation is transformed to *cepstral coefficients*, in which the coefficients are nearly *orthogonal*. This property is desirable as it is beneficial for modeling purpose and leads to significant reduction in the number of parameters to be estimated. Particularly, when using *Gaussian* or *Gaussian mixtures* model, diagonal covariance matrices with *uncorrelated* components could be used instead of full covariance matrices.

Given these alternative features, however, MFCCs, sometimes with their first and/or second derivatives, are more widely adopted in the community of speaker diarization researches. In contrast to speech recognition, higher order cepstral coefficients are retained since they capture more speaker-specific information, yet there is no consensus on the order of MFCCs. Typically, 16 to 20 coefficients are used in most of the current state-of-the-art diarization systems [11, 35–37]. Nonetheless, Ajmera and Wooters [38] reported the diarization results using both LPCC and MFCC features. They observed that LPCCs perform better during clean speech, while MFCCs work better in case of noisy conditions. In another attempt, Wooters et al. [39] compared the performance of MFCC and PLP features, with the empirical evidence that MFCCs slightly outperform PLPs.

Apart from these common spectral features, some lesser known features were also explored for speaker diarization task in the literature. The LIA system [40, 41] performed speaker segmentation using 20th order *linear cepstral features* (LFCC) augmented by the energy.

2.2.1.2 Prosodic features

Prosodic features are supra-segmental, they are not confined to any one segment, but occur in some higher level of an utterance. Prosodic units are marked by phonetic cues including pause, pitch, stress, volume, and tempo. The most important prosodic feature is the *fundamental frequency* ($F0$). Other common prosodic features are: duration, speaking rate, and energy distribution/modulations [42]. Combining prosodic features with spectral features has been shown to be effective for speaker verification, especially in noisy condition. Recently, these features have been adopted in several speaker diarization systems and showed promising results. In El-Khoury et al. [43], a difference between the averages of the $F0$ between speech segments was calculated and used as merging criterion for bottom-up clustering. In Friedland et al. [44], the authors investigated the speaker discriminability of 70 long-term features, most of them prosodic features. They applied *Fisher Linear Discriminant Analysis* (LDA) to rank these 70 prosodic and long-term features by their speaker discriminative power. The authors showed improvement in speaker diarization results when combining the top-ten ranked prosodic and long-term

features with regular MFCCs. In a recent paper, Imseng and Friedland [45] proposed the use of prosodic features to obtain initial clusters, which are crucial in many state-of-the-art agglomerative speaker diarization systems. The proposed approach achieved significant improvement over the baseline system.

2.2.2 Time delay of arrival

When multiple microphone recordings are accessible, the relative *time delay of arrival* (TDOA) between the different microphones can be estimated. Assuming the speakers are not changing position, those features can be used in speaker diarization [46]. It has been shown that TDOA improves the speaker diarization significantly in combination with conventional spectral features [47]. There have been many techniques proposed in the past for estimating the TDOA, and *generalized cross correlation with phase transform* (GCC-PHAT) [26] is one the most commonly used method.

Given a pair of microphones i and j , let $x_i(n)$ and $x_j(n)$ be the windowed signals from microphone i and j respectively. The cross-correlation between the two signals is defined as

$$R_{x_i x_j}(\tau) = \mathbf{E} [x_i(n) \cdot x_j^*(n - \tau)] \quad (2.4)$$

where $\mathbf{E}[\cdot]$ denotes the expectation. In practice, it is estimated as

$$R_{x_i x_j}(\tau) = \frac{1}{2N} \sum_{n=-N}^N x_i(n) \cdot x_j^*(n - \tau) \quad (2.5)$$

where N is the length of the windowed signals (in terms of number of samples), for reliable estimation, the window size is typically at least 500ms. It is generally assumed that the signals picked up by the two microphones i and j are similar with one being the delayed version of the other by a time T_{ij} . T_{ij} is then estimated by maximizing the cross-correlation function

$$T_{ij} = \arg \max_{-N \leq \tau \leq N} R_{x_i x_j}(\tau) \quad (2.6)$$

In real applications, however, there are many external factors such as ambient noises, reverberation etc. that could affect the estimation of time delay and it is shown that

cross-correlation is not robust against these issues. To address this problem, Knapp and Carter [26] introduced a general version named the Generalized Cross-Correlation (GCC), which is defined as

$$R_{x_i x_j}(\tau) = \mathbf{E} \left[\left(h_i(n) * x_i(n) \right) \cdot \left(h_j(n - \tau) * x_j^*(n - \tau) \right) \right] \quad (2.7)$$

where $h_i(n)$ and $h_j(n)$ are the filter coefficients. It is apparent from the GCC equation that it is simply the cross-correlation computed on the filtered signals. Generally, GCC for long windowed signals is computed in the frequency domain for efficiency. The generalized cross power spectral density (GXPSD) [26] can be expressed as

$$\Phi_{x_i x_j}(k) = [H_i(k)X_i(k)] \cdot [H_j(k)X_j(k)]^* \quad (2.8)$$

where X_i , X_j , H_i , H_j are the discrete Fourier transform of x_i , x_j , h_i , and h_j correspondingly, with k being the discrete frequency index. Rearranging the above equation

$$\Phi_{x_i x_j}(k) = H_i(k)H_j^*(k)X_i(k)X_j^*(k) \quad (2.9)$$

$$= \Psi_{ij}X_i(k)X_j^*(k) \quad (2.10)$$

where

$$\Psi_{ij} = H_i(k)H_j^*(k)$$

being the weighting function. Various weighting functions have been studied in the literature including: *Roth filter* [48], *the smoothed coherence transform* (SCOT) [49], *the phase transform* (PHAT) [26], *the Eckart filter* [50], and *the Hannon and Thomson filter* [26].

In the scope of this thesis, PHAT will be used as weighting function since it is widely adopted and is shown to be robust against a wide range of conditions [26]. PHAT is defined as

$$\Psi_{ij}^{PHAT}(k) = \frac{1}{|X_i(k)X_j^*(k)|} \quad (2.11)$$

In summary, the time delay T_{ij} between microphone i and j can be estimated as

$$\Phi_{x_i x_j}(k) = \frac{X_i(k)X_j^*(k)}{|X_i(k)X_j^*(k)|} \quad (2.12)$$

$$R_{x_i x_j}(\tau) = \mathcal{F}^{-1}[\Phi_{x_i x_j}(k)] \quad (2.13)$$

$$T_{ij} = \arg \max_{-N \leq \tau \leq N} R_{x_i x_j}(\tau) \quad (2.14)$$

with $\mathcal{F}^{-1}$ denoting the inverse Fourier transform.

In a typical meeting room setup, there are usually more than two microphones. Thus, at a particular time frame, it is possible to extract multiple TDOA values, one for each microphone pair. These values are then concatenated to form a vector of TDOA, and in this thesis, these vectors are referred to as delay features. These features are often used in conjunction with acoustic features (e.g. MFCC) to augmented the performance of the speaker diarization system.

2.2.3 Feature normalization techniques

RASTA filtering RASTA filtering [51] is mainly applied to removes slow channel variations. It is equivalent to a band-pass filtering of each frequency channel through an IIR filter with the transfer function:

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4} \times (1 - 0.98z^{-1})} \quad (2.15)$$

Filtering could be performed in either log spectral or cepstral domain. In the cepstral domain, the low and high cut-off frequency define the frequency range in which the cepstral change within this range is preserved. RASTA is applied on the MFCC features before estimating speaker models in the MIT Lincoln Laboratory diarization systems [52].

Cepstral Mean Normalization *Cepstral Mean Normalization* (CMN) is typically employed to minimize the effect of session variations, which occur with the change of channel characteristics. It is calculated by first estimating the cepstral mean across an utterance or a window of N frames and then subtracting the mean from each cepstral vector to obtain the normalized vector. As a result, the long-term average of any observation sequence (the first moment) is zero. When the audio stream is processed online, a dynamic CMN approach is applied, where the cepstral mean μ_t at time t is updated as follows:

$$\mu_t = \alpha C(t) + (1 - \alpha)\mu_{t-1} \quad (2.16)$$

where α is a time constant (typically, around 0.001), $C(t)$ is the cepstral vector at time t and μ_{t-1} is the dynamic cepstral mean at time $(t - 1)$. In Reynolds and Torres-Carrasquillo [52], MFCC features for each cluster is processed with CMN to increase robustness against channel distortion in their offline speaker diarization systems. While in Zamalloa et al. [53], the dynamic CMN approach is applied in their online speaker tracking system.

Feature warping In order to avoid the influence of background noises and other non-speaker related events, *feature warping* is proposed to condition and conform the individual feature streams such that they follow a specific target distribution over a window of speech frames. Normally, the target distribution is chosen to be following Gaussian shape [54]. In the context of speaker diarization, Sinha et al. [55] and Zhu et al. [56] apply this normalization technique for each short segment using a sliding window of 3 seconds in the clustering stage.

2.3 Speech activity detection

Speech activity detection (SAD) identifies audio regions containing speech from any of the speakers present in the recording. Depending on the domain of the data being used, the non-speech regions may contain silence, laughing, music, room noise, or background noise. The use of a speech/non-speech detector is an important part of speaker diarization system. The inclusion of non-speech frames into the clustering process makes it difficult to correctly differentiate between two speaker models. SAD could be broadly classified into four categories: (1) *energy-based speech detection*, (2) *model based speech detection*, (3) *hybrid speech detection*, and (4) *multi-channel speech detection*.

2.3.1 Energy-based speech detection

Many energy-based speech detectors are proposed in the literature, however, with the diverse environments of audio recordings, the non-speech can be from a variety of noise sources, like paper shuffling, coughing, laughing, etc. energy-based methods have shown to be relatively ineffective in speaker diarization task [57, 58]. Nevertheless, with its

simplicity and speed, this approach has been adopted in several systems. In [59], Cassidy defines a threshold based on *root mean square* (RMS) and *zero crossing rate* (ZCR) of the audio signal to separate speech and silence.

2.3.2 Model based speech detection

With the limitation of energy-based approach, in general, *model based* speech/non-speech detectors are frequently used in many speaker diarization systems as they are able to characterize various acoustic phenomena. The simplest system uses just two models for speech and non-speech such as in Wooters et al. [39]. A more complex system is described in Nguyen et al. [60] with four speech models including gender/bandwidth combinations. Noise and music are explicitly modeled in Gauvain et al. [61], and Zhu et al. [62]; the systems comprise of five classes: speech, music, noise, speech + music, and speech + noise. The speech + music and speech + noise models are used to help minimize the false rejection of speech occurring in the presence of music or noise, and this data is subsequently reclassified as speech [55, 61–63]. The classes can be broken down further, as in Liu and Kubala [64], there are five models for non-speech (music, laughter, breath, lip-smack, and silence) and three for speech (vowels and nasals, fricatives, and obstruents). In Meignier et al. [2], the acoustic segmentation system are designed in a hierarchical approach to provide finer classification. First, speech/non-speech is detected then the speech class is further classified as clean speech, speech with music, and telephone speech. Each category is subsequently separated by gender and two additional models representing female and male speech recorded under degraded conditions are then included to refine the final segmentation.

2.3.3 Hybrid speech detection

The model-based approach, however, has its own limitation: its models need to be trained with pre-labeled data using training set. This requires the data to be annotated with class labels and this process takes much effort. The performance of these models on unseen data is also an important issue especially in the case where testing data is substantially different from development data. In the framework of NIST Rich Transcription Evaluation for the meeting domains, each meeting is generally recorded from different

meeting rooms at different sites. Thus, the environment, ambiance and room acoustics are expected to vary across recordings. In such scenario, training a model-based *speech activity detection* (SAD) is difficult and it tends not to work well for all meetings as reported in [23]. A better solution would be hybrid speech/non-speech detectors, which have been widely used in many state-of-the-art systems.

The simplest form of hybrid SAD is a combination of energy-based detector and model-based decoder, as was first suggested in [65] and later used in [9, 10, 22, 66]. In the first stage, an energy-based detector finds all segments with low energy, while applying minimum segment duration. The low-energy segments and high-energy segments are used to train the initial non-speech and speech model respectively. Then several iterations of Viterbi segmentation and models retraining are performed to produce the final speech/non-speech segments. Figure 2.2 illustrates the described procedures. In Anguera et al. [65, 67], the authors first use a derivative filter in combination with a finite state machine(FSM) to detect speech and non-speech regions. These initial labels are then used to build two HMMs for speech/non-speech; the system iteratively segments and trains both models until the overall likelihood stops increasing. In some systems, not all frames are utilized to build the initial models and some additional constraints may be imposed to ensure the correctness of selected frames. For e.g. Sun et al. [9] selected 10 percent frame features with the highest energies and relative low zero-crossing rates for the speech modeling, while the 20 percent frame features with the lowest energies and relative higher zero-crossing rates were selected for non-speech modeling.

The drawback of using energy however is that it is not possible to use this approach when the audio contains fragments with high energy levels that are non-speech. More complex hybrid SAD detectors were also suggested in [11, 15], where a pre-trained model-based detector was first used to create an initial segmentation of speech and silence region. The energy-based detector was then applied on the silence region to classify into two groups: silence segments with low energy and silence segments with high energy and high zero-crossing rates. In total, 3 initial models are created in this approach, then iterative retrain and re-segment procedure is performed as described. It is also noted that the acoustic features for SAD module might not be the same with the ones for speaker clustering, e.g. Friedland et al. [11] used 13 MFCC with the first and second derivatives

Figure 2.2: *Hybrid energy-based and model-based speech activity detector*

appended as features for SAD while the same system employed 19 MFCC features for speaker segmentation and clustering. A probable explanation is that for speech modeling, higher cepstral coefficients should be removed since they contain more information about speaker characteristic rather than general speech properties, while for speaker modeling, higher cepstral coefficients should be retained.

2.3.4 Multi-channel speech detection

In recent years, with the increasing availability of multi-channel audio, there have been a number of related efforts toward *multi-speaker speech activity detection*. In Wrigley et al. [68, 69], the authors perform a systematic analysis of features for classifying multi-channel audio into four sub-classes: local channel speech, crosstalk speech, local channel and crosstalk speech, and non-speech. They look at the frame-level classification accuracy for each class with the various features selected for analysis. A key result from this work is that, from among the twenty features examined, the single best performing feature for each class is one derived from cross-channel correlation. This fact shows evidence of the

importance of cross-channel information for multi-channel detection task. Pfau et al. [70] propose an ergodic HMM (eHMM) speech activity detection and as a post-processing step, the authors identify and remove crosstalk speech segment by thresholding cross-channel correlations which yields 12% relative frame error rate (FER) reduction. In [71], Laskowski et al., propose a scheme using a cross-channel correlation speech/non-speech detection. This scheme is later used in a multi-channel speech activity detection system that models vocal interaction between meeting participants with joint multi-participant models [72–74].

2.4 Clustering architecture

Figure 2.3: *Hierarchical clustering architecture*

Figure 2.4: *Sequential clustering architecture*

Speaker clustering seeks to group all audio frames, segments from the same speakers together. Ideally, this process produces one cluster for each speaker with all segments of a given speaker assigned to a single cluster. Different diarization systems adopt different strategies for speaker clustering. However, in a broad sense, the clustering architectures fall into one of these categories: (1) *offline architecture*, or (2) *online architecture*. In offline architecture, the collection of all feature vectors are observable by the system at

all time and the algorithm could optimize through multiple iterations with no constraint on the execution time. While in online architecture, the features are presented to the system only when the data is available, the algorithm has no knowledge about the future and generally there is constraint on the latency, the time difference between when the result is obtained and when the data is available. Figure 2.3 and Figure 2.4 show the *hierarchical clustering* and *sequential clustering* approaches, which are the representatives for offline and online speaker diarization systems correspondingly. In our work, the focus of interest is on offline speaker diarization techniques. The following sections will discuss the components of various clustering architectures, with focus on well-established approaches.

2.4.1 Speaker modeling

At the heart of many speaker diarization systems is the choice of speaker modeling technique. As diarization is a part of speaker recognition, many modeling techniques in speaker verification and identification are also applicable. In this section, however, we only include those which have been adopted and shown to be effective for diarization tasks.

2.4.1.1 Gaussian Mixture Model

Since *Gaussian Mixture Model* (GMM) was initially introduced in the context of speaker modeling by Reynolds et al. [75] in 1995, it has become the standard reference method in speaker recognition. A GMM is a probability distribution that is a convex combination of several *Gaussian distributions*. The mixture density is:

$$f(x) = \sum_{k=1}^K a_k f_k(x) \quad (2.17)$$

where

- K is the number of mixtures.
- a_k is the prior probability of mixture k such that $\sum_{k=1}^K a_k = 1$

- $f_k(x)$ is the component density of Gaussian distribution parameterized by mean μ_k and covariance Σ_k :

$$f_k(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp \left(\frac{-(x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k)}{2} \right) \quad (2.18)$$

with d is the dimension of feature vector.

Given a sequence of observation vectors, the parameters of a GMM can be trained via the *Expectation Maximization* (EM) algorithm [76] to maximize the likelihood of the data. For speech, we assume that each observation in sequence $X = \{x_1, \dots, x_n\}$ is *independent and identically distributed* (i.i.d.). Accordingly, the likelihood of a GMM parameterized by θ given observations sequence X is computed as:

$$p(X|\theta) = \prod_{i=1}^n p(x_i|\theta) \quad (2.19)$$

In practice, many systems restrict the covariance matrices of the GMM to be diagonal, since it is computationally expensive and requires more training data to estimate the parameters of a full-covariance GMM.

The choice of K , the number of mixtures, is highly empirical as there is no consistency among different systems [10, 15, 23, 77]. In the ISCI agglomerative speaker diarization systems [38, 39, 78], the authors suggest the use of variable complexities. The systems are initialized with a fixed number of Gaussian mixtures for all cluster models. Upon merging any two clusters, the new cluster model is generated with the complexity as the sum of both parent's Gaussian mixtures. Later, in [79], Anguera et al. propose the concept of *cluster complexity ratio* (CCR), which assumes that the number of mixtures is linearly proportional to the number of features in the clusters, to initialize this parameter. Based on the similar assumption, in Leeuwen and Konecny [23], the *constant seconds per Gaussian* (CSPG) is used to determine the number of mixtures.

Mono-gaussian model uses a single Gaussian component with either full or diagonal covariance matrix as speaker model. Modeling with mono-gaussian is computationally efficient since only a small number of parameters need to be estimated. Although the accuracy is clearly behind GMM, it is sometimes the model of choice due to: lack of training data, or limitation on computational resource. In many speaker segmentation

systems [80, 81], since the speaker segments are relatively short, mono-gaussian models with full covariance matrices are employed to detect speaker change points. In others [40, 41], diagonal covariance matrices are used. Reynolds and Torres-Carrasquillo [52] perform bottom-up clustering with BIC metric and mono-gaussian models with full covariance.

2.4.1.2 Hidden Markov Model

The *Hidden Markov Model* (HMM) [82] is a generative probabilistic model comprising of a finite number internal hidden states and these states are not visible to observer. Each hidden state is associated with an emission probability distribution and an observation can be generated according to this distribution. In speech processing, Gaussian or mixtures of Gaussian are commonly used to model the emission probabilities. The transitions among hidden states are assumed to follow the first-order Markov process, they are specified by a transition probability matrix and an initial state distributions.

Formally, a HMM is completely specified by $\{\mathbf{\Pi}, \mathbf{A}, \mathbf{\Theta}\}$ with:

- A set of parameters of emission distribution conditioned on the hidden states: $\mathbf{\Theta} = \{\mathbf{\Theta}_1, \dots, \mathbf{\Theta}_N\}$ with N being the number of states.
- A matrix of transition probabilities

$$\mathbf{A} = \begin{pmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \ddots & a_{NN} \end{pmatrix}$$

where a_{ij} is the transition probability from state i to state j .

- The initial state distributions $\mathbf{\Pi} = \{\pi_1, \dots, \pi_N\}$

In this thesis, it is denoted $\mathbf{\Lambda} = \{\mathbf{\Pi}, \mathbf{A}, \mathbf{\Theta}\}$ as the parameters of the HMM. When used for speech, the HMM usually has a left-to-right topology. Given a sequence of observation vectors X , the parameters of the HMM are trained using EM algorithm to maximize the likelihood:

$$\mathbf{\Lambda}^* = \arg \max_{\mathbf{\Lambda}} p(X|\mathbf{\Lambda}) \quad (2.20)$$

The best hidden state sequence (q_{best}) is derived using the Viterbi algorithm [83], i.e.:

$$q_{best} = \arg \max_q p(X, q | \Lambda) = \arg \max_q p(X | q, \Lambda) \cdot p(q | \Lambda) \quad (2.21)$$

The likelihood of an observation vector x_n given state q_k , $p(x_n | q_k)$, are generally modeled by a GMM.

The HMM-based speaker clustering framework was first presented by Ajmera et al. in [84]. Since then, it has been widely adopted in most state-of-the-art speaker diarization systems [15, 23, 77]. The LIA speaker diarization system [40, 41] also use HMM with different topology for speaker modeling. In their system, there is no duration constraint, each state of the HMM characterizes a speaker and the transitions model the speaker turns. On the other hand, Kim et al. [85] perform re-segmentation by applying a HMM-based classifier on segments of 1.5 seconds each with the assumption that no speaker change within each segment.

2.4.1.3 Total factor vector

With the success of the *total variability* approach in the task of speaker verification [86], it has been recently adapted to the problem of speaker diarization [87]. In this modeling technique, a speaker utterance is represented by a *supervector* (M) that consists of component from the *total variability subspace*, contains the speaker and channel variabilities simultaneously.

$$M = m + Tw + \epsilon \quad (2.22)$$

where M is a speaker and session dependent supervector, a supervector in this context is a stacked mean vectors from a GMM [88], and m is the speaker and session independent supervector commonly adapted from the *Universal Background Model* (UBM) supervector. In the speaker recognition terminology, the UBM is a large GMM (in terms of 512 – 2048 mixtures), trained on speech from many speakers (several hundred to several thousand), to represent the speaker independent distribution of acoustic features [89]. T is the rectangular matrix of low rank which contains the eigenvectors with the largest eigenvalues of the total variability covariance matrix, w is a low-dimensional random vector having a standard normal distribution $\mathcal{N}(0, I)$, and the residual noise term ϵ covers the variabilities not captured by T [90]. The vector w , with dimension in order of hundreds

compared to the dimension in order of thousands of a supervector, is referred to as a *total factor vector*, an *identity vector* or an *i-vector*. In short, this modeling technique can be seen as a simple factor analysis for projecting a speech utterance from high-dimensional supervector space to the low-dimensional total variability space. Once projected to low-dimensional space, the applicability of many machine learning algorithms are then more straight forward.

Probably, the first attempt to make use of i-vector in the context of speaker diarization was presented by Shum et al. in [87]. In this paper, good diarization results on summed-channel telephone data with two speakers were reported with various dimensions of the i-vector from 40 to 600, in conjunction with *Principal Component Analysis* (PCA) for further dimension reduction and cosine distance metric for scoring. In the latter work also by Shum et al. [91], the authors applied i-vector in the framework of spectral clustering [92] to extend the solution to diarization of telephone data with unknown number of participating speakers. In [93], with the motivation that the estimation of i-vectors for short segments is not reliable and may harm the clustering process especially at early phases, Silovsky and Prazak employed the two-stage clustering approach, using i-vector in the second stage, while the first stage using GMM for speaker modeling. They reported performance improvement over the standalone i-vector system. In [94], Rouvier and Meignier re-defined the speaker clustering as a problem of *Integer Linear Programming* (ILP) based on i-vectors and conclude that i-vector models are more robust than GMMs.

2.4.1.4 Other modeling approaches

Supervector *Supervector* often refers to combining many low dimensional vectors into a higher dimensional vector. In speaker recognition terminology, supervector typically refers to Gaussian supervector [88], formed by stacking all the mean vectors of an adapted GMM. Supervector is widely used in many speaker verification systems together with *support vector machine* (SVM) classifier. These combinations have been shown to be effective and robust in many situations, probably due to the ability to capture the speech utterance statistics of supervector as well as the generalization capability of SVM in high dimensional space. In Tang et al. [95], supervector was used with either Euclidean or cosine distance metric to measure the distance among different speakers. These distances

are then used to learn the speaker-discriminative acoustic feature transformation and the discriminative speaker subspace. They reported that the speaker clustering methods based on the GMM mean supervector and vector-based distance metrics outperform traditional methods based on statistical model and statistical model-based distance metrics.

Eigen vector space model *Eigen vector space model* (EVSM) [96] is inspired from the eigenvoice approach [97]. Each cluster is first modeled by a supervector, then all the super vectors are projected to the lower subspace by applying *Principal Component Analysis* (PCA) to obtain new supervectors with lower dimension. These newly obtained vectors are termed eigen vector space models. It has been shown experimentally in [96] that clustering with EVSM and cosine distance metric consistently yielded higher cluster purity and lower Rand Index than the GLR-based method. In a later work, EVSM was also used in El-Khoury et. al. [43] for cluster modeling in their hierarchical bottom-up clustering system.

2.4.2 Distance measures

Many speaker diarization systems employ some kinds of distance metrics in one way or the other. In agglomerative systems, they are used to decide which clusters to merge and when to stop the clustering process. While in speaker segmentation, distance metrics are often used in conjunction with sliding windows to detect the speaker change points. On the other hand, these metrics also find some applications in cluster refinement and purification.

Many distance measures were proposed in the past and they can be broadly classified into two categories: (1) *template-based*, and (2) *likelihood-based*. The first kind compares the parameters of the models which are applied to the data. These are generally very fast to compute and often used as initial estimation or in real-time systems. The representatives of this kind are: *Symmetric Kullback-Leibler distance* (KL2) [98], *Divergence Shape Distance* (DSD) [99] and *Arithmetic Harmonic Sphericity* (AHS) [100]. The second group of distances require the evaluation of the fitness (likelihood) of the data given the representing models. These distances are slower to compute since the likelihood-score need to be evaluated for each data point, however their performance is better than those

in the first category. In this work, the metrics in the second group could be referred to as likelihood-based techniques, which are also the main focus of this chapter. Among them, the more popular distances are: the *Generalized Likelihood Ratio* (GLR) [101], the *Bayesian Information Criterion* (BIC) [102], the *Cross Likelihood Ratio* (CLR) [103] and the *Normalized Cross Likelihood Ratio* (NCLR) [104].

Let's consider two audio segments (i, j) with feature vectors $\mathbf{X}_i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_{N_i}^i\}$ and $\mathbf{X}_j = \{\mathbf{x}_1^j, \mathbf{x}_2^j, \dots, \mathbf{x}_{N_j}^j\}$ respectively. For brevity, from now on we will refer to the audio segments as $\mathbf{X}_i$ and $\mathbf{X}_j$. It is desirable that a proper distance metric would have a smaller value if these two segments belong to the same speaker and have a larger value if these two segments belong to different speakers.

2.4.2.1 Symmetric Kullback-Leibler distance

Kullback-Leibler (KL) divergence between two random variables A and B is an information theoretic approach to measure the expected number of extra bits required to encode random variable A with a code that was designed for optimal encoding of B [105].

$$KL(A||B) = \int_{-\infty}^{\infty} p_A(x) \log \frac{p_A(x)}{p_B(x)} dx \quad (2.23)$$

where p_A and p_B denote the pdf of A and B .

Symmetric Kullback-Leibler (KL2) is simply a symmetric version of KL and is defined as:

$$KL2(A, B) = KL(A||B) + KL(B||A) \quad (2.24)$$

When both A and B have Gaussian distributions, we can obtain the closed form solution [99]:

$$KL2(A, B) = \frac{1}{2} \text{tr}((\mathbf{C}_A - \mathbf{C}_B)(\mathbf{C}_B^{-1} - \mathbf{C}_A^{-1})) + \frac{1}{2} \text{tr}((\mathbf{C}_A^{-1} + \mathbf{C}_B^{-1})(\mu_A - \mu_B)(\mu_A - \mu_B)^T) \quad (2.25)$$

where $\mathbf{C}_A, \mathbf{C}_B, \mu_A, \mu_B$ are respectively the covariance matrices and means of p_A and p_B .

Given any two audio segments $\mathbf{X}_i$ and $\mathbf{X}_j$, they can be considered as random variables A and B and therefore the distance can be computed using the above formula.

$KL2$ is used in the work by Siegler et al. [98] to compute the distance between two sliding windows for speaker change point detection. It is also employed as cluster distance

metric in several agglomerative speaker clustering systems [98, 106]. For their system, the authors [98] show that $KL2$ distance works better than the Mahalanobis distance.

When using $KL2$ as a distance metric for speaker clustering, the speaker models are often assumed to be Gaussian distributed since there is closed-form expression to compute $KL2$. This assumption may make the speaker models too simple to be able to capture the characteristics of individual speakers. There are some works in this direction to adapt $KL2$ metric to more complex models. In Ben et al. [107], a novel distance between GMMs is derived from the $KL2$ distance for the particular case where all the GMMs are mean adapted from a common GMM based on the principles of Maximum A Posteriori (MAP) adaptation. The speaker diarization results using this metric are shown to be better than BIC when the segmentation is of high quality. However, due to the sensitivity to segmentation errors, this metric is not as robust as BIC in general. In Rogui et al. [108], the author propose a divergence measurement method between GMMs which is based on KL divergence, which allows fast computation of distance between GMMs.

2.4.2.2 Divergence Shape Distance

Observe that equation (2.25) comprises of two components in which one of them involves the mean vectors. As the mean vectors are easily biased by environment conditions, the *Divergence Shape Distance* (DSD) [109] is derived from the KL distance by eliminating this part. Therefore, the corresponding expression for DSD is:

$$DSD(A, B) = \frac{1}{2} \text{tr} ((\mathbf{C}_A - \mathbf{C}_B)(\mathbf{C}_B^{-1} - \mathbf{C}_A^{-1})) \quad (2.26)$$

Kim et al. [85] use DSD for speaker change detection; they show that the DSD method is more accurate than the BIC approach in presence of short segments, while both approaches are equivalent on long segments.

2.4.2.3 Arithmetic Harmonic Sphericity

The *Arithmetic Harmonic Sphericity* (AHS) [100] assumes the distributions of random variables are Gaussian and it can be viewed as an arithmetic harmonic sphericity test on covariance matrices of pdfs of two random variables. The AHS is defined as:

$$AHS(A, B) = \log \left(\text{tr}(\mathbf{C}_A \mathbf{C}_B^{-1}) \cdot \text{tr}(\mathbf{C}_B \mathbf{C}_A^{-1}) \right) - 2 \log(d) \quad (2.27)$$

where d is the dimension of feature vector.

2.4.2.4 Generalized Likelihood Ratio

Given two audio segments $\mathbf{X}_i$ and $\mathbf{X}_j$, we consider the following hypothesis test:

- H_0 : both segments are generated by the same speaker.
- H_1 : each segment is from a different speaker.

Assuming that the feature vectors of each speaker K are distributed according to the pdf $\mathcal{M}(\beta_K)$ with parameter set β_K .

- Under hypothesis H_0 : $\mathbf{X}_i \cup \mathbf{X}_j \sim \mathcal{M}(\beta_{ij})$
- Under hypothesis H_1 : $\mathbf{X}_i \sim \mathcal{M}(\beta_i)$ and $\mathbf{X}_j \sim \mathcal{M}(\beta_j)$

The *Generalized Likelihood Ratio* (GLR) between two hypotheses is defined by:

$$R = \frac{\mathcal{L}(\mathbf{X}_i \cup \mathbf{X}_j | \mathcal{M}(\beta_{ij}))}{\mathcal{L}(\mathbf{X}_i | \mathcal{M}(\beta_i)) \cdot \mathcal{L}(\mathbf{X}_j | \mathcal{M}(\beta_j))} \quad (2.28)$$

being $\mathcal{L}(\mathbf{X} | \mathcal{M})$ the likelihood of the data $\mathbf{X}$ given the model $\mathcal{M}$. The GLR differs from the standard *likelihood ratio* test (LLR) in that the pdf for the GLR are unknown and must be estimated from the data, whereas in the LLR the models are supposed to be known a priori.

$$R = \frac{\max_{H_0} p_{\mathcal{M}}(\mathbf{x}_1^i, \dots, \mathbf{x}_{N_i}^i, \mathbf{x}_1^j, \dots, \mathbf{x}_{N_j}^j | \beta_{ij})}{\max_{H_1} p_{\mathcal{M}}(\mathbf{x}_1^i, \dots, \mathbf{x}_{N_i}^i | \beta_i) p_{\mathcal{M}}(\mathbf{x}_1^j, \dots, \mathbf{x}_{N_j}^j | \beta_j)} \quad (2.29)$$

The feature vectors are assumed to be independent with each other, thus:

$$R = \frac{\max_{H_0} \prod_{k=1}^{N_i} p_{\mathcal{M}}(\mathbf{x}_k^i | \beta_{ij}) \prod_{k=1}^{N_j} p_{\mathcal{M}}(\mathbf{x}_k^j | \beta_{ij})}{\max_{H_1} \prod_{k=1}^{N_i} p_{\mathcal{M}}(\mathbf{x}_k^i | \beta_i) \prod_{k=1}^{N_j} p_{\mathcal{M}}(\mathbf{x}_k^j | \beta_j)} \quad (2.30)$$

The maximum values in both the numerator and denominator are generally found by applying Expectation-Maximization (EM) algorithm to estimate the parameter sets of

the models. The distance d_{GLR} is then computed from the logarithm of the previous expression:

$$d_{GLR} = -\log R \quad (2.31)$$

In Bonastre et al. [110], the GLR is used to segment the signal into speaker turns. In Adami et al. [111], an algorithm is specifically designed for two-speaker segmentation with GLR as distance metric. Another system for two-speaker segmentation is proposed by Gangadharaiah et al. [112], with GLR metric in the first segmentation step.

2.4.2.5 Bayesian Information Criterion

Bayesian Information Criterion (BIC) is a Bayesian approach to the model selection problem which is proposed by Schwarz [113]. The BIC value for a model M is defined as:

$$BIC_M = \log p(X|M) - \lambda \frac{\#(M)}{2} \log N \quad (2.32)$$

where $p(X|M)$ denotes the likelihood of data X given model M , $\#(M)$ denotes the number of free parameters in M and N denotes the number of observations in X , λ is a tunable parameter dependent on the data. BIC is an approximation to the posterior distribution on model classes. It is shown in [113] that maximizing BIC value also results in maximizing the expected value of the likelihood over the set of parameters of M . Thus, BIC is commonly used to choose the best parametric model among the set of models with different number of parameters.

Considering the same hypotheses as in GLR, under hypothesis H_0 we have:

$$BIC_{H_0} = BIC_{\mathcal{M}(\beta_{ij})} = \log \mathcal{L}(\mathbf{X}_i \cup \mathbf{X}_j | \mathcal{M}(\beta_{ij})) - \lambda \frac{1}{2} \#(\mathcal{M}(\beta_{ij})) \log(N) \quad (2.33)$$

Likewise, under hypothesis H_1 :

$$BIC_{\mathcal{M}(\beta_i)} = \log \mathcal{L}(\mathbf{X}_i | \mathcal{M}(\beta_i)) - \lambda \frac{1}{2} \#(\mathcal{M}(\beta_i)) \log(N) \quad (2.34)$$

$$BIC_{\mathcal{M}(\beta_j)} = \log \mathcal{L}(\mathbf{X}_j | \mathcal{M}(\beta_j)) - \lambda \frac{1}{2} \#(\mathcal{M}(\beta_j)) \log(N) \quad (2.35)$$

$$BIC_{H_1} = BIC_{\mathcal{M}(\beta_i)} + BIC_{\mathcal{M}(\beta_j)} \quad (2.36)$$

The BIC distance metric is then defined as:

$$d_{BIC} = BIC_{H_1} - BIC_{H_0} \quad (2.37)$$

The above expression can be re-written in terms of d_{GLR} as:

$$d_{BIC} = d_{GLR} - \lambda \frac{1}{2} (\#(\mathcal{M}(\beta_i)) + \#(\mathcal{M}(\beta_j)) - \#(\mathcal{M}(\beta_{ij})) \log(N) \quad (2.38)$$

$$= d_{GLR} - \lambda \frac{1}{2} (\Delta \mathcal{M}_{ij}) \log(N) \quad (2.39)$$

where $\Delta \mathcal{M}_{ij}$ is the difference between the number of free parameters of models in hypothesis H_1 and hypothesis H_0 . From this expression, we can view a BIC distance as a penalized GLR distance, with the penalty depending on the free parameter λ , number of parameters as well as number of observations. The selection of free parameter λ has been subject of constant study.

BIC is introduced for the case of speech and specifically for acoustic change detection and clustering by Chen and Gopalakrishnan [102], where the problem is formulated as that of model selection. In this paper, the authors introduced a tunable parameter λ in the penalty term which is used to improve the performance of the system for a particular condition in practice. This parameter therefore implicitly defines a threshold which needs to be tuned to the data and its correct setting has been subject of constant study [114–118]. Ajmera [38, 119] proposes a method to cancel the penalty term by adjusting the number of free parameters in the models accordingly. The authors use a GMM with diagonal covariance matrices for each of the models $\mathcal{M}(\beta_i)$, $\mathcal{M}(\beta_j)$ and $\mathcal{M}(\beta_{ij})$. By making the number of mixtures in $\mathcal{M}(\beta_{ij})$ equals to the number of mixtures in $\mathcal{M}(\beta_i)$ plus the number of mixtures in $\mathcal{M}(\beta_j)$, they enforce $(\Delta \mathcal{M}_{ij}) = 0$, and the penalty term is eliminated. In this case:

$$d_{BIC} = d_{GLR} \quad (2.40)$$

In Chen and Gopalakrishnan [102], it is shown that BIC value increases according to data size. This presents in general a problem when there is a big mismatch between clusters or windows with different data sizes. Thus, Perez-Freire [120] introduces the penalty weight which depends on the data size in order to achieve better robustness. Vandecatseyes [118] normalizes the BIC score by the total number of frames and shows that it consistently outperforms non-normalized BIC.

2.4.2.6 Cross Likelihood Ratio

The *Cross Likelihood Ratio* (CLR) measure was first used in Reynolds et al. [103] to compute the distance between two adapted speaker models and it was defined as:

$$d_{CLR} = \log \left(\frac{\mathcal{L}(\mathbf{X}_i | \mathcal{M}(\beta_j))}{\mathcal{L}(\mathbf{X}_i | \mathcal{M}(\beta_U))} \right) + \log \left(\frac{\mathcal{L}(\mathbf{X}_j | \mathcal{M}(\beta_i))}{\mathcal{L}(\mathbf{X}_j | \mathcal{M}(\beta_U))} \right) \quad (2.41)$$

where $\mathcal{M}(\beta_U)$ is the *Universal Background Model* (UBM); $\mathcal{M}(\beta_i)$, $\mathcal{M}(\beta_j)$ are adapted speaker models for speaker i and speaker j , respectively. The UBM is trained with a huge amount of audio data according to the gender (male, female) and the channel conditions. The speaker models are derived by adapting the UBM parameters with speaker speech data. The adaptation method often used is the Maximum A Posteriori (MAP) [121] adaptation. The CLR is commonly employed as distance metric in agglomerative speaker clustering systems including Barras et al. [122], Reynolds and Torres-Carrasquillo [52], and Sinha et al. [55].

2.4.2.7 Normalized Cross Likelihood Ratio

Normalized Cross Likelihood Ratio (NCLR) was presented as a distance measure between two speaker models [104]. Given two speaker models $\mathcal{M}(\beta_i)$ and $\mathcal{M}(\beta_j)$, the NCLR distance is defined as:

$$d_{NCLR} = \frac{1}{N_i} \log \left(\frac{\mathcal{L}(\mathbf{X}_i | \mathcal{M}(\beta_i))}{\mathcal{L}(\mathbf{X}_i | \mathcal{M}(\beta_j))} \right) + \frac{1}{N_j} \log \left(\frac{\mathcal{L}(\mathbf{X}_j | \mathcal{M}(\beta_j))}{\mathcal{L}(\mathbf{X}_j | \mathcal{M}(\beta_i))} \right) \quad (2.42)$$

2.4.2.8 Other distance measures

Gish-distance Gish et al. [123] propose a distance measure, which is referred to as Gish-distance in the literature. This distance is based on *likelihood ratio* with the assumption of multivariate Gaussian models and is used as clustering metric in [123] and [124]. Van Leeuwen [58] uses Gish distance for agglomerative clustering in the TNO speaker diarization system. Jin et al. [125] perform agglomerative clustering using Gish-distance with scaling heuristic to favour merging of consecutive segments.

Vector quantization distortion Mori and Nakagawa [116] use a *vector quantization* (VQ) distortion criterion. The experimental results in their paper demonstrate superior performance of VQ metric in both speaker segmentation and speaker clustering comparing to GLR and BIC . However, the database is too small (175 utterances) and restrictive (only clean speech) to deduce any conclusions.

XBIC In Anguera [80], a $XBIC$ metric, which is based on *cross-likelihood* between each data segment and the model trained on data from the other segment, is introduced for speaker segmentation and is shown to behave similar or better to BIC with reduction in computation.

Probabilistic pattern matching In Malegaonkar et al. [81], they employ a probabilistic pattern matching approach for detecting speaker changes and study different likelihood normalization techniques to make the metric more robust, achieving better results than BIC for speaker segmentation.

2.4.3 Speaker segmentation

Speaker segmentation, with the aim to split the audio stream into speaker homogenous segments, is a fundamental process to any speaker diarization systems. While many state-of-the-art systems tackle the problem of segmentation and clustering iteratively, traditional systems usually perform speaker segmentation or *acoustic change point detection* independently and prior to the clustering stage. Various segmentation algorithms have been investigated in previous works, which can be categorized in one of the following groups: (1) *silence detection based methods*, (2) *metric-based segmentation*, and (3) *hybrid segmentation*.

2.4.3.1 Silence detection based methods

Some of the speaker segmentation techniques are based on silence detection in speech signal. In these methods, it is supposed that there exists a silence region at change points between speaker turns. The silence is detected either by a decoder [64] or directly by measuring and thresholding the audio energy [124, 126]. The segments are then generated

by cutting the input at silence locations. However, the accuracy of such naive techniques is poor [124]. Moreover, the correlation between the existence of a silence in a recording and a change of speaker is arbitrary at most. Therefore, such techniques are usually used to detect hypothetical change points, which will then be confirmed by more advanced techniques in the latter stage.

2.4.3.2 Metric-based segmentation

The favoured approach to speaker segmentation is to observe adjacent windows of data and calculating a distance metric between the two, then deciding whether the windows originated from the same or different speakers. The decisions generally base on a threshold/penalty term and this threshold is set empirically by using an additional development data. Various *metric-based segmentation* algorithms have been proposed in the literature, the differences among them lie mainly in the choice of distance metrics, the size of two windows, the time increment of the shifting of the two windows, and the threshold decisions.

Fixed-size sliding window In the pioneered work by Siegler et al. [98], the authors represent each window as a Gaussian and computing the distance between the two distributions using the symmetric $KL2$ distance. To accomplish this, means and variances are estimated for a 2 second window placed at every point in the audio stream. When the $KL2$ distance between bordering windows reaches a local maximum, a new segment boundary is generated. The same framework is applied in Bonastre et al. [110] with the GLR as the distance metric and a tuned threshold to avoid missed detection to the detriment of false alarms. In Adami et al. [111], one speaker model is estimated from the small segment at the beginning of the conversation and the segment that has the largest GLR distance from the initial segment is used to train second speaker model. The segment boundaries are defined at the points where the GLR distances with respect to both speakers are equal; each segment in the conversation is assigned to the speaker with the smallest distance. Kim et al. [85] use DSD for speaker change detection with the covariances estimated for two sliding windows of 3 second and 2.5 second overlapping. They show that the DSD metric is more accurate than the BIC approach in presence of short

segments, while both approaches are equivalent on long segments. In a more recent work, inspired by speaker verification techniques, a probabilistic pattern matching method with several likelihood normalization methods are investigated for speaker segmentation task in [81]. The proposed bi-lateral scoring scheme is shown to be more effective than BIC and XBIC, mainly due to the inclusion of score normalization techniques.

Variable-size sliding window Later, Chen and Gopalakrishnan [102] formulate the problem of speaker change detection as a model selection problem and use BIC for this purpose. This technique looks for potential change points in a window of frames by testing two hypotheses: the first hypothesis assumes the data in the window belong to one speaker and therefore is better represented by that speaker distribution, on the other hand the second hypothesis assumes that there are two different speakers, hence the data are better modelled by two different distributions. In case there is no change point detected within the window, its size is increased by a certain number of frames depending on the algorithm and the process is repeated. Tritschler and Gopinath [114] propose another variable window scheme in which the size of the window is increased adaptively in contrast to a fixed amount as in Chen and Gopalakrishnan [102]. They also devise some rules to eliminate some of the BIC tests in the window, when they correspond to locations where the detection of a boundary is very unlikely. These heuristics make the algorithm faster and give importance to detecting short changes. In Sivakumaran et al. [127] and Cettolo and Vescovi [128], by significantly reducing the number of operations involved in the estimation of the means and covariance matrices, the segmentation process are sped up. In Roch and Cheng [129], a MAP-adapted version of the models is presented, which allows for shorter change points to be found at the cost of being slightly worse than EM-trained models when longer hypothesis windows are used. A notable variation to BIC has been proposed by Ajmera and Wooters in [119] and consists of fixing the number of parameters between the two BIC hypotheses so as to eliminate the need for tuning the BIC penalty term.

Multi-step segmentation There is also works which attempt to make the detection procedure faster by applying a distance measure prior to BIC. DIST-BIC [115, 130] is a

work in this direction. A log-likelihood ratio (LLR) based distance computation prior to BIC is proposed in this work, which is faster than BIC. Then, only selected change points are passed through the BIC test. In Zochova et al. [131], the same framework is used with some modifications in speaker change candidate detection and speaker change position location. They report better results in a majority of tests. Also in this direction, Zhou and Hansen [106] propose applying T^2 -statistics prior to BIC. The authors claim to improve the algorithm speed by an order of 100 compared to Chen and Gopalakrishnan [102] without sacrifice the overall performance. Lu and Zhang [132] apply $KL2$ distance on *line spectrum pair* (LSP) frequency features; the speaker change detection scheme is able to meet the requirement of real-time processing in multimedia applications. Vandecatseye [118] use a measure called *normalized log likelihood ratio* (NLLR) to generate potential change points in the first stage and then use normalized BIC in second stage to eliminate false alarm turns. All of these algorithms perform in a bottom-up manner where there are many short speaker turns in the first step which will be eliminated subsequently in the second step. However, Wang [133] proposed a method which perform in top-down manner. A long sliding windows is first used to segment a long audio stream into shorter sub-segments, and then applies the sequential divide and conquer segmentation to each sub-segment with shorter windows to detect the remaining change points. Both stages use BIC as the distance metric. In Gangadharaiah et al. [112], a two-speaker segmentation is performed in two steps. In the first step, a standard approach for segmentation is applied using GLR with same size adjacent windows, fixed step shifting. In the second step, several segments are selected to train a GMM for each speaker and the rest are assigned to either speaker with a *maximum likelihood* (ML) approach.

2.4.3.3 Hybrid segmentation

Liu and Kubala [64] introduce a two-stage hybrid segmentation system combining model-based and metric-based approach. The output of a phone-based decoder is used as the initial segments and a new penalized GLR criterion is employed to accept/reject change-points previously found. Kemp et al. [124] chopped the input signal into short segments of 1 second, then performed bottom-up clustering using Gish distance until a predetermined number of clusters remains. GMMs are trained for each cluster and a model-based segmenter is then applied.

2.4.3.4 Segmentation evaluation

In evaluating segmentation performance, two kinds of error measures are commonly computed, *false alarm rate* (FAR) and *miss detection rate* (MDR):

$$FAR = \frac{\text{Number of false alarms}}{\text{Number of detected change points}} \quad (2.43)$$

$$MDR = \frac{\text{Number of miss detections}}{\text{Number of actual change points}} \quad (2.44)$$

where a *false alarm* refers to a change point is detected but it does not exist, a *miss detection* refers to an existing change point but is not detected by the algorithm. On the other hand, one may use the *recall* (RCL) and *precision* (PRC) defined as:

$$RCL = 1 - FAR \quad (2.45)$$

$$PRC = 1 - MDR \quad (2.46)$$

In order to consider the trade-off between these two metrics, the F measure can be used:

$$F = \frac{2 \times PRC \times RCL}{PRC + RCL} \quad (2.47)$$

2.4.4 Speaker clustering

2.4.4.1 Agglomerative hierarchical clustering

Most state-of-the-art systems use an *agglomerative hierarchical clustering* (AHC) approach, also known as *bottom-up clustering*, where the systems start with an over-determined number of segments/clusters and via merging procedures to converge to the optimum number of clusters determined by some stopping criteria.

Step-by-step speaker segmentation and clustering This is the classical approach where change point detection is typically used to segment the recording into speaker homogeneous segments and then these segments are grouped together according to a distance measure until a stopping criterion is satisfied.

Jin et al. [125] is probably one of the earliest research done in speaker clustering with intention for speaker adaptation in *automatic speech recognition* (ASR) systems. After segmentation, the system built a distance matrix using *Gish-distance* based on the

Gaussian models of the acoustic segments and hierarchical clustering was performed on this distance matrix, in which the distance between consecutive segments was scaled by a factor to increase the probability of merging these segments. As stopping criterion, the optimal clustering was selected by minimizing the within-cluster dispersion with some penalty against too many clusters. However, no systematic way to deduce the optimal value of the penalty term was proposed in this work. With the same purpose of speaker adaptation, in Siegler et al. [98], the KL2 distance was used as a distance metric. The authors showed that the KL2 distance works better than the Mahalanobis distance for speaker clustering. In this work, the stopping criterion was determined with a merging threshold, which was presumably tuned on development data set however no training procedure was specified. Solomonoff et al. [134] used the GLR distance matrix for speaker clustering. The authors proposed a cluster purity metric to evaluate the quality of a partition, which can also be used to determine the appropriate number of clusters. However, this metric requires knowing the true speaker of each segments, thus a method to estimate cluster purity without true labels was also presented in the paper. The estimation method involved a tunable parameter, but the authors did not clearly indicate how to obtain this parameter value. In Tsai et al. [135], they also used the same metric to determine number of speakers, however with an entirely different inter-cluster distance measurement. Instead of training models for each speaker clusters as classical measurement, they projected the segments/clusters into a speaker reference space, in which the distances between segments/clusters are claimed to be more effective and reliable. Although fairly good performance has been obtained, they also raised concern about the correlation between speaker reference bases, which ideally should be statistically independent with each other. Jin et al. [136] also used GLR metric in their work with the modification in speaker models. Instead of training each speaker model independently, they constructed a UBM from all the speech segments of that recording then used MAP adaptation for each individual speaker.

Chen and Gopalakrishnan [102] introduced BIC metric for speaker segmentation and clustering. In this work, starting from each individual segment as a cluster, hierarchical clustering was performed by calculating the BIC measure for every pair of clusters and merging the two clusters with the highest BIC measure. The clustering is stopped when

no two clusters resulted into an increase in BIC measure, when merged. Zhou and Hansen [106] proposed a multi-step segmentation approach using T^2 -statistic to select potential change points and validate with BIC to speed up the segmentation procedure. Then a GMM classifier was employed to automatically label the segments as male or female speech, and bottom-up clustering was performed on the segments of each gender independently, with BIC as distance metric and stopping criterion. It was shown that with gender labeling, the resulting clusters have higher purity in comparison with no gender labeling. This algorithm has been applied for audio indexing tasks in [137]. In Cassidy [59], Mahalanobis was used as cluster distance metric to merge all segments longer than 1.5 second and BIC was used as stopping criterion. Once speaker clustering has been performed, a Gaussian model was then trained on each cluster and these models were used to classify all speech segments. The similar framework was employed in Van Leeuwen [58] with BIC for speaker segmentation, stopping criterion and Gish distance for agglomerative clustering. Zhu et al. [56] applied Viterbi re-segmentation after the initial segmentation with *Gaussian divergence measure*. Then, a two-stage clustering method was performed, with BIC agglomerative clustering preceding a speaker identification module. The boundaries of speech segments are kept unaltered during the clustering process.

In some later works, the additional re-segmentation step was implemented after the clustering procedure, to refine the segment boundaries. In [122], the authors used BIC as both the distance metric and stopping criterion with the addition module of speaker identification (SID) clustering. The system first classified each cluster for gender and bandwidth and used MAP adaptation to derive speaker models from each cluster. In SID clustering process, agglomerative clustering was performed separately for each gender and band condition, the speaker models were compared using a metric between clusters named cross likelihood distance [103]. It was shown that with the addition of SID clustering, the diarization error rate could be reduced nearly 50% relatively. In Kim et al. [85], a multi-stage approach was proposed which includes: speaker segmentation using DSD metric, initial clustering with BIC as distance metric, clustering by HMM models likelihood scores, and finally HMM-based re-segmentation. The combined method was shown to outperform any individual approach. The LIUM speaker diarization system [57] was

also based upon a standard framework composed of three modules: signal split into small homogeneous segments with GLR metric, speaker clustering without changing the boundaries with BIC metric, and boundaries adjustment with Viterbi decoding. Other than HMM and Viterbi, Reynolds and Torres-Carrasquillo [52] employed the iterative re-segmentation using GMM for frame-based classification with a smoothed window of 100 frames. Moreover, they also included non-speech models in the re-segmentation step.

Iterative speaker segmentation and clustering Typically, in this framework, initial clusters are obtained by some initialization procedures. Models are then trained on these clusters and Viterbi decoding is performed to identify when the speaker changes occur as well as the hypothesized speaker identity of each segment. The procedure is repeated: the latest segmentation is used to train the hypothesized speaker models, then Viterbi decoding is run to perform speaker segmentation and speaker clustering.

In Gauvain et al. [61], the authors proposed an iterative GMM clustering method which uses an objective function based on penalized log-likelihood. For each iteration, all possible pairs of segments are considered for merging and its corresponding likelihood loss is calculated. Eventually, the pair with the smallest loss is merged and the GMM statistics were reevaluated. The process is reiterated until the likelihood loss crosses a specified threshold, then the data is segmented using a Viterbi decoder with the newly estimated GMMs. The whole process is repeated until the segmentation converges or a maximum number of iterations are reached. Two parameters are introduced in this function to penalize number of segments and number of clusters. The function therefore could be used both to determine which clusters should be merged and to determine when to stop merging. However, the selection of parameters is too ad-hoc and there are two parameters to tune, this method is coupled with robustness issue. Sinha et al. [55] also used a similar *iterative agglomerative clustering* (IAC) scheme with the addition of *speaker identification* (SID) clustering after IAC. In the SID clustering phase, CLR metric is employed to select the closest pair of clusters to be merged and a threshold is defined to stop the merging process.

The paper by Ajmera and Wooters [38] was probably the first to suggest an iterative, agglomerative clustering technique based on a HMM framework. A uniformly initial

segmentation is used to train speaker models that iteratively decode and retrain on the acoustic data. Pairs of closest clusters are merged in successive iterations and merging stops automatically using a threshold-free BIC metric. Later, in the ICSI-SRI fall 2004 diarization system, Wooters et al. [39] used the same framework with the introduction of Viterbi segmentation likelihood scores as stopping criterion. The Viterbi stopping criterion was reported to be slightly better than BIC which is contributed by the reduction in speaker error rate. In Anguera et al. [78], they improved this framework with an addition of purification algorithm to split clusters that are not acoustically homogeneous.

Information bottleneck *Agglomerative Information Bottleneck* (aIB) is another bottom-up algorithm used to perform speaker diarization. The goal of the aIB system is to iteratively merge uniform short segments $S = s_1, s_2, \dots, s_N$ into clusters $C = C_1, C_2, \dots, C_K$ which simultaneously maximize the mutual information $I(Y, C)$ of a set of relevance variables Y and a set of clusters C , while minimizing the mutual information $I(C, X)$ of C and a set of segments S , as shown in Equation (2.48). The merging continues until the stopping criterion is met. After which, Viterbi decoding is performed in order to determine the segment boundaries.

$$\max \left[I(Y, C) - \frac{1}{\beta} I(C, X) \right] \quad (2.48)$$

where Y is a set of components of a background GMM trained on the entire audio recording, and β is a Lagrange multiplier. Thus, Equation (2.48) is used to determine a cluster representation C which is useful for describing the relevance variables Y (maximize $I(Y, C)$) and simple (minimize $I(C, X)$). The aIB is more computationally efficient than the HMM-GMM speaker diarization system since new models are not trained for each potential merging of two clusters. Instead, for the aIB framework subsequent statistics are taken to be averages of previously defined statistics. Speaker diarization systems which employed aIB are predominantly implemented by Vijayasenan et al. [138–143].

Multi-stream clustering With the available of multi-channel recordings, the time delays between microphone pairs can be computed. In Ellis and Liu [144], the delay feature vectors are then classified into individual speakers using spectral clustering. However,

the system misses many speaker turns which incurs miss speaker errors and leads to high overall error rate. In Pardo et al. [46], the authors use the same framework as in Ajmera and Wooters [38] with time delay feature in place of acoustic feature. The system is compared to that of [144] and significant improvement is reported. However if comparing these results with the results obtained from the same system using standard acoustic feature, there is still a big gap to cover. Luque et al. [145] analyze the TDOA distribution of a recording and exploit the most likely and stable pairs of TDOA to obtain an initial clustering of speakers. An iterative agglomerative clustering framework similar to [38] is then employed with MFCC as features. The authors report better performance comparing to uniform initialization of clusters. In Anguera [67], the inter-delay feature and acoustic-feature are cleverly combined in a multi-stream HMM framework and the performance is greatly improved. In this framework, each feature stream is assigned a weight which reflects the relative contribution of individual feature stream; the weights are learned from development data. Later in [146], the same authors propose an automatic weighting for the combination of these two feature streams. The scheme is later used in ICSI RT07s Speaker Diarization System [15] and this is the state-of-the-art system thus far. Following the same clustering framework, the AMIDA speaker diarization system [23], in contrast to uniform initialization, starts with typically 40 initial clusters by performing segmentation and clustering using BIC. The system employs CLR as clustering distance measure using both cepstral and delay features. The weight of each feature stream is fixed, with higher contribution given to acoustic features.

Apart from time delay features, other features may also combine with conventional spectral features in multi-stream speaker diarization system. Vinyals and Friedland [147] propose the use of *modulation spectrogram* [148] as an additional stream of features to the commonly used MFCCs. In this work, the clustering framework follows the ICSI agglomerative clustering approach [15, 38, 78] with fixed weighting for each feature stream. In [44, 149], the authors investigated a large set of 70 prosodic and long-term features and applied Fisher criterion to rank these features by their ability to discriminate speakers. It is shown in the paper that the combination of MFCC features with the additional top-ten ranked prosodic and long-term features leads to improvement in terms of *diarization error rate*.

Clusters initialization With respect to agglomerative clustering framework, initial clusters play an important role to the final performance of speaker diarization systems. As demonstrated in the work of Koh et al. [F], good initial clusters lead to good diarization results. In that paper, the authors proposed a bootstrap clustering procedure which use *direction of arrival* (DOA) information to obtain initial clusters with very high accuracy. Relying on DOA information, however, severely restricts the applicability of this technique, particularly when this information is not readily available in many situations, for e.g. single channel diarization system. Several studies have been taken in the direction of initialization methods with attempt to improve the quality of initial clusters. A brief overview of these initialization techniques will be presented in this section, with focus on single channel techniques.

1. **Uniform initialization** *Uniform initialization* is the simplest and surprisingly is quite an effective technique. The speech frames obtained from the output of the SAD module are uniformly grouped into K_0 clusters, where the value of K_0 is empirically determined from development data set, typically K_0 is from 15-20. Despite being such a naive strategy, many systems [12, 15, 22, 66, 78, 150] have adopted this method and reported competitive results. Another variation of uniform initialization was suggested in Sun et al. [10], where the segments from the SAD output were used. These segments are further segmented if they are longer than m seconds, the segmentation points are set to be the lowest energy frame in the range of $m - \Delta m$ seconds and $m + \Delta m$ seconds. These short segments of approximately m seconds are then linearly grouped into K_0 such that each cluster contains about the same number of segments.
2. **Speaker segmentation and initial clustering** A plausible approach, at least in theory, for clusters initialization is first to detect the speaker turns and then to perform clustering to group these short segments into K_0 clusters. This approach was predominantly used in many early systems, however it does not seem to be superior to uniform initialization when the final results are considered and it also takes more execution time as speaker segmentation is required.

2.4.4.2 Divisive hierarchical clustering

Divisive hierarchical clustering, also known as top-down clustering, starts with very few clusters and proceed to split the clusters iteratively until the desired number of clusters is reached. In the current literature there are few systems following this clustering framework.

A top-down split-and-merge speaker clustering frame work is proposed in Johnson and Woodland [151] to enhance the accuracy of ASR in broadcast news by improving the unsupervised speaker adaption. The clustering algorithm starts with one node consisting of the whole speech recording. At each stage, a node is considered to be split into four child nodes if some segments belong to that node might move to other nodes using the *maximum likelihood* (ML) criterion. The splitting is continued until the algorithm converges or the maximum number of iterations is reached. At each stage of splitting, clusters that are very similar to each other are allowed to merge. Two different implementations of the algorithm are proposed: one is based on direct maximization of MLLR and one is based on AHS metric. In Johnson [152], the same framework with AHS distance metric is applied for speaker diarization using the stopping criterion similar to the one introduced in Solomonoff et al. [134].

In Meigner et al. [153], an iterative approach combining both segmentation and clustering in a top-down manner named evolutive HMM (e-HMM) is proposed. Initially, the system starts with one HMM trained on all the acoustic data available. The best subset features of this model (in terms of maximum likelihood scores) are taken out to train a new model using MAP adaptation. According to the subset selected, a segmentation is performed using Viterbi decoding. This process is repeated until the gain in likelihood score is insignificant, which is controlled with a tunable parameter. This parameter significantly influences the segmentation error as reported in the paper. In Anguera and Hernando [154] a similar approach is followed and a repository model is further introduced, which shows an improvement of 20% relatively.

2.4.4.3 Other approaches

Self Organizing Map In [155], Lapidot presented an approach for speaker clustering based on *Self Organizing Map* (SOM) given a known number of speakers. In this ap-

proach, SOM is used as likelihood estimators for speaker model and BIC is applied for estimation of the number of clusters.

Genetic algorithm In Tsai and Wang [156], they formulated the problem of speaker clustering as that of maximizing the overall within-cluster homogeneity. The within-cluster homogeneity is defined as the likelihood probability that a cluster model, trained using all the utterances within a cluster, matches each of the within-cluster utterances. This probability is maximize using genetic algorithm with initial random cluster assignment and iterative evaluation of the likelihood and mutation. In order to select the optimum amount of clusters they used BIC computed on the resulting models.

Variational Bayesian In [157, 158], Valente and Wellekens explored the use of Variational Bayesian (VB) learning, which has the capacity of model parameter learning and model complexity selection at the same time, for speaker clustering. With the proposed VB approach, the initial speaker models could be modeled as GMMs with any number of Gaussians, VB automatically prunes together with the cluster number, the best Gaussian model at the same time, resulting in smaller models where few observations are available and in bigger models where more observations are available.

Dirichlet Process Mixture Model In the previous works, the model learning methods (EM, ML, MAP) require the model space (such as number of mixtures, components, states etc.) is known a priori. Recently, Valente [159] proposed the use of infinite models for speaker clustering, in which the segmentation is obtained through a Dirichlet Process Mixture Model (DPMM). DPMM is a flexible model of unknown complexity with a prior on the parameters follow Dirichlet Process [160] which avoids fixing the number of observation modes. The experiments on broadcast news data showed improvements over ML/BIC, MAP/BIC and VB. In [161], Fox et al. extended the original work on hierarchical Dirichlet process hidden Markov model (HDP-HMM) [162] and apply this framework for speaker diarization on the NIST meeting database. The reported result was comparable to that of the state-of-the-art system [15], which use agglomerative BIC clustering, in NIST Rich Transcription Evaluations 2007.

2.4.4.4 Multiple systems combination

The speaker diarization systems presented thus far use either top-down or bottom-up technique for clustering. There are some works on algorithms to combine multiple systems and obtain an improved speaker diarization. In Tranter [163], the author presents a cluster-voting scheme designed to reduce the diarization error rate by combining information from two different diarization systems. Improvements are shown on broadcast news database when combining two bottom-up systems and two top-down systems. In Moraru et al. [40, 41], two strategies for systems combination are presented, those are: hybridization strategy, and merging strategy. The hybridization strategy consists in using segmentation results of the bottom-up system to initialize the top-down system. This solution associates the advantages of longer and quite pure segments of the agglomerative hierarchical approach with the HMM modeling and decoding power of the integrated approach. The merging strategy proposes a matching of common resulting segments followed by a re-segmentation of the data to assign the non-common segments.

2.4.4.5 Speaker clustering evaluation

Consider N_s speakers that are clustered into N_c groups, where n_{ij} is the number of frames in cluster i spoken by speaker j , n_{c_i} is the number of frames in cluster i , n_{s_j} is the number of frames spoken by speaker j , and N is the total number of frames.

Average cluster purity The *average cluster purity* (*acp*) [84] gives a measure of how well a cluster is limited to only one speaker; it reduces when a cluster includes segments from two or more speakers. The *acp* is based on cluster purity which is defined as:

$$p_{c_i} = \sum_{j=1}^{N_s} \frac{n_{ij}^2}{n_{c_i}^2} \quad (2.49)$$

where p_{c_i} is the purity of cluster i . Then the *acp* is computed as:

$$acp = \frac{1}{N} \sum_{i=1}^{N_c} p_{c_i} n_{c_i} \quad (2.50)$$

Average speaker purity On the other hand, the *average speaker purity* (*asp*) [84] gives a measure of how well a speaker is limited to only one cluster; it reduces when speech of a single speaker is split to more than one cluster. The *asp* is based on the speaker purity:

$$p_{s_j} = \sum_{i=1}^{N_c} \frac{n_{ij}^2}{n_{s_j}^2} \quad (2.51)$$

where p_{s_j} is the purity of speaker j . The *asp* is computed as:

$$asp = \frac{1}{N} \sum_{j=1}^{N_s} p_{s_j} n_{s_j} \quad (2.52)$$

K measure To balance the trade off between *acp* and *asp*, as well as to facilitate comparison between systems, Ajmera [84] propose the K measure, which is a geometrical mean of *acp* and *asp* :

$$K = \sqrt{acp * asp} \quad (2.53)$$

Rand index The *Rand index* [164] is a widely used measure for comparing partitions. It gives the probability that two randomly selected frames are from the same speaker but grouped in different clusters, or the two frames are in the same cluster but from different speakers. Rand index is defined as:

$$R = \frac{1}{\binom{N}{2}} \left[\frac{1}{2} \left(\sum_{i=1}^{N_c} n_{c_i}^2 + \sum_{j=1}^{N_s} n_{s_j}^2 \right) - \sum_{i=1}^{N_c} \sum_{j=1}^{N_s} n_{ij}^2 \right] \quad (2.54)$$

Rand index value changes from 0 to 1. The lower the index, the higher the agreement is between two partitions. However, it does not provide any information on how the partitions are distributed and how the two partitions are related.

2.5 Summary

The chapter has given an introduction to speaker diarization system in general and diarization in meetings in particular. The presentation focuses on off-line speaker diarization systems with hierarchical clustering framework as these approaches are thus far

the most popular in the literature. The system structures and their components were examined thoroughly and relevant works related to each component were reviewed. This chapter would provide the readers with the necessary backgrounds and understanding on the topic of speaker diarization, which would then be explored further in subsequent chapters.

Chapter 3

Speaker diarization systems in meetings

Over the last decade, NIST has held eight Rich Transcription (RT) evaluations [8] to investigate the speaker diarization problem. The ultimate purpose of the RT evaluations is to automatically produce transcriptions which are more informative to humans and more useful for machines [8]. Initiated originally within the telephony domain, and subsequently in broadcast news, in recent years, speaker diarization receives the most attention in the domain of conference meetings. These NIST evaluations, which specify standard experimental protocols and databases, give speaker diarization researchers the possibility to meaningfully compare different approaches on unseen datasets. In this thesis, special emphasis will be placed on established technologies within the context of the NIST RT benchmark evaluations, which has become a reliable indicator for the current state-of-the-art in speaker diarization.

3.1 NIST Rich Transcription Evaluation in Meetings

In a typical meeting room setup [165–168], multiple microphones are available for recording, from headwear microphones or lavalier microphones equipped for each meeting participant to microphones placed around the room and/or on the table surface. The former ones are referred to as close-talking microphones while the later ones are distant microphones. Speaker diarization systems are often reported to use distant microphones, either single channel or multiple channels, since these microphones are considered to be

less intrusive and cumbersome to users. The convenience, however, is not without issues. The speech signals picked up by distant microphones are generally at much lower *signal to noise ratio* (SNR) than the signals recorded by close-talk microphones due to the distance to the speech source. Moreover, the present of interference noises such as shuffling, knocking, and coughing etc. is more noticeable in the recordings. As a result, speaker diarization using distant microphones is considered to be more challenging and is one of the primary benchmark conditions in NIST Rich Transcription evaluation [17–19].

3.1.1 Meeting room setup

Within the series of RT evaluations, the meetings are captured in different meeting rooms with different microphone configurations. Typically, the most common sensors located within the meeting room are :

- **Individual Headset Microphone:** The individual headset microphone (IHM) is a head-mounted microphone positioned very close to the participant’s mouth. The microphone is usually a cardioid or super-cardioid microphone and has the best quality signal for each speaker.
- **Lapel Microphone:** The lapel microphone (LM) is another type of individual microphone, but is placed on the participant’s clothing. The microphone is generally omnidirectional or cardioid and is more susceptible to interfering speech from other participants.
- **Tabletop Microphone:** The tabletop microphone is typically an omni-directional microphone and is placed between participants on a table or other at surface. The number and placement of such microphones varies based on table geometry and the location and number of participants.
- **Microphone Array:** The microphone array is a collection of omni-directional microphones with a fixed topology (linear or circular). Depending on the sophistication of the setup, the array composition can range from four to sixty-four microphones.

The first two types comprise the sensors for the near-field or close-talking microphone condition and the last two sensors for the far-field or distant microphone condition. In theory, the speech from a near-field microphone should be that of the wearer of the microphone, making diarization unnecessary. Hence, speaker diarization is generally limited to the distant microphones.

3.1.2 Databases for speaker diarization in meetings

3.1.2.1 Official evaluation datasets

NIST realeased several standard sets of conference speech databases, which were the official evaluation datasets for NIST RT evaluations [17–19]. These datasets were also used to evaluate the performance of different speaker diarization systems in this thesis. Details of these data sets are summarized in Table 3.1, 3.2, and 3.3. In the scope of this thesis, the data from RT 2005 and RT 2006 were used as development data for many experiments, unless stated otherwise.

Table 3.1: *Conference speech databases. N: number of meeting excerpts. T: average length of each meeting excerpt in seconds*

	RT 2005	RT 2006	RT 2007	RT 2009
N	10	8	8	7
T	722	1080	1352	1543

Table 3.2: *Summary of data sources in terms of number of speakers. S: number of speakers. N: number of meeting excerpts*

S	4	5	6	7	8	9	11
N	20	5	3	1	1	2	1

Table 3.3: *Summary of data sources in terms of recording sites. C: recording sites. N: number of meeting excerpts*

C	AMI	CMU	EDI	ICSI	NIST	VT
N	4	6	6	2	9	6

3.1.2.2 Development datasets

Besides the official evaluation datasets, there are several available databases for meeting recordings in which the speaker segments are accurately transcribed to serve the need for speaker diarization task:

- **The ISL Meeting Corpus** [169]: 104 meetings with a total of 103 hours. Each meeting lasts an average of 60 minutes, with an average of 6.4 participants.
- **The ICSI Meeting Corpus** [170]: 75 meetings collected during the years 2000-2002. The recordings range in length from 17 to 103 minutes, but generally about 1 hour each. There are a total of 53 unique speakers in the corpus. Meetings involve from 3 to 10 participants, averaging 6.
- **NIST Meeting Pilot Corpus** [171]: 19 meetings collected between 2001 and 2003. Approximately 15 hours of data are recorded simultaneously from multiple microphones and video cameras.
- **The AMI Meeting Corpus** [172]: 100 hours of meeting recordings. The meetings are recorded in English using three different rooms with different acoustic properties, and include mostly non-native speakers.

3.1.3 Speaker diarization system evaluation

Typically, two types of errors are reported for each diarization system, those are: (1) *speech activity detection error rate* (SER), and (2) *diarization error rate* (DER). The SER measures the performance of *speech activity detection* (SAD) module while the DER measures the overall performance of the speaker diarization system.

3.1.3.1 Speech activity detection

The *speech activity detection* (SAD) error rate (SER) is a performance indicator of SAD module, it measures the ratio of misclassified speech and non-speech to the total amount of speech under evaluation. The lower this error rate, the better the SAD module is. Formally, SER [8] is defined as:

$$SER = \frac{T_{FA} + T_{miss}}{T_{speech}} \quad (3.1)$$

where T_{FA} , T_{miss} , and T_{speech} are the duration of silence misclassified as speech, the duration of speech wrongly classified as silence, and the total duration of speech under evaluation correspondingly.

3.1.3.2 Diarization error rate

The standard performance metric of the speaker indexing and diarization systems is the *diarization error rate* (DER) [8]. To evaluate the performance, an optimum mapping from the reference speakers in the conversation to the system speakers of the system should be found. The criterion for this mapping optimality is the percentage of the speech parts which are common to both the reference speaker and the system speaker. This optimality metric is calculated for all segments and all speakers. The mapping should map each reference speaker to at most one system speaker and vice versa. Once the optimal mapping is found, the DER is then evaluated as a time-based score which calculates the percentage of speaker time which is not mapped correctly to a reference speaker.

$$DER = \frac{\sum_s dur(s) \cdot (\max(N_{ref}(s), N_{sys}(s)) - N_{correct}(s))}{\sum_s dur(s) \cdot N_{ref}(s)} \quad (3.2)$$

where s is the longest continuous segments for which the reference and system speakers do not change, $dur(s)$ is the duration of s , $N_{ref}(s)$ is the number of reference speakers in s , $N_{system}(s)$ is the number of system speakers in s and $N_{correct}(s)$ is the number of mapped reference speakers which match the system speakers.

DER can also be expressed in an equivalent form as:

$$DER = \frac{T_{FA} + T_{miss} + T_{SpkErr}}{T_{speech}} \quad (3.3)$$

where T_{FA} , T_{miss} , T_{SpkErr} , and T_{speech} are the duration of silence misclassified as speech, the duration of speech wrongly classified as silence, the duration of speech misclassified to other speakers e.g. the speech segment belongs to speaker A but is classified to speaker B , and the total duration of speech under evaluation correspondingly. As seen from equation 3.1 and equation 3.3, DER is inclusive of SER, thus SAD is important to speaker diarization system.

3.1.4 Related Projects in Meeting Room

The Interactive Multimodal Information Management (IM2) [173] aims at the study of multimodal interaction, covering a wide range of activities and applications, including the recognition and interpretation of spoken, written and gestured languages, computer vision, and the automatic indexation and management of multimedia documents. One of the most important and challenging applications is Smart Meeting Management. The overall objective of this application is the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analysed meetings, which are captured from rooms equipped with multimodal sensors including: close-talk microphones, distant microphones, microphone arrays as well as cameras.

The Computers In the Human Interaction Loop (CHIL) [168] aims at improving the interactions between users and computers by making computers more usable and receptive to the user's needs and realizing computer services that are delivered to people in an implicit, indirect and unobtrusive way. Several intelligent meeting rooms with audio and video sensors are built where data is collected and research is performed on the lecture-type meetings.

The project Augmented Multi-party Interaction (AMI) [174,175] focuses on enhancing the productivity of meetings by changes in technologies and changes in business processes. The AMI Consortium studies the human behaviour in meetings using advance signal processing, machine learning models and social interaction dynamics. Within the scope of the project, Consortium members have developed a very large database of pre-processed meeting recordings of multiple sources of information (contained in audio, video and images captured). The actions, words and all data (slides, white board drawings and hand written notes) associated with a set of scripted meetings are captured using highly instrumented meeting rooms.

3.2 State of the art

In this section, speaker diarization systems, which have consistently performed well in the recent NIST RT evaluations, are reviewed and analyzed. Through the analysis, certain

system limitations, which lead to the motivation of this work, would be highlighted and discussed.

The AMIDA speaker diarization system [23] is one among many systems that employed the AHC framework, as described in section 2.4.4.1. BIC (section 2.4.2.5) was used for speaker segmentation and initial clustering, with small value of λ ($\lambda = 1$) to obtain *over-segmentation* and *under-clustering* initial clusters. AHC was then performed on these clusters, with CLR (section 2.4.2.6) as the distance metric and stopping criterion. In this work, the authors did not use tunable threshold for stopping criterion, but rather a theoretical motivated value of 0 (the merging process stops when the maximum value of CLRs fall below 0). This value is originated from speaker verification literature and although being hypothetically applauded, it is well known that this threshold does not work well in practice as many score calibration techniques and some tunings are required on particular dataset.

At the front-end, the AMIDA system extracted the speech frames by applying a two-state HMM for SAD, with diagonal covariance GMMs for speech and silence state. MFCCs (section 2.2.1.1) were used to perform speaker diarization for SDM condition, while both MFCCs and TDOA (section 2.2.2) features obtained from BeamformIt [27] were used for MDM condition. Fixed stream weighting scheme was employed where the weight for MFCC stream was set to 0.9 ($\alpha = 0.9$) and the weight for TDOA stream was set to 0.1. These weights were obtained with the help of development data.

Another state of the art system, which also utilized the AHC framework, is the ICSI speaker diarization system [11]. With different initialization method, the ICSI system relies on long-term features, and based on some heuristics, the system determines both the number of initial clusters and the number of mixtures in each initial cluster. The modified d_{BIC} metric described in Equation (2.40) was used to determine which clusters to merge as well as when to stop merging clusters. In this strategy, the total number of free parameters must be kept constant after merging two clusters (i.e. in case of modeling with GMM diagonal covariance, this means that the number of mixtures after merging is equal to the sum of the number of mixtures in each of the clusters to be merged). By satisfying this condition, it is noted that the term $\Delta M_{ij} = 0$ in Equation (2.39), and in theory, this modified d_{BIC} is not depending on λ , as opposed to the conventional BIC.

The clustering and merging process would continue until the maximum value of d_{BIC} (for all pairs of cluster) falls below 0. Although the empirical results show that the overall performance of the ICSI is one of the best, it is not clear whether the system is really threshold free or the threshold is implicitly specified by the number of initial clusters and the number of mixtures. To my knowledge, there is no analytical or experimental analysis on the behavior of the modified BIC metric regarding the cluster size and the number of mixtures. And, it seems that the stopping criterion could not be easily adapted to other clustering framework, i.e. how would it work if the top-down clustering framework is used ?

Similar to most state of the art system, the hybrid SAD was implemented in the ICSI speaker diarization system. The implementation is more complicated than the AMIDA's approach, as there are 3 initial models: speech, silence with low energy and silence with high energy and high zero crossing rate. This SAD module would be able to detect audible non-speech sounds.

The IDIAP speaker diarization system [36] used an agglomerative information bottleneck (aIB) approach, as described in Section 2.4.4.1, with multiple features streams including: MFCC, TDOA, modulation spectrum, and frequency domain linear prediction features. To combine these features streams, the weights were tuned by selecting the ones which minimized the DER on development data.

The LIA-Eurocom system [77] followed the divisive hierarchical clustering scheme as presented in Section 2.4.4.2. However, only one feature stream, which is the normalized Linear Frequency Cepstrum Coefficients (LFCCs), was used in both SDM and MDM conditions. In this system, the model-based SAD described in Section 2.3.2 with two models: speech and non-speech, was employed to filter out silence segments. At the clustering stage, the system would iteratively analyze each cluster to decide if the new speaker model should be added according to some heuristics on the speaker segment duration. Once there is no more minimum candidate segments which may be used to add a new speaker, the stopping criterion is reached [176]. Could this stopping criterion be used in other clustering framework? The answer is not obvious, since it seems that the technique is rather particular to LIA current framework.

Having analyzed the state of the art speaker diarization systems, it is observed that there are several recurrent issues which are common to all or some systems. These issues together with our motivations are presented below:

1. **Selecting the proper distance metric:** the metric should be robust against various factors including data size, and which could be used in either segmentation, initialization or clustering stage.
2. **Selecting the effective stopping criterion:** ideally, the criterion could choose the optimal partitioning without depending on any particular clustering framework and which does not rely on any development data.
3. **Selecting the appropriate streams weights:** the weights should be automatically estimated without external development sets.

The abovementioned issues and motivations prompt for solutions and this work attempts to address these in the subsequent chapters.

3.3 Baseline speaker diarization systems

To address the robustness issues which were mentioned in the previous section, it is therefore necessary to set up a competitive base line speaker diarization system. In this work, the IAHC framework was chosen, since it is still by far the most popular clustering framework. Other modules including the feature extraction, speech enhancement and speech activity detection were largely adapted from the state of the arts.

3.3.1 Single distance microphone speaker diarization system

Figure 3.1 shows the typical structure of the state-of-the-art IAHC speaker diarization systems using a single distant microphone.

3.3.1.1 Front end processing

The speech signals are recorded using multiple sensors from close-talking microphones to distant microphones and microphone arrays. The signals are typically sampled at 16

Figure 3.1: *Typical structure of the state-of-the-art single channel speaker diarization systems*

kHz and each sample is encoded with 16 bits. For some meeting rooms equipped with microphone arrays, the sampling rate could be up to 48 kHz or 96 kHz. In a single channel speaker diarization system, generally the most centrally located distant microphone is selected for processing.

As mentioned previously, the signals recorded with distant microphone have low SNR and to improve this, Wiener filtering is generally applied. In the speaker diarization community for the NIST Rich Transcription evaluation, the Qualcomm-ICSI-OGI front end [21] is commonly used to perform Wiener filtering, with the assumption that the noise is additive and uncorrelated.

Once the speech signal is enhanced by Wiener filtering, the output signal from the filter is then used to extract the acoustic features. Similar to other tasks in speaker recognition, short-term spectral features are also prominently used in speaker diarization. Particularly, in meetings domain, *Mel-frequency cepstrum coefficients* (MFCC) are widely adopted by many systems including the state-of-the-arts. Normally, the time sampled speech signal is segmented into overlap frames of 20ms to 30ms in length and 10ms shifting between subsequent frames. The frames are then multiplied with a window function such as Hann [177, pp. 447-448] or Hamming [177, pp. 447-448]. A MFCC feature vector is calculated for each windowed frame, and only the first N coefficients are retained, N is typically set to 16 – 19 for speaker diarization task, since it has been shown that higher cepstral coefficients capture speaker characteristics and are therefore helpful for speaker diarization. For the baseline system, the enhanced signal was processed to extract 19 MFCC features using the windowed frame of 30ms and 10ms shifting.

3.3.1.2 Speech activity detection

The speech activity detector implemented in the baseline system, which is a hybrid energy and model based detector, was adopted from [10]. The procedures are summarized here for clarity. For each audio frame (window size is 30ms with shift size is 15ms), 36 MFCC features are extracted (12 MFCCs with their first and second order derivatives), frame energy, and the zero crossing rate. The initial speech GMM $\Theta_S^{(0)}$ and non-speech GMM $\Theta_{NS}^{(0)}$ are trained with EM algorithm using all the MFCC features extracted from each recording. The 10 percent frame features with the highest energies and relative low zero-crossing rates are selected as training data set for the bootstrap speech model $\Theta_S^{bootstrap}$,

while the 20 percent frame features with the lowest energies and relative higher zero-crossing rates as the training data for bootstrap non-speech model $\Theta_{NS}^{bootstrap}$ [15]. The speech and non-speech are modeled using GMM with 16 and 4 mixture components, correspondingly. Based on these bootstrap models, all the frames are re-classified to either speech or non-speech using maximum likelihood criterion. These classified frames are then used to re-train the speech model $\Theta_S^{(1)}$ and non-speech model $\Theta_{NS}^{(1)}$ by MAP adaptive training from $\Theta_S^{(0)}$ and $\Theta_{NS}^{(0)}$ respectively. The algorithm then iteratively performs as follows:

1. Using the speech model $\Theta_S^{(i)}$ and non-speech model $\Theta_{NS}^{(i)}$ to classify each frame to either speech or non-speech where i indicates the iteration number.
2. With the newly classified frames, re-estimate the speech model $\Theta_S^{(i+1)}$ and non-speech model $\Theta_{NS}^{(i+1)}$ by MAP adaptive training from $\Theta_S^{(i)}$ and $\Theta_{NS}^{(i)}$ respectively.
3. Step (1) and (2) are repeated until the relative change of detected speech to non-speech ratio is less than 1%. Typically, the procedure converges within 10 iterations.

3.3.1.3 Clusters initialization

The uniform initialization approach was employed in the baseline system. The speech frames obtained from the output of SAD module were grouped uniformly into 20 initial clusters. Basic refinement was performed to obtain higher cluster purity. A GMM with 16 mixtures was trained for each cluster using EM algorithm and a HMM is constructed with duration constraint of 0.5 second with insertion penalty of -150 to penalize the switching of speakers. The iterative process of Viterbi decoding and models retraining were repeated several times until the frame labels are stabilized. The values chosen here are based on the observations from development data.

3.3.1.4 Iterative agglomerative hierarchical clustering framework for single distant microphone

The IAHC was initially proposed by Ajmera and Wooters [38] for speaker clustering and since then, the approach has been adopted by many others [11, 15, 22, 23, 178]. For completeness, this section provides the detail description of the IAHC clustering procedures.

To facilitate the description, some notations are to be defined:

- N_s the actual number of speakers in the audio recording
- K_i the number of clusters at iteration i
- $C_k^{(i)}$ the k th cluster at iteration i , with $k = 1, \dots, K_i$
- $X_k^{(i)}$ the set of feature vectors assigned to cluster $C_k^{(i)}$ at iteration i

Initially, right after the clusters initialization step, $i = 0$ and K_0 initial clusters are given. The IAHC procedures are as follows:

1. For each cluster $C_k^{(i)}$, a *Gaussian mixture model* (GMM) $\Theta_k^{(i)}$ is trained using all the feature vectors $X_k^{(i)}$ with $k = 1, \dots, K_i$. The training algorithm is *expectation maximization* (EM) [89] or *maximum a posteriori* (MAP) [89] adaption from a common *universal background model* (UBM). Typically, for MAP adaptive training, only the mean vectors are updated while the mixture weights and covariance matrices are unaltered.
2. Construct a *hidden Markov model* (HMM) $\Lambda^{(i)}$ with duration constraint for each cluster $C_k^{(i)}$ as shown in Figure 3.2 [84]. Each cluster $C_k^{(i)}$ is modeled by a HMM $\Lambda_k^{(i)}$ with S sub-states, where S determines the minimum duration of each speaker segment. Typically, the segments are constrained to be about 3 seconds, corresponding to approximately $S = 300$ sub-states. Each sub-state shares the same GMM $\Theta_k^{(i)}$ estimated from data of their respective cluster $C_k^{(i)}$ in the previous step.
3. Perform *Viterbi decoding* [83] on the feature stream to obtain the new cluster assignment $X_k^{(i+1)}$.
4. Repeat step (1) to (3) until the cluster assignment is stabilized (typically when the change in cluster membership or when the improvement in likelihood score is insignificant).
5. Assuming that it is now at iteration j , the distance between every pairs of clusters are computed:

$$d_{k,l}^{(j)} = \text{dist} \left(C_k^{(j)}, C_l^{(j)} \right) \quad (3.4)$$

Figure 3.2: *HMM topology for speaker clustering with duration constraint*

where $dist$ denotes the distance metric, which can be one of those presented in section 2.4.2. The pair with minimum distance is then identified:

$$d_{min}^{(j)} = \min_{1 \leq k < l \leq K_j} d_{k,l}^{(j)} \quad (3.5)$$

$$(k_{min}, l_{min}) = \arg \min_{1 \leq k < l \leq K_j} d_{k,l}^{(j)} \quad (3.6)$$

6. Check the stopping criterion. There are various stopping criteria proposed in the literature, the popular method is thresholding: $d_{min}^{(j)}$ is compared to a predetermined threshold T_{stop} , which is derived from development data sets. The clustering procedure stops when

$$d_{min}^{(j)} > T_{stop}$$

The other popular stopping criterion is *Bayesian information criterion* (BIC) which is presented in section 2.4.2.5. The clustering procedure stops when

$$dist_{BIC} (C_{k_{min}}^{(j)}, C_{l_{min}}^{(j)}) > 0$$

If the stopping criterion is not satisfied, cluster $C_{k_{min}}^{(j)}$ and $C_{l_{min}}^{(j)}$ are merged together to form a new combined cluster. Obviously, after this step the number of cluster $K_{j+1} = K_j - 1$. Step (1) to step (6) are then repeated until the stopping criterion is met or until the clustering algorithm converges to the trivial case of one cluster.

For the baseline system, GLR metric was used as cluster distance measure and GLR threshold was used as stopping criterion. In the main procedure, each cluster was modeled by a GMM with 32 mixtures and the same HMM topology as in the initialization step was employed for iterative clustering.

3.3.2 Multiple distant microphones speaker diarization system

The speaker diarization system using multiple distant microphones are similar to the SDM system in many aspects. Many components are shared among the systems including: Wiener filtering, acoustic feature extraction, speech activity detection, and clustering framework. The notable differences are the acoustic enhancement with beamforming and the addition of the TDOA feature stream. Figure 3.3 shows the structure of the baseline IAHC speaker diarization systems using multiple distant microphones. The highlighted components are distinctive to the MDM system and will be presented in the following sections. For the other components, which are similar to the SDM system, readers could refer to section 3.3.1.1 for the details.

3.3.2.1 Front-end processing

Wiener filtering was first applied to all the microphone channels and the BeamformIt toolkit [27] then performed beamforming and extracted time delay features using the filtered signals. For beamforming, the toolkit parameters were set to default values with the window size of 500ms and overlapping of 250ms. To extract the delay features, however, the window size was set to 500ms with the shifting size of 10ms to ensure synchronization with the acoustic features. The beamformed signal was then processed to extract 19 MFCC features using the windowed frame of 30ms and 10ms shifting.

Figure 3.3: *Typical structure of the state-of-the-art multichannel speaker diarization systems*

3.3.2.2 Clusters initialization

Following the same approach as the SDM system, the speech frames obtained from the output of SAD module are grouped uniformly into 20 initial clusters. Each cluster is then modeled by two GMMs: one for acoustic features with 16 mixtures and one for delay features with 1 mixture. Similarly, two HMMs are constructed with duration constraint of 0.5 second with insertion penalty of -150. Several iterations of multi-stream Viterbi re-segmentation and models retraining are then performed until the frame labels are stabilized. The stream weights are empirically set to 0.9 for acoustic stream and 0.1 for delay stream as these weights were shown to be the best combination with development data.

3.3.2.3 Iterative agglomerative hierarchical clustering framework for multiple distant microphones

Following the same convention as in the SDM systems, let:

- N_s be the actual number of speakers in the audio recording
- K_i be the number of clusters at iteration i
- $C_k^{(i)}$ be the k th cluster at iteration i , with $k = 1, \dots, K_i$
- $X_k^{(i)}$ be the set of acoustic feature vectors assigned to cluster $C_k^{(i)}$ at iteration i
- $Y_k^{(i)}$ be the set of delay feature vectors assigned to cluster $C_k^{(i)}$ at iteration i

Initially, right after the clusters initialization step, $i = 0$ and K_0 initial clusters are given. The IAHC procedures for MDM system are summarized as follows:

1. For each cluster $C_k^{(i)}$, GMM $\Theta_{X,k}^{(i)}$ and GMM $\Theta_{Y,k}^{(i)}$ are trained using all the acoustic feature vectors $X_k^{(i)}$ and all the delay feature vectors $Y_k^{(i)}$, correspondingly, with $k = 1, \dots, K_i$.
2. Construct HMM $\Lambda_X^{(i)}$ and HMM $\Lambda_Y^{(i)}$ with duration constraint for each cluster $C_k^{(i)}$ as shown in Figure 3.2. For HMM $\Lambda_X^{(i)}$ and $\Lambda_Y^{(i)}$, each cluster $C_k^{(i)}$ is modeled by a HMM $\Lambda_{X,k}^{(i)}$ and $\Lambda_{Y,k}^{(i)}$ with S sub-states, respectively. Each sub-state of $\Lambda_{X,k}^{(i)}$ and $\Lambda_{Y,k}^{(i)}$ shares the same GMM $\Theta_{X,k}^{(i)}$ and $\Theta_{Y,k}^{(i)}$ correspondingly.

3. Perform multi-stream *Viterbi decoding* [83] on the two feature streams: acoustic features and delay features to obtain the new cluster assignment $X_k^{(i+1)}$ and $Y_k^{(i+1)}$. The Viterbi decoding find the optimum path by considering the joint log-likelihood for any given frame as

$$\mathcal{L}\left(x[n], y[n] | \Theta_{X,k}^{(i)}, \Theta_{Y,k}^{(i)}\right) = W_X \mathcal{L}\left(x[n] | \Theta_{X,k}^{(i)}\right) + W_Y \mathcal{L}\left(y[n] | \Theta_{Y,k}^{(i)}\right) \quad (3.7)$$

where $x[n]$, $y[n]$ are the acoustic feature vector and delay feature vector at frame n respectively. W_X and W_Y are the weighting of acoustic feature stream and delay feature stream correspondingly, with $W_X + W_Y = 1$. The weights are to reflect the contribution and usefulness of each feature stream, they may be fixed by using development data to calibrate. In [28, 146], an adaptive method for estimation the stream weight was proposed with promising result on development set, however it did not generalize well on test data as the performance was worse than fixed weighting. More details on this technique are given in section 5.3 as it will be compared to a novel weight adaptation technique proposed in this thesis.

4. Repeat step (1) to (3) until the cluster assignment is stabilized.
5. Assuming that it is now at iteration j , the distance between every pairs of clusters are computed:

$$d_{k,l}^{(j)} = \text{dist}\left(C_k^{(j)}, C_l^{(j)}\right) \quad (3.8)$$

where *dist* denotes the distance metric, which can be one of those presented in section 2.4.2. With multiple feature streams, the distance are computed as

$$d_{k,l}^{(j)} = W_X \cdot \text{dist}\left(X_k^{(j)}, X_l^{(j)}\right) + W_Y \cdot \text{dist}\left(Y_k^{(j)}, Y_l^{(j)}\right) \quad (3.9)$$

with W_X and W_Y being the weights of acoustic feature stream and delay feature stream respectively, with $W_X + W_Y = 1$. Typically, the same set of weights are used for both Viterbi decoding and distance computation. The pair with minimum distance is then identified:

$$d_{min}^{(j)} = \min_{1 \leq k < l \leq K_j} d_{k,l}^{(j)} \quad (3.10)$$

$$(k_{min}, l_{min}) = \arg \min_{1 \leq k < l \leq K_j} d_{k,l}^{(j)} \quad (3.11)$$

6. Check the stopping criterion. If the stopping criterion is not satisfied, cluster $C_{kmin}^{(j)}$ and $C_{lmin}^{(j)}$ are merged together to form a new combined cluster. Step (1) to step (6) are then repeated until the stopping criterion is met or until the clustering algorithm converges to the trivial case of one cluster.

The baseline system implements used GLR metric as cluster distance measure and GLR threshold as stopping criterion. For the main iterative clustering loop, each cluster is also modeled by two GMMs: one for acoustic features with 32 mixtures and one for delay features with 2 mixtures. The same HMM topology and stream weights as in the initialization step is employed for iterative clustering.

3.4 Experiments

This section reports the performance of baseline speaker diarization systems and these results are to be used as references in the later chapters. Table 3.4, and Table 3.5 show the SER for the baseline single channel system and multichannel system respectively. To better demonstrate the complete picture, the diarization performance the SDM and MDM baseline system are plotted in Figure 3.4 and Figure 3.5, correspondingly, at all possible stopping thresholds. It is noted that the MDM system extracted the acoustic features from the beamformed channel while the SDM system extracted from a single Wiener filtered channel. There seems not much advantage to apply beamforming just for the sake of SAD since the SER of two systems are almost identical. As for the effect of beamformed signal on the diarization performance, no significant influence was reported in the literature [57]. On the other hand, the positive contribution of delay features are clearly perceived as reflecting in the performance of MDM system, which performed much better than SDM system. It may be perceived from the obtained results that the inclusion of delay feature stream did not help in RT 2005 and RT 2006 data sets as the error rates of the two systems differ negligibly. However, it is shown in later chapters that these results are due to the limitation of using GLR as threshold for stopping, detailed analyses and comparisons are given in section 5.4.1.

Table 3.4: *Performance of Speech/Non-speech Detector of SDM system*

	RT 2005	RT 2006	RT 2007	RT 2009
SER (%)	3.2	4.8	3.4	2.6

Table 3.5: *Performance of Speech/Non-speech Detector of MDM system*

	RT 2005	RT 2006	RT 2007	RT 2009
SER (%)	3.0	4.5	3.1	2.7

Figure 3.4: *Speaker diarization performance of SDM system with various GLR stopping thresholds.*

Figure 3.5: *Speaker diarization performance of MDM system with various GLR stopping thresholds.*

3.5 Summary

This chapter reviews the state-of-the-art speaker diarization systems. Throughout the review, it is highlighted that some common robustness issues such as: threshold for determining number of speakers and stream weights need to be learned from development data. Thus, the performance on unseen data is a concern and several robust techniques would be investigated in subsequent chapters to address these issues.

This chapter also presents the baseline speaker diarization systems: one for single channel and one for multi-channel speaker diarization. Both systems employ the IAHC framework, and as a result, they share many common components. The major differences between the two systems is the addition of the delay feature stream in the MDM system. The results of these systems are also presented and would serve as the baseline and reference for subsequent chapters. The major differences between the two systems is the addition of the delay feature stream in the MDM system. In overall, by combining both acoustic features and delay features the MDM systems perform better than the SDM system.

Chapter 4

Analysis of distance measures

In many applications, there is a requirement to measure the similarity between two or more identities. For example, in information retrieval, the user inputs a text or voice query, the system returns the documents containing the similar texts or voice segments. In speaker verification, the similarity between a speech utterance and a speaker model is evaluated to verify whether the utterance is spoken by the speaker which the model represents. In agglomerative hierarchical clustering, the similarities among all possible pairs of clusters are computed and similar pairs which satisfy a given threshold are merged. The need to quantify these similarities requires the definition of distance functions, which are formulated for each problem specifically. In the scope of this thesis, the focus is on distance functions for speaker clustering.

In state-of-the-art speaker diarization systems, *iterative agglomerative hierarchical clustering* (IAHC) framework is the most commonly used clustering architecture and in the IAHC framework, the critical element is the distance metric since it is used to measure the similarities among pairs of clusters for merging. The *Generalized Likelihood Ratio* (GLR) based distance metrics are commonly used in many state-of-the-art systems and the first part of the chapter (section 4.1) is reserved to analyze these widely used metrics, their weaknesses and discuss various attempts to tackle them. As a result, a novel distance metric is proposed in section 4.2 to address existing issues in the GLR-based metric. The experimental results are then reported in section 4.3 and finally the chapter is summarized in section 4.4.

4.1 Analysis of GLR-based metrics

4.1.1 Motivation

GLR-based distance metrics such as GLR and BIC are widely used in many state-of-the-art speaker diarization systems. However, it is widely known that these metrics are not robust against data size variations. In Chen and Gopalakrishnan [102], it is shown that BIC value increases according to data size. In Han et al. [14], the authors showed that GLR and BIC values are linearly proportional to data size with firm mathematical proof, at least in the special case of normal distribution. Figure 4.1a and Figure 4.1b show the relationship between average inter-speaker distances and average intra-speaker distances with different cluster sizes using GLR and BIC metric correspondingly. In this context, inter-speaker distance refers to the distance between two cluster, where each cluster belongs to a different speaker, while intra-speaker distance refers to the distance between two clusters of the same speaker. As can be seen from the figures, both BIC and GLR values increase as the data size increases. This presents in general a problem when there is a big mismatch between clusters or windows with different data sizes. It is observed from the figures that at the values indicated by the dash line, it is impossible to tell whether the distance belongs to the intra-speaker class or inter-speaker class. Thus, when measuring the distances among pairs of clusters to decide which pair to be merged, it is probable that the decision is wrong since there are certain distance values which it is not possible to ascertain whether the two clusters belong to the same speaker or from two different speakers. To mitigate this issue, one may suggest to compare clusters of the same size, however, this is not practical in IAHC framework. Even with uniform initialization, after a few iterations of merging and re-segmentation, the clusters are typically of different sizes. The difference might be significant, as not everyone would speak the same amount of time, someone may talk more while the others talk less. To confirm the observations, a mathematical proof will be given in this section to show that GLR and BIC values are indeed linearly proportional to the data size. The proof presented here does not make assumption of Gaussian models as opposed to the one shown in Han et al. [14].

4.1.2 Generalized Likelihood Ratio

Given two clusters $C_{\mathbf{x}}$ and $C_{\mathbf{y}}$ with their feature vectors $\mathbf{x} = \{x_1, \dots, x_M\}$ and $\mathbf{y} = \{y_1, \dots, y_N\}$, respectively, the following hypothesis tests are considered:

- H_0 : both clusters are from the same speaker.
- H_1 : each cluster is from a different speaker.

The feature vectors of each speaker K are assumed to be distributed according to the generating *probability density function* g_K .

- Under hypothesis H_0 : $\mathbf{x} \cup \mathbf{y} \sim g_{XY}$
- Under hypothesis H_1 : $\mathbf{x} \sim g_X$ and $\mathbf{y} \sim g_Y$

Since the generating density functions g_X , g_Y , and g_{XY} are unknown, these functions are therefore required to be estimated from the observed data by means of *maximum likelihood* (ML) optimization. Denote f_X , f_Y , and f_{XY} respectively the ML estimated models of the generating densities g_X , g_Y , and g_{XY} . The *Generalized Likelihood Ratio* (GLR) between two hypotheses is then defined by:

$$R = \frac{\mathcal{L}(\mathbf{x} \cup \mathbf{y} | f_{XY})}{\mathcal{L}(\mathbf{x} | f_X) \cdot \mathcal{L}(\mathbf{y} | f_Y)} \quad (4.1)$$

The feature vectors are assumed to be *independently and identically distributed* (i.i.d.), thus:

$$R = \frac{\prod_{i=1}^M f_{XY}(x_i | \theta_{f_{XY}}) \prod_{i=1}^N f_{XY}(y_i | \theta_{f_{XY}})}{\prod_{i=1}^M f_X(x_i | \theta_{f_X}) \prod_{i=1}^N f_Y(y_i | \theta_{f_Y})} \quad (4.2)$$

being θ_{f_X} , θ_{f_Y} , and $\theta_{f_{XY}}$ the parameter sets of the models f_X , f_Y , and f_{XY} correspondingly. The distance d_{GLR} is the negative logarithm of the previous expression:

$$d_{GLR} = -\log R \quad (4.3)$$

$$\begin{aligned} &= \sum_{i=1}^M \log f_X(x_i | \theta_{f_X}) + \sum_{i=1}^N \log f_Y(y_i | \theta_{f_Y}) \\ &\quad - \sum_{i=1}^M \log f_{XY}(x_i | \theta_{f_{XY}}) - \sum_{i=1}^N \log f_{XY}(y_i | \theta_{f_{XY}}) \end{aligned} \quad (4.4)$$

Considering:

$$-\frac{1}{M} \sum_{i=1}^M \log f_X(x_i|\theta_{f_X}) = \frac{1}{M} \sum_{i=1}^M [-\log f_X(x_i|\theta_{f_X}) + \log g_X(x_i) - \log g_X(x_i)] \quad (4.5)$$

$$= \frac{1}{M} \sum_{i=1}^M \left[\log \frac{g_X(x_i)}{f_X(x_i|\theta_{f_X})} - \log g_X(x_i) \right] \quad (4.6)$$

$$= \frac{1}{M} \sum_{i=1}^M \log \frac{g_X(x_i)}{f_X(x_i|\theta_{f_X})} - \frac{1}{M} \sum_{i=1}^M \log g_X(x_i) \quad (4.7)$$

Notice that the *Kullback-Leibler divergence* or the *relative entropy* [105] between g_X and f_X which reflects the separation between these two distributions is defined as:

$$D_{KL}(g_X||f_X) = \mathbf{E}_g \left[\log \left(\frac{g_X(x)}{f_X(x)} \right) \right] \quad (4.8)$$

where $\mathbf{E}_g[\cdot]$ denotes the expectation under the generating density function. By the *strong law of large numbers* [105], the first term in equation (4.7) converges to D_{KL} :

$$\Pr \left(\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{i=1}^M \log \frac{g_X(x_i)}{f_X(x_i|\theta_{f_X})} = D_{KL} \right) = 1 \quad (4.9)$$

It is known that Kullback-Leibler divergence is non-negative and is zero only when $f_X = g_X$ almost surely. Examine the second term in equation (4.7), the *asymptotic equipartition property* (AEP) [105] states that:

$$-\frac{1}{M} \sum_{i=1}^M \log g_X(x_i) \rightarrow \mathbf{H}(g_X) \quad (4.10)$$

in probability with $\mathbf{H}(g_X)$ denoting the entropy or differential entropy depending on whether g_X is discrete or continuous respectively. In other words:

$$\lim_{n \rightarrow \infty} P \left(\left| -\frac{1}{M} \sum_{i=1}^M \log g_X(x_i) - \mathbf{H}(g_X) \right| > \epsilon \right) = 0 \quad (4.11)$$

with every $\epsilon > 0$. Thus, it is observed that the mean negative log-likelihood converges to the differential entropy under the generating distribution g_X plus the Kullback-Leibler divergence between the generating distribution g_X and the estimated distribution f_X . In the work of Cavanaugh [179], it was shown that:

$$\mathbf{E}_g [-\log g_X(x)] - \mathbf{E}_g [-\log f_X(x|\theta_{f_X})] = \frac{k}{2} + \mathcal{O}(1) \quad (4.12)$$

under the strong assumption that g_X and f_X must belong to the same parametric class $\mathcal{F}(k)$ where k denotes the number of free parameters. That is to say, if the generating function is not known, at least the parameter class and the number of free parameters of this function must be known for the equation (4.12) to be justifiable. From all the assumptions and aforementioned derivations, it is shown asymptotically that:

$$\mathbf{E}_g [-\log f_X(x|\theta_{f_X})] = \mathbf{E}_g [-\log g_X(x)] - \frac{k}{2} + \mathcal{O}(1) \quad (4.13)$$

$$= \mathbf{H}(g_X) - \frac{k}{2} + \mathcal{O}(1) \quad (4.14)$$

Back to the discussion of GLR metric, with the assumption that all density functions having the same number of free parameters k and all the equations are understood in the asymptotic sense, it can be seen that:

$$\begin{aligned} d_{GLR} &= -M\mathbf{H}(g_X) + \mathcal{O}(M) - N\mathbf{H}(g_Y) + \mathcal{O}(N) \\ &\quad + (M + N)\mathbf{H}(g_{XY}) + \mathcal{O}(M + N) \end{aligned} \quad (4.15)$$

1. *If cluster C_X and C_Y are from the same speaker:*

$$\begin{aligned} \mathbf{H}(g_X) &= \mathbf{H}(g_Y) = \mathbf{H}(g_{XY}) \\ d_{GLR} &= \mathcal{O}(M) + \mathcal{O}(N) + \mathcal{O}(M + N) \end{aligned} \quad (4.16)$$

2. *If cluster C_X and C_Y are from different speakers: without loss of generality, assuming that $\mathbf{H}(g_X) \geq \mathbf{H}(g_Y)$*

$$\mathbf{H}(g_{XY}) \leq \mathbf{H}(g_X) + \mathbf{H}(g_Y) \quad (4.17)$$

$$\mathbf{H}(g_{XY}) \geq \max\{\mathbf{H}(g_X), \mathbf{H}(g_Y)\} = \mathbf{H}(g_X) \quad (4.18)$$

$$\begin{aligned} d_{GLR} &\leq N\mathbf{H}(g_X) + M\mathbf{H}(g_Y) + \mathcal{O}(M + N) \\ &\leq (N + M)\mathbf{H}(g_X) + \mathcal{O}(M + N) \end{aligned} \quad (4.19)$$

$$d_{GLR} \geq N[\mathbf{H}(g_X) - \mathbf{H}(g_Y)] + \mathcal{O}(M + N) \quad (4.20)$$

As can be seen from equation (4.16) to equation (4.20), the GLR metric is linearly proportional to the cluster sizes.

$$d_{GLR} \propto (M + N) \quad (4.21)$$

4.1.3 Bayesian information criterion

For a more comprehensive overview of BIC, please refer to section 2.4.2.5. In brief, the BIC metric is a variation of GLR with the additional penalty term:

$$d_{BIC} = d_{GLR} - \lambda \frac{1}{2} (\Delta \mathcal{M}) \log(M + N) \quad (4.22)$$

where $\Delta \mathcal{M}$ is the difference between the number of free parameters of models in hypothesis H_1 and hypothesis H_0 . In other words:

$$\Delta \mathcal{M} = |\theta_{f_X}| + |\theta_{f_Y}| - |\theta_{f_{XY}}| \quad (4.23)$$

where $|\theta_{f_X}|$, $|\theta_{f_Y}|$, and $|\theta_{f_{XY}}|$ denote the number of free parameters in the distribution functions f_X , f_Y , and f_{XY} respectively and λ is an adjustable constant, optimized for each data set specifically. As can be seen from equation (4.22):

$$\lambda \frac{1}{2} (\Delta \mathcal{M}) \log(M + N) \propto \log(M + N) \quad (4.24)$$

It is shown in previous section that

$$d_{GLR} \propto (M + N) \quad (4.25)$$

Thus

$$d_{BIC} \propto (M + N) \quad (4.26)$$

4.1.4 Information change rate

This section examines the ICR [14] metric, which attempts to tackle the effect of data size variations by normalizing with the number of frames. In short:

$$d_{ICR} = \frac{d_{GLR}}{M + N} \quad (4.27)$$

where M and N are the number of frames in each cluster. Theoretically, this normalization should help to eliminate the discrepancy due to size, especially when M and N are sufficiently large.

4.2 Proposed distance measure

Using the same notations as previous section, denoting:

$$\mathbf{z}_1 = \{\log f_X(x_i|\theta_{f_X}), \forall x_i \in \mathbf{x}\} \cup \{\log f_Y(y_j|\theta_{f_Y}), \forall y_j \in \mathbf{y}\} \quad (4.28)$$

$$\mathbf{z}_2 = \{\log f_{XY}(x_i|\theta_{f_{XY}}), \forall x_i \in \mathbf{x}\} \cup \{\log f_{XY}(y_j|\theta_{f_{XY}}), \forall y_j \in \mathbf{y}\} \quad (4.29)$$

Regarding $\mathbf{z}_1$ and $\mathbf{z}_2$ as the sequences of i.i.d. random variables drawn according to the probability density functions of random variable Z_1 and Z_2 correspondingly. From the basic property of variance:

$$\text{Var}(Z_1 - Z_2) = \text{Var}(Z_1) + \text{Var}(Z_2) - 2\text{Cov}(Z_1, Z_2) \quad (4.30)$$

with $\text{Var}(\cdot)$ and $\text{Cov}(\cdot)$ be the variance and covariance operator respectively. If $\mathbf{x}$ and $\mathbf{y}$ belong to the same speaker then $f_X = f_Y = f_{XY}$ and

$$\text{Var}(Z_1 - Z_2) = 0 \quad (4.31)$$

On the other hand, if $\mathbf{x}$ and $\mathbf{y}$ belong to different speakers then $f_X \neq f_{XY}$, $f_Y \neq f_{XY}$ and

$$0 < \text{Var}(Z_1 - Z_2) \leq \text{Var}(Z_1) + \text{Var}(Z_2) \quad (4.32)$$

From the above observation, it is expected that the variance is also a useful metric to indicate whether the two clusters belong to the same speakers or not. Refer to the ICR metric presented in the previous section, it can be expressed as

$$d_{ICR} = \frac{1}{|\mathbf{z}_1|} \sum_{z \in \mathbf{z}_1} z - \frac{1}{|\mathbf{z}_2|} \sum_{z \in \mathbf{z}_2} z \quad (4.33)$$

Let

$$\mathbf{s}_1 = \{\log f_X(x_i|\theta_{f_X}) - \log f_{XY}(x_i|\theta_{f_{XY}}), \forall x_i \in \mathbf{x}\} \quad (4.34)$$

$$\mathbf{s}_2 = \{\log f_Y(y_j|\theta_{f_Y}) - \log f_{XY}(y_j|\theta_{f_{XY}}), \forall y_j \in \mathbf{y}\} \quad (4.35)$$

The ICR metric can also be expressed as

$$\mathbf{s} = \mathbf{s}_1 \cup \mathbf{s}_2 \quad (4.36)$$

$$d_{ICR} = \mu_s = \frac{1}{|\mathbf{s}|} \sum_{s \in \mathbf{s}} s \quad (4.37)$$

where $|\mathbf{s}|$ be the number of elements in $\mathbf{s}$. It is noted that $\mathbf{s}$ can be regarded as the i.i.d. random variables drawn according to the probability density functions of random variable $Z_1 - Z_2$ and

$$\sqrt{\text{Var}(Z_1 - Z_2)} = \sigma_{\mathbf{s}} = \sqrt{\frac{1}{|\mathbf{s}| - 1} \sum_{s \in \mathbf{s}} (s - \mu_{\mathbf{s}})^2} \quad (4.38)$$

The ICR metric can be seen as the mean, the first-order statistic, of the random variable $Z_1 - Z_2$. It is shown above that the variance is also a useful indicator of speaker similarity, therefore a novel distance measure is proposed to exploit both the first-order and second-order statistic

$$d_s = \mu_{\mathbf{s}} + \sigma_{\mathbf{s}} \quad (4.39)$$

Similar to ICR, the proposed metric is, in theory, invariant to the data sizes. Empirical evidences will be presented in the experiments for comparison of different metrics on real data set.

4.3 Experiments

4.3.1 Experiments on distance measures

Objective The subsequent sets of experiments were designed with the intention to compare the effectiveness of various distance measures in terms of speaker discriminability. More precisely, consider a distance between two clusters evaluated by a given metric, is it possible to deduce from this value whether the two clusters are from the same speakers? And how accurate is this deduction? In speaker verification research, the *equal error rate* (EER) is often reported and its formal definition is presented in section 4.3.1.1. For ease of understanding, assume a distance metric is reported with an EER of k percent, the implication is that if a detection threshold is set at the EER, then using this metric to predict whether the two clusters are from the same speakers or not has the same error rate of k percent. Therefore, in the following sets of experiments, various distance measures will be compared in terms of EER.

Database The RSR2015 database [180] comprises of 300 speakers, of which 143 female and 157 male speakers. The participants were required to record in nine sessions using three portable devices. Each session consists of 73 sentences including: 30 short sentences, 20 voice commands, and 23 digit sequences.

Feature extraction and speech activity detection 19 MFCC features were extracted for each speech segment using window size of 30ms and shift size of 10ms. A simple energy based speech/non-speech detector was then applied to discard non-speech frames. This technique is satisfactory for this database since the recording condition is clean.

Feature extraction and speech activity detection 19 MFCC features were extracted for each speech segment using window size of 30ms and shift size of 10ms. A simple energy based speech/non-speech detector was then applied to discard non-speech frames. This technique is satisfactory for this database since the recording condition is clean.

Experimental setup The experiment was limited to investigation of distances among speakers of the same gender since it is well known that the voice characteristics of male and female are highly different. Without any preference, the female speakers were selected for the subsequent experiments. Since the original sentences are short, they are concatenated to form longer speech segments. For each speaker, every k sentences recorded from the same device were concatenated to form a new segment and this new segment was put into set S_k where $k = \{5, 10, 15\}$, which correspond to speech segments of approximately 10, 20 and 30 seconds respectively. In summary, the data set was constructed such that each speaker has 8 speech segments in S_k set. In the first experiment (section 4.3.1.2), various distance metrics including: GLR, BIC, NCLR, ICR, and d_s were used to measure the similarity between all possible pairs of segments and the results were reported. All distance values are belong to either one of the two classes: intra-speaker class (distances between segments of the same speakers) and inter-speaker class (distances between segments of different speakers).

4.3.1.1 Performance evaluation

The performance of all experiments in this chapter are reported in terms of *equal error rate* (EER) [181] or *detection error tradeoff* (DET) [182] curve.

Equal error rate *Equal error rate* (EER) is the miss rate at the operating point at which the miss rate and false alarm rate are equal. The miss rate refers to the percentage of intra-speaker distances being wrongly classified as inter-speaker class. The false alarm rate is the percentage of inter-speaker distances being wrongly classified as intra-speaker class.

DET curve Instead of reporting a single error number such as EER, the DET curve provides more information by plotting all the operating points of a system in a graph, where the two error rates: miss rate and false alarm rate are plotted on the $\mathbf{x}$ and $\mathbf{y}$ axes.

4.3.1.2 Comparative study of distance metrics

Experiment In this experiment, the mono Gaussian model with full covariance matrix was selected for modeling speaker speech segments. One of the consideration to choose this simpler model is the execution time, since the number of evaluated distances are in the order of millions. The performance reported, nevertheless, should be generalizable to more complex modeling technique such as GMM. Table 4.1 summarizes the performance of different distance metrics in terms of EER. The results are reported for individual sets S_5 , S_{10} , and S_{15} , where all segments in the same set have approximately the same length. Experiments were also carried out on mix sets to verify the robustness of each distance metric in terms of data size variations.

Table 4.1: EER(%) of different speaker distance measures

	S_5	S_{10}	S_{15}	$S_5 \cup S_{10}$	$S_5 \cup S_{15}$	$S_{10} \cup S_{15}$	$S_5 \cup S_{10} \cup S_{15}$
GLR	15.95	14.58	13.54	20.28	25.10	15.79	22.83
BIC	15.76	14.62	13.54	19.65	23.93	15.63	21.93
NCLR	21.74	20.57	19.49	21.88	21.69	20.21	21.47
ICR	17.06	16.15	14.97	17.94	19.96	15.41	18.01
d_s	16.72	14.84	13.74	16.47	16.47	14.84	16.09

(a) GLR

(b) BIC

Figure 4.1: Average inter-speaker distances and intra-speaker distances with different cluster sizes using GLR and BIC metric

(a) ICR

(b) d_s

Figure 4.2: Average inter-speaker distances and intra-speaker distances with different cluster sizes using ICR and d_s metric

Figure 4.3: *DET for various distance metrics evaluated on set $\{S_5 \cup S_{10} \cup S_{15}\}$*

Discussion Figure 4.1 and Figure 4.2 show the relationship between the distance values and cluster sizes. It is evident from the figures that the GLR and BIC distances increase as cluster size increases, while the ICR and d_s distance are remaining steady when the cluster size is sufficiently large. These experimental results therefore reconfirm the analysis made in section 4.1. Figure 4.3 then plots the DET curves for different distance metrics, which were evaluated on the set $\{S_5 \cup S_{10} \cup S_{15}\}$. From the obtained results in Table 4.1, several observations can be made:

1. As expected, for experiments with consistent data size, normalization did not help, it even degraded the performance: ICR performed worse than GLR, d_s is slightly worse than GLR but better than ICR, while GLR performed equally well as BIC.
2. Normalization did help when there are variations in data size. The effect is more prominent when the inconsistency is more obvious (comparing column $S_5 \cup S_{10}$ with $S_5 \cup S_{15}$).
3. In terms of robustness to data size mismatch, GLR is the worst. BIC is slightly better than GLR, which could be explained by the existing of the penalty term $P =$

$\lambda_2^1(\Delta\mathcal{M})\log(M+N)$ in the computation of BIC. In this case, P serves as a form of normalization. Interestingly, although the performance of NCLR is not good, this metric nonetheless does not seem to be affected by data size inconsistency, probably due to the fact that NCLR is also normalized by the number of frames as seen in equation (2.42).

4. The performance of ICR is reasonably good in general with certain extent of robustness to size mismatch. However, the normalizing technique is far from ideal, which presents many opportunities for improvement.
5. d_s consistently outperforms ICR, CLR in every cases and it also outperforms other metrics when there is discrepancy in sizes.

4.3.2 Distance measures and speaker diarization performance

Objective In previous sections, various distance metrics have been studied in standalone experiments in the context similar to speaker verification. This section aims to investigate the influence of these metrics on the performance of a speaker diarization system, particularly the state-of-the-art system with IAHC framework. As noted, the IAHC framework has the capability of correcting mistakes made in early stages by means of iterative models retraining and Viterbi re-segmentation. As a consequence, it is not straight forward to relate the EER to the diarization error rate (DER) which are used to evaluate diarization performance. Nevertheless, these measurements are expected to have some correlations.

Experiments In this experiment, the baseline SDM system was used to perform speaker diarization with an oracle stopping criterion. The oracle stopping method relies on the reference labels (ground truth) to scan through all possible stopping points and select the one with minimum DER. In doing so, it is isolated the errors caused by making the wrong stopping decision, thus the obtained results could truly reflect the performance of different distance metrics. The obtained results are reported in Table 4.2 and Figure 4.4.

Table 4.2: *DER (%) of the baseline speaker diarization system with various clustering distance measures*

	RT 2005	RT 2006	RT 2007	RT 2009
GLR	11.11	16.57	12.54	17.34
BIC	10.91	16.51	12.41	17.02
NCLR	10.83	14.88	12.29	16.79
ICR	8.15	14.72	11.36	16.58
d_s	8.06	14.47	10.62	15.91

Figure 4.4: *Performance of baseline speaker diarization system with various clustering distance measures*

Discussion It is noted from the experimental results that there some correlations between the EER and the contribution of distance metrics in the performance of speaker diarization system with the exception of NCLR metric. Surprisingly, the system using NCLR metric performed better than BIC and GLR. The proposed distance measure performed consistently well across different data sets and is always better than the others. However, the improvements are not as significant as in the speaker verification task. A probable explanation for this observation is that the speaker diarization system performed multiple iterations of refinement and re-segmentation, as such even some merging mistakes occurred in the early steps, the algorithm is capable of recovering to certain extent. Another reason might be due to the number of distances considered are substantially less than in verification tasks (several hundreds comparing to several millions distances), thus it is not statistically significant to deduce any conclusions.

4.4 Summary

This chapter addressed the issue of selecting the appropriate distance measure for speaker clustering. As the common practice, GLR-based distance metrics such as GLR and BIC, are widely adopted in many speaker diarization systems. However, with the detailed analysis and proof presented in this chapter, it is shown that GLR and BIC are not robust against data size mismatch. It is therefore suggested that some forms of normalization should be applied and as a result, ICR metric and the proposed distance metric d_s are recommended. A comparative study on a real data set was then taken for validation and the obtained results confirmed the theoretical analysis. The obtained results clearly indicates that normalization helps in case of data size variation and in general, the novel metric outperforms the rest. When these distance metrics are employed in a speaker diarization system, the experimental results exhibit some correlations between the speaker discriminability of the distance metric and the diarization error rate.

Chapter 5

Robust techniques for speaker diarization

Current state-of-the-art speaker diarization systems are rather complex with many integrated components. Different components may have different parameter sets, which need to be defined appropriately, for the system to achieve the best possible results and the tuning of parameters is typically done with the help of development data set. Relying on external data, however, imposes certain limitation on the performance of the systems, especially on unseen data that do not match with development set. This chapter aims to study several techniques for speaker diarization, with the focus on robustness. In this context, robustness maybe interpreted as: (1) *less dependent on development data*, and/or (2) *adaptive to the current states*, e.g. the parameters are not fixed but adjusted accordingly, dependent on the meeting recordings, clustering iterations, etc. The topics that are covered in this chapter include: (1) *stopping criteria*, (2) *spectral subspace*, and (3) *automatic feature stream weights estimation*.

5.1 Stopping criteria

Most state-of-the-art speaker diarization systems use an *iterative agglomerative hierarchical clustering* (IAHC) scheme for simultaneous speaker segmentation and clustering. In such approach, the systems start with a number of initial clusters which are obtained from some initialization methods. The clusters are then iteratively merged until some stopping criteria are satisfied. In each iteration, cluster membership re-assignment is usually perform to incrementally improve the performance. Generally, there are two major

issues in this clustering architecture: determining number of clusters (stopping criteria), and finding the cluster pair to be merged (merging criteria). Thus far, in chapter 4, the merging criteria have been investigated by studying various distance metrics and their characteristics. As a property of hierarchical clustering, the minimum distance between all possible pairs of clusters is monotonically increasing (in practice this might not always be true since the systems perform re-segmentation) after each iteration of merging, therefore the indicated distance metrics can also be functioned as stopping criteria by setting the threshold on their values. Selecting the appropriate threshold is typically done by minimizing the error rate on the development data set, however, the robustness of this threshold is highly dependent on the chosen distance metric and the data set itself. Figure 3.4 shows that the optimal stopping points for GLR threshold are highly varied across different data sets. Besides, Figure 5.1 demonstrates that even within a data set, the threshold may need to be tuned specifically for each meeting. The limitation of threshold method prompts for the study of threshold-free techniques, and as a result two novel stopping criteria are proposed in this section. These criteria have the simple interpretation that they aim to maximize the separation between the two distributions: one for the distances between segments of the same speakers and one for the distances between segments of different speakers.

5.1.1 Proposed criteria

Problem formulation Given N data points $\{x_1, x_2, \dots, x_N\}$ which are grouped into K disjoint clusters $\{C_1, C_2, \dots, C_K\}$. There are many possible ways to cluster these N data points into K groups. How should each partitioning be evaluated in case there is no ground truth available? Which metric should be used to rank or judge different cluster partitions? This section introduces two metrics namely T_s and ρ to measure the quality or goodness of a cluster partition. These metrics could be used as clustering criteria, where the objective is to maximize either T_s or ρ .

Let $d(x_i, x_j)$ be the similarity measure between two points x_i and x_j . Define:

$$D_{intra} = \{d(x_i, x_j) | \forall i, j \exists k : x_i \in C_k, x_j \in C_k\} \quad (5.1)$$

$$D_{inter} = \{d(x_i, x_j) | \forall i, j \exists k \neq l : x_i \in C_k, x_j \in C_l\} \quad (5.2)$$

Figure 5.1: Meetings with highly varied optimal stopping thresholds

where D_{intra} is the set of intra-cluster distances and D_{inter} is the set of inter-cluster distances.

5.1.1.1 T_s metric

Let $m_1, \sigma_1, n_1, m_2, \sigma_2, n_2$ be respectively the mean, standard deviation, size of D_{intra} and D_{inter} . T_s is defined as:

$$T_s(D_{intra}, D_{inter}) = \frac{|m_2 - m_1|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (5.3)$$

The interpretation of this metric: T_s measures the difference between the mean distance among points in the same clusters and the mean distance among points across different clusters taking into account the variance of these distances. Smaller value of T_s indicates D_{intra} is closer to D_{inter} , it implies that either the partition is not good since some data points might be assigned to the wrong clusters or the actual (ground truth) data clusters are of different densities as illustrated in Figure 5.2.

 Figure 5.2: *Clusters with different densities*

5.1.1.2 ρ metric

The ρ metric is a linear transform of Mann-Whitney U value [183]. To compute ρ , the values of $D = \{D_{intra} \cup D_{inter}\}$ are first sorted in ascending order and assign a ranking order for each element of D .

$$R_1 = \sum_{x_i \in D_{intra}} rank(x_i) \quad (5.4)$$

$$U_1 = R_1 - \frac{|D_{intra}| (|D_{inter}| + 1)}{2} \quad (5.5)$$

$$\rho = \left| \frac{U_1}{|D_{intra}| \cdot |D_{inter}|} - 0.5 \right| \times 2 \quad (5.6)$$

where $rank(x_i)$ is the order of x_i in the sorted sequence of D . $|\cdot|$ denotes the cardinal of the set. ρ can take values between 0 and 1. $\rho = 0$ represents complete overlap while $\rho = 1$ represents complete separation. This metric has a simple interpretation: ρ measures the overlap between the set of distances among points in the same clusters and the set of distances among points across different clusters based on the ranking order of these distances, the actual values of the distances are not important. ρ , unlike T_s is less sensitive to outliers. However, the separation between two sets defined by ρ is not as

discriminative as T_s . In cases when two sets are completely separated, ρ has value of 1 regardless of the difference between the means of the two sets.

Figure 5.3: Illustration of T_s and ρ values for different distributions of D_{intra} and D_{inter}

Figure 5.3 illustrates the values of T_s and ρ for various distributions of D_{intra} and D_{inter} . As can be seen from the Figure 5.3c and Figure 5.3d, when D_{intra} and D_{inter} are well separated, ρ might not be as effective as T_s since its values are not as discriminative. However, it has the benefit over T_s in the sense that no implicit distribution is assumed, while T_s implicitly assumes that D_{intra} and D_{inter} have normal distributions.

5.1.2 Application of the proposed criteria in speaker diarization system

The previous section introduced the two novel criteria for a general clustering problem. This section describes the procedures to incorporate these criteria into a speaker diarization system with IAHC framework to determine the number of clusters. Using the same notation for IAHC as in section 3.3.1.4, let:

- K_i the number of clusters at iteration i
- $C_k^{(i)}$ the k th cluster at iteration i , with $k = 1, \dots, K_i$
- $X_k^{(i)}$ the set of feature vectors assigned to cluster $C_k^{(i)}$ at iteration i

It is noted that, at the beginning, when the iterative clustering module starts, $i = 0$ and K_0 initial clusters are given (which are initialized by the clusters initialization module described in Section 3.3.1.3).

Procedures This section reiterates the summary of IAHC procedure with focus on how the proposed criteria could be used to determine the number of clusters. Please refer to section 3.3.1.4 for comprehensive details about the IAHC.

1. For each cluster $C_k^{(i)}$, a *Gaussian mixture model* (GMM) $\Theta_k^{(i)}$ is trained using all the feature vectors $X_k^{(i)}$ with $k = 1, \dots, K_i$.
2. Construct a *hidden Markov model* (HMM) $\Lambda^{(i)}$ with duration constraint for each cluster $C_k^{(i)}$. Each cluster $C_k^{(i)}$ is modeled by a HMM $\Lambda_k^{(i)}$ with S sub-states, where S determines the minimum duration of each speaker segment. Each sub-state shares the same GMM $\Theta_k^{(i)}$ estimated from data of their respective cluster $C_k^{(i)}$ in the previous step.
3. Perform *Viterbi decoding* [83] on the feature stream to obtain the new cluster assignment $X_k^{(i+1)}$.
4. Repeat step (1) to (3) until the cluster assignment is stabilized.

5. Assuming that it is now at iteration j , each cluster $C_k^{(j)}$ is then split into segments of L seconds. In total, there are N_l segments $\{s_1, s_2, \dots, s_{N_l}\}$ for the whole audio recording. The distance $d(s_m, s_n)$ between any pair of segments s_m and s_n can be computed with the selected distance metric. Follow equation 5.1 and equation 5.2:

$$D_{intra} = \{d(s_m, s_n) | \forall m, n \exists k : s_m \in C_k^{(i)}, s_n \in C_k^{(i)}\} \quad (5.7)$$

$$D_{inter} = \{d(s_m, s_n) | \forall m, n \exists k \neq l : s_m \in C_k^{(i)}, s_n \in C_l^{(i)}\} \quad (5.8)$$

6. Once D_{intra} and D_{inter} are specified, the value of T_s and ρ could be evaluated with equation 5.3 and equation 5.6 respectively. These values are then recorded together with the cluster partitioning.
7. The cluster pair with minimum distance is identified:

$$d_{min} = \min_{1 \leq m < n \leq K_j} d(s_m, s_n) \quad (5.9)$$

$$(m_{min}, n_{min}) = \arg \min_{1 \leq m < n \leq K_j} d(s_m, s_n) \quad (5.10)$$

8. Repeat the whole procedures from step (1) to step (7), until there are only two clusters. Since the value of T_s and ρ are noted at each iteration in step (6), the cluster partitioning corresponding to the maximum value of T_s or ρ is finally selected. Figure 5.4 illustrates the example procedures to determine whether the partition with 2 clusters or the partition with 3 clusters is better with ρ criterion.

5.2 Spectral subspace

In clustering techniques, selecting the appropriate distance measure is certainly a crucial decision. Nonetheless, it is not the only way for improving the performance of clustering. By exploiting the relationship of data points to their neighbors, spectral subspace approach [92, 184] has been shown to be capable of handling difficult data set with different structures and sparsity. The main principle is to construct an *affinity matrix*, a symmetrical square matrix of distances between all pairs of clusters, and to project this matrix to the subspace of lower dimension. In this subspace, the data points generally form tight groups and the groups are well separated. This section describes one of the most popular

 Figure 5.4: Determine number of clusters with ρ criterion

implementation based on the paper by Ng et al [92] on spectral clustering. The idea is, however, not to perform clustering but to learn the procedures to project to the spectral subspace. Later, these procedures are then applied to study the speaker discriminability and the newly proposed clustering criteria in the spectral subspace, comparing to the original space.

5.2.1 Spectral clustering algorithm

Problem formulation To group N data points $\{x_1, x_2, \dots, x_N\}$ into K disjoint clusters $\{C_1, C_2, \dots, C_K\}$, knowing that the distance between any two points x_i and x_j is $d(x_i, x_j)$.

Spectral clustering procedures

1. Form the affinity matrix $\mathbf{A} \in \mathcal{R}^{N \times N}$ defined by

$$\mathbf{A}_{ij} = \exp\left(\frac{-d(x_i, x_j)^2}{\sigma_i \sigma_j}\right) \quad (5.11)$$

where $\sigma_i = d(x_i, x_M)$ is the distance from point x_i to its M^{th} nearest neighbor.

2. Define Q to be a diagonal matrix with

$$Q(i, i) = \sum_{j=1}^N A_{ij} \quad (5.12)$$

and construct the normalized affinity matrix

$$L = Q^{-1/2} A Q^{-1/2} \quad (5.13)$$

3. Find $v_1, v_2, \dots, v_K$, the K largest eigenvectors of L , largest eigenvectors are those vectors corresponding to the largest eigenvalues. Form the matrix $V = [v_1 v_2 \dots v_K] \in \mathcal{R}^{N \times K}$
4. Re-normalize V to obtain matrix Y such that

$$Y_{ij} = \frac{V_{ij}}{\sqrt{\sum_{j=1}^N V_{ij}^2}} \quad (5.14)$$

5. Treat each row of Y as a point in $\mathcal{R}^K$ and cluster via k -means.

5.2.2 Spectral clustering algorithm with proposed metrics

In the original spectral clustering technique, the k -mean algorithm is typically employed for clustering once the data points are projected to the affinity subspace. In this section, the procedures to integrate the proposed criteria in the spectral clustering framework are given:

5.2.2.1 Algorithm

1. Step (1) to step (4) are similar to the original spectral clustering in section 5.2.1.
2. Instead of performing k -mean clustering, let $Z \in \mathcal{R}^{N \times N}$ be the matrix in which each element of Z is the cosine distance between row i and row j of Y

$$Z(i, j) = \sum_{l=1}^N Y_{il} Y_{jl} \quad (5.15)$$

3. Define D_{intra} and D_{inter} following equation 5.1 and equation 5.2 respectively, as the set of intra-cluster distances and inter-cluster distances

$$D_{intra} = \{Z(i, j) | \forall i, j \exists k : x_i \in C_k, x_j \in C_k\} \quad (5.16)$$

$$D_{inter} = \{Z(i, j) | \forall i, j \exists k \neq l : x_i \in C_k, x_j \in C_l\} \quad (5.17)$$

Then, the values T_s and ρ could be evaluated.

5.2.2.2 Analysis

This section will provide an analytical analysis of the proposed algorithm in the ideal case where $A_{ij} = 0$ if x_i and x_j are in different clusters and $A_{ij} > 0$ otherwise. In other word, the points in different clusters are assumed to be infinitely far apart. Without loss of generality, it is supposed that there are $K = 3$ actual clusters for ease of discussion.

The matrix A and matrix L are block-diagonal matrices:

$$L = \begin{pmatrix} L^{(1)} & 0 & 0 \\ 0 & L^{(2)} & 0 \\ 0 & 0 & L^{(3)} \end{pmatrix} \quad (5.18)$$

where $L^{(k)}$ is the sub-matrix corresponding to cluster C_k . There are three possible scenarios:

1. *The actual number of clusters is known*

In this case

$$V = \begin{pmatrix} v^{(1)} & 0 & 0 \\ 0 & v^{(2)} & 0 \\ 0 & 0 & v^{(3)} \end{pmatrix}$$

where $v^{(k)}$ is an eigenvector of the sub-matrix $L^{(k)}$. When V is normalized such that each row of V has unit length, we obtain:

$$Y = \begin{pmatrix} \vec{1} & 0 & 0 \\ 0 & \vec{1} & 0 \\ 0 & 0 & \vec{1} \end{pmatrix}$$

When each data point x_i is assigned to its correct cluster: D_{intra} comprises of all 1s and D_{inter} comprises of all 0s. Thus, $T_s = +\infty$ and $\rho = 1$. In all other cases where there is at least one data point in a wrong cluster, D_{intra} and D_{inter} both comprises of some values equal to 0 and some values equal to 1 which will lead to $T_s < +\infty$ and $\rho < 1$.

2. *The estimated number of clusters is less than the actual number of clusters*

Supposed $K' = 2$ largest eigenvectors are selected

$$Y = \begin{pmatrix} \vec{1} & 0 \\ 0 & \vec{1} \\ 0 & 0 \end{pmatrix}$$

In this case, D_{intra} will definitely comprise of some values equal to 0 and some values equal to 1, regardless of how we group the data into K' clusters. Hence, $T_s < +\infty$ and $\rho < 1$.

3. *The estimated number of clusters is more than the actual number of clusters*

Supposed $K'' = 4$ largest eigenvectors are selected

$$Y = \begin{pmatrix} \vec{a} & 0 & 0 & \vec{b} \\ 0 & \vec{1} & 0 & 0 \\ 0 & 0 & \vec{1} & 0 \end{pmatrix}$$

where $\vec{a}, \vec{b}$ are the two eigenvectors from the same sub-matrix, $\vec{a} \cdot \vec{b} = 0$ and at least one of them is strictly positive [92]. Regardless of how we group the data into K'' clusters, D_{intra} has some values equal to 1 and some values less than 1, D_{inter} has some values equal to 0 and some values greater than 0. Thus, $T_s < +\infty$, however, the value of ρ is not conclusive although in practice it works well as shown later in the experiments.

5.3 Automatic stream weights estimation

Considering that multiple feature streams are accessible, it is often beneficial to utilize more than one feature stream as they may contain complementary information. In the scope of this thesis, as it has been shown in section 3.3.2.3, multichannel speaker diarization systems employ weighting scheme at the log-likelihood or the distance to reflect the contribution and effectiveness of each stream. More useful streams are given larger weights and vice versa. The complication arises when each stream carries different information, whether and how these streams are useful for the task at hand need to be evaluated. The straightforward approach is to try these features on development data

set and the obtained results could be used to calibrate the contribution of each feature stream. An obvious problem with this simple strategy is the generalization capability on the unseen data, one fixed set of calibrated weights may not work well for all conditions and different sets may be required for different scenarios. This complicates the issue, and furthermore one may add additional feature streams or remove the existing ones and the whole system needs to be re-calibrated. On the other hand, adaptive or dynamic weighting maybe more robust towards unseen scenarios. In the context of IAHC, the dynamic weighting might be even more helpful since the weights could be adaptively changed from iteration to iteration. This section presents a novel scheme for adaptive weighting based on Fisher criterion.

Let:

- K be the number of clusters.
- C_k be the k th cluster with $k = 1, \dots, K$
- $\mathbf{w}_l$ be the unknown weight of features stream l .

Algorithm Considering K clusters $\{C_1, \dots, C_K\}$, they are further divided into segments of T seconds to obtain the segments $\{S_1, \dots, S_M\}$, with M being the total number of segments.

1. For each segment S_m , and for each features stream l , build a GMM Θ_m^l , with $m = 1, \dots, M$ and $l = 1, \dots, L$.
2. For each pair of segments S_m and S_n , and for each features stream l , build a GMM Θ_{mn}^l , with $1 \leq m < n \leq M$ and $l = 1, \dots, L$.
3. Given a pair of frames i and j , let $\mathbf{x}_i^l$ and $\mathbf{x}_j^l$ be the feature vector of the features stream l at frame i and j respectively. If i and j belong to segment S_m and S_n , $m \neq n$, correspondingly then compute

$$s_i^l = \log \mathcal{L}(\mathbf{x}_i^l | \Theta_m^l) - \log \mathcal{L}(\mathbf{x}_i^l | \Theta_{mn}^l) \quad (5.19)$$

$$s_j^l = \log \mathcal{L}(\mathbf{x}_j^l | \Theta_n^l) - \log \mathcal{L}(\mathbf{x}_j^l | \Theta_{mn}^l) \quad (5.20)$$

where $\mathcal{L}(\mathbf{x}|\Theta)$ denoting the likelihood of $\mathbf{x}$ given the model Θ . Noted that s_i^l and s_j^l are scalar values. Stacking L features stream together to form a L -dimensional vector

$$s_i = \begin{bmatrix} s_i^1 \\ s_i^2 \\ \vdots \\ s_i^L \end{bmatrix} \quad (5.21)$$

$$s_j = \begin{bmatrix} s_j^1 \\ s_j^2 \\ \vdots \\ s_j^L \end{bmatrix} \quad (5.22)$$

If S_m and S_n belong to the same cluster then s_i and s_j are assigned to the set s_{intra} otherwise, they are assigned to the set s_{inter} .

4. Suppose that s_{intra} and s_{inter} have means $\boldsymbol{\mu}_{intra}$, $\boldsymbol{\mu}_{inter}$ and covariances $\boldsymbol{\Sigma}_{intra}$, $\boldsymbol{\Sigma}_{inter}$. The linear combination of scores with weights $\mathbf{w}$ will have means $\mathbf{w} \cdot \boldsymbol{\mu}_{intra}$, $\mathbf{w} \cdot \boldsymbol{\mu}_{inter}$ and variances $\mathbf{w}^T \boldsymbol{\Sigma}_{intra} \mathbf{w}$, $\mathbf{w}^T \boldsymbol{\Sigma}_{inter} \mathbf{w}$ respectively. The separation between these two distributions is defined to be the ratio of the variance between the classes to the variance within the classes according to Fisher [185]

$$S = \frac{\sigma_{between}^2}{\sigma_{within}^2} \quad (5.23)$$

$$= \frac{(\mathbf{w} \cdot \boldsymbol{\mu}_{intra} - \mathbf{w} \cdot \boldsymbol{\mu}_{inter})^2}{\mathbf{w}^T \boldsymbol{\Sigma}_{intra} \mathbf{w} + \mathbf{w}^T \boldsymbol{\Sigma}_{inter} \mathbf{w}} \quad (5.24)$$

$$= \frac{(\mathbf{w} \cdot (\boldsymbol{\mu}_{intra} - \boldsymbol{\mu}_{inter}))^2}{\mathbf{w}^T (\boldsymbol{\Sigma}_{intra} + \boldsymbol{\Sigma}_{inter}) \mathbf{w}} \quad (5.25)$$

It is shown that the maximum separation occurs when

$$\mathbf{w} \propto (\boldsymbol{\Sigma}_{intra} + \boldsymbol{\Sigma}_{inter})^{-1} (\boldsymbol{\mu}_{inter} - \boldsymbol{\mu}_{intra}) \quad (5.26)$$

By constraining $\sum_{l=1}^L w_l = 1$, equation 5.26 leads to a unique solution $\mathbf{w}_0$, which are also the features stream weights.

Figure 5.5 illustrates the features stream weights for different distributions of s_{intra} and s_{inter} for the case of two features streams.

Figure 5.5: Illustration of features stream weights for different distributions of s_{intra} and s_{inter} . The red line indicates the projection hyperplane to achieve maximum separation

5.4 Experiments

5.4.1 Determine number of clusters in speaker diarization systems

Objective As depicted in Figure 3.4, the error rates of a speaker diarization system may fluctuate in a wide range depending on how accurate the stopping points are. The optimal stopping points, however, are data related and are different for each data set or even different for each meeting in the same data set. This experiment studies the performance of estimating the number of clusters using threshold method and the proposed technique. For the threshold method, the considered distance metrics include GLR, BIC, NCLR, ICR and d_s .

Experiments The experiments using baseline SDM speaker diarization systems with different distance metrics were conducted and the performances on NIST RT meetings data were recorded for a wide range of possible stopping thresholds. Figure 5.6, Figure 5.7, Figure 5.8 and Figure 5.9 show the DER of the baseline systems using BIC, NCLR,

ICR and d_s metric correspondingly. Table 5.1 and Table 5.2 present the results of the improved speaker diarization system using d_s as distance metric with T_s and ρ as stopping criteria respectively.

Figure 5.6: Speaker diarization performance of SDM system with various BIC stopping thresholds.

Table 5.1: DER (%) of the SDM speaker diarization system with d_s as distance metric and T_s as stopping criterion

	RT 2005	RT 2006	RT 2007	RT 2009
d_s/T_s	13.06	19.47	21.62	18.91

Table 5.2: DER (%) of the SDM speaker diarization system with d_s as distance metric and ρ as stopping criterion

	RT 2005	RT 2006	RT 2007	RT 2009
d_s/ρ	14.62	22.94	20.64	23.16

Figure 5.7: Speaker diarization performance of SDM system with various NCLR stopping thresholds.

Figure 5.8: Speaker diarization performance of SDM system with various ICR stopping thresholds.

Figure 5.9: *Speaker diarization performance of SDM system with various d_s stopping thresholds.*

Discussion Figure 3.4, 5.6, 5.7 and 5.9 show the performance of the baseline SDM speaker diarization systems with various distance metrics: GLR, BIC, NCLR and d_s respectively. It is observed from these figures that for the investigating metrics, there is no such universal threshold which works equally well across all the databases. To illustrate this inconsistency, let's examine how the performance would vary when different development dataset was used. Table 5.3 shows the DER of the system when the stopping threshold was tuned on RT2006 dataset, while Table 5.4 shows the DER when the threshold was tuned on RT2007 dataset.

The tables clearly show that there are large variances in both the threshold values and the DERs when different development sets are used. This problem, in practice, will make the task of selecting the proper parameters for the system to be more challenging, especially when there is no prior knowledge about the data to which the system is being applied. On the other hand, the proposed stopping criteria have been shown to outperform the conventional threshold approach, especially for certain databases (RT2007 and RT2009), the improvement is up to 20% relatively. T_s criterion seems to perform better

Table 5.3: *DER (%) of the baseline SDM speaker diarization system with various clustering distance measures. The stopping threshold was tuned on RT2006 database. The value in the parenthesis is the optimal threshold obtained from the development data.*

	RT 2005	RT 2006	RT 2007	RT 2009
GLR (2500)	16.99	17.58	28.25	37.24
BIC (2.25)	17.59	13.53	27.42	38.44
NCLR (0.85)	16.96	15.97	27.82	43.09
ICR (0.085)	18.52	18.20	28.62	39.87
d_s (0.41)	17.62	18.43	24.14	39.04

Table 5.4: *DER (%) of the baseline SDM speaker diarization system with various clustering distance measures. The stopping threshold was tuned on RT2007 database. The value in the parenthesis is the optimal threshold obtained from the development data.*

	RT 2005	RT 2006	RT 2007	RT 2009
GLR (4000)	17.12	23.33	13.97	29.26
BIC (3.75)	17.12	18.15	13.58	29.65
NCLR (1.25)	11.32	18.58	16.86	31.50
ICR (0.135)	14.21	31.69	19.54	30.65
d_s (0.49)	14.58	22.90	19.64	31.29

than ρ criterion, particularly on RT2009 dataset. Both T_s and ρ criteria do not required any threshold or development data, the clustering partition which corresponding to the maximum value of T_s or ρ should be picked.

And how well are the proposed techniques comparing to the ideal stopping points? This information would allow us to gauge how far are we from the theoretical limit and whether there is extent for improvement. Table 5.5 presents the performance of baseline SDM speaker diarization system using d_s as distance metric and oracle stopping criterion. The oracle stopping method relies on the reference labels to analyze all possible stopping points and select the one with minimum DER for the investigating database.

Table 5.5: *DER (%) of the baseline SDM speaker diarization system with d_s as distance metric and oracle stopping criterion*

	RT 2005	RT 2006	RT 2007	RT 2009
d_s	8.06	14.47	10.62	15.91

Comparing these results with the one in Table 5.1 or Table 5.2, there are obviously significant gaps which leave much potential for advancement in the future works.

5.4.2 Distance metrics in spectral subspace

Objective In section 4.3.1, several experiments were conducted to study the speaker discriminative capability of various distance metrics. This section extends that study to the spectral subspace as it is generally believed that better separation among clusters could be achieved. It is hypothesized that when projected to the spectral subspace, improvements in terms of *equal error rate* (EER) could be observed across all distance metrics since the speakers are more separated.

Experimental setup With the same setup as in section 4.3.1, the distances between every pair of speaker segments were calculated using these distance metrics: GLR, BIC, NCLR, ICR, and d_s . In this set of experiments, all the distances were evaluated on the mixed set $\{S_5 \cup S_{10} \cup S_{15}\}$. Given the distances, they were then used to form the affinity matrices following the equation 5.11

$$\begin{aligned} \mathbf{A}_{ij}^{GLR} &= \exp\left(\frac{-d_{GLR}(i, j)^2}{\sigma_i \sigma_j}\right) \\ \mathbf{A}_{ij}^{BIC} &= \exp\left(\frac{-d_{BIC}(i, j)^2}{\sigma_i \sigma_j}\right) \\ \mathbf{A}_{ij}^{NCLR} &= \exp\left(\frac{-d_{NCLR}(i, j)^2}{\sigma_i \sigma_j}\right) \\ \mathbf{A}_{ij}^{ICR} &= \exp\left(\frac{-d_{ICR}(i, j)^2}{\sigma_i \sigma_j}\right) \\ \mathbf{A}_{ij}^s &= \exp\left(\frac{-d_s(i, j)^2}{\sigma_i \sigma_j}\right) \end{aligned}$$

where $i = 1, \dots, N$ and $j = 1, \dots, N$ denote the i th and j th segment respectively, N denotes the total number of segments, σ_i is the distance between segment i to its M^{th} nearest neighbor. These affinity matrices were subsequently projected to the spectral subspace following the procedures described in section 5.2 to obtain the matrices: $\mathbf{Y}^{GLR}$, $\mathbf{Y}^{BIC}$, $\mathbf{Y}^{NCLR}$, $\mathbf{Y}^{ICR}$, and $\mathbf{Y}^s$ as shown in equation 5.14 correspondingly. Treating each row of these matrices as a new data point and Euclidean metric were employed to measure the similarity between all pairs of points (each point in the subspace is a one-to-one correspondent of a segment in the original space).

Number of neighbors In this experiment, the parameter K which is the number of eigenvalues to retain in spectral clustering was set to the actual number of speakers in the data set. Figure 5.10 summarizes the effect of choosing the value of M neighbors.

Figure 5.10: $EER(\%)$ of different distance metrics when projected to the spectral subspace with M being the number of neighbors used in spectral clustering algorithm

Number of eigenvalues This experiment was set up to analyze the performance of spectral clustering in case the number of clusters are not known a priori and must be estimated from the data. Ideally, parameter K should be set to the actual number of clusters since it is often the practice to assume that large eigenvalues corresponding to signal component and small eigenvalues are corresponding to noise components. Nonetheless, the estimated $\hat{K}$ might be different than the actual number of clusters and the effects of selecting different values of $\hat{K}$ are reported in Figure 5.11 with the number of neighbors M fixed to 50.

Figure 5.11: $EER(\%)$ of different distance metrics when projected to the spectral subspace with varying number of eigenvalues

5.4.3 Determine number of clusters in spectral subspace

Experiments This set of experiments was conducted to study the performance of the proposed technique to determine number of clusters when the distances are projected to the spectral subspace. The procedures for projection are described in section 5.2.2. Table 5.6 and Table 5.7 show the DER for two systems using T_s and ρ metric correspondingly.

Table 5.6: $DER(\%)$ of the SDM speaker diarization system with d_s as distance metric and T_s as stopping criterion in spectral subspace

	RT 2005	RT 2006	RT 2007	RT 2009
d_s	12.56	19.91	16.67	17.05

Discussion Comparing the results in Table 5.6, Table 5.7 with those in Table 5.1 and Table 5.2, it is noted that projecting to the spectral subspace does improve the system performance, however the slight improvement which is not consistent across all

Table 5.7: *DER (%) of the SDM speaker diarization system with d_s as distance metric and ρ as stopping criterion in spectral subspace*

	RT 2005	RT 2006	RT 2007	RT 2009
d_s	13.35	21.56	18.24	21.36

databases, might not justify the complexity and extra steps required to implement this algorithm. The probable explanation regarding the shortcoming of this technique is due to the fact that the subspace projection is sensitive to the number of chosen eigenvalues as depicted in Figure 5.11. How should this value be selected since the number of clusters K is not known and need to be estimated using this algorithm? This problem becomes the “chicken and egg” issue as K needs to be known for the subspace projection to be effective. Thus, to mitigate this issue, in the proposed technique, the hypothesis $\hat{K}$ is used, which is the current number of clusters and which may vary according to the clustering iterations. This presents the new challenge as the projections are to different subspaces with different dimensionality and the values of T_s or ρ need to be compared over all these subspaces, which may explain why this algorithm might not be so effective.

5.4.4 Adaptive stream weights in multichannel speaker diarization system

Objective In MDM speaker diarization system, the acoustic feature stream and delay feature stream are both utilized to improve the performance. However, for different meeting room, the usefulness of each feature stream may not be the same e.g. when the participants move or when the microphones are too close, the delay features should contribute less since they are not reliable. It is therefore hypothesized that adaptive weighting should perform better than fixed weighting as the algorithm can adapt to new environment and adjust the contribution correspondingly. This set of experiments was taken to test the hypothesis.

Experiments Two MDM systems: one with fixed weighting and one with adaptive weighting were evaluated on Rich Transcription data sets. For fixed weighting system, the weights were calibrated to minimize the error rates on RT 2005 and RT 2006 data sets. The resulting weights are 0.9 for acoustic stream and 0.1 for delay stream.

	RT 2005	RT 2006	RT 2007	RT 2009
Fixed weighting	8.83	15.55	10.68	13.97
Adaptive weighting	7.47	12.59	9.37	12.19

Table 5.8: *DER (%) of the MDM speaker diarization system with d_s metric and proposed T_s criterion in spectral subspace*

5.5 Summary

The chapter first proposed two novel criteria for measuring cluster partitioning quality. The technique to determine number of clusters in an IAHC framework is then described, by using the proposed criteria to select the partition with highest quality. In contrast to traditional methods, the proposed criteria are not monotonically increasing or decreasing, instead they converge to the local maximum at the correct number of clusters. These criteria are then investigated in the spectral subspace, in which the performance of the proposed technique is further improved since the clusters are more separated. The obtained results show that the presented approach perform better than conventional methods and no prior threshold is required. Finally, in the last section the issue of multiple stream fusion is addressed, particularly the issue of weighting for each feature stream. A technique is proposed to select the weights such that the fusion scores between two classes: within clusters scores and between cluster scores are maximized based on Fisher criterion. The weighing scheme is then applied to the MDM speaker diarization system and its performance is shown to be better than fixed weighting scheme.

Chapter 6

Conclusions and Future Work

This thesis documents the development of robust speaker diarization systems, where the objective is to annotate a continuous audio recording with appropriate speaker identities corresponding to the region where they spoke, with special focus on meetings domain. Robustness is a key issue for diarization in meetings since it is expected that the systems should perform well across different meeting rooms, with different microphone types and characteristics, without any knowledge about the number of participants or their identities. Just as important, the systems should work in the condition of low *signal to noise ratio* (SNR) as the signals are typically captured by distant microphones. Many state-of-the-art systems perform unsatisfactorily in such challenging scenarios, particularly those relying on *thresholds* calibrated from development data. This thesis aims to develop robust speaker diarization techniques to address these issues. The major contributions of this thesis can be broken down into three parts, related to different stages of speaker diarization systems, namely (1) *distance metric*, (2) *stopping criteria*, and (3) *adaptive feature streams weighting*. While the proposed techniques are module specific, they follow the same underlying concept throughout the thesis: *robust, threshold-free* and *unsupervised*.

To conclude the work, this chapter begins with the summary of the thesis contributions in section 6.1, followed by the possible directions for further exploration in section 6.2.

6.1 Contributions

Iterative agglomerative hierarchical clustering (IAHC) is the most widely used framework in state-of-the-art speaker diarization systems. This clustering architecture has formed the basis for the contributions in this thesis, since the proposed techniques are designed with intentions to improve the robustness of IAHC speaker diarization systems. The following sections summarize the proposed techniques and their contributions, which are focusing on three different aspects of the IAHC framework.

6.1.1 Distance metric

IAHC framework requires a distance metric to be specified in order to measure the similarities between clusters, which are then used to decide the cluster pair to be merged at each clustering iteration. A distance metric is also a requisite in many speaker diarization systems where threshold is employed as stopping criterion. Consequently, adopting an appropriate distance metric is critical to the success of an IAHC speaker diarization system. In many state-of-the-art diarization systems, GLR-based metrics such as GLR or BIC are adopted frequently, however very few published works ever analyzed systematically the performances and characteristics of these metrics. Recently, Han et al. [14] showed that GLR-based metrics are not robust to data size variation since these distances grow proportionally to the size of the clusters. The authors have given the mathematical proof in a special case when the distributions of feature vectors are Gaussians. In this thesis, the work in chapter 4, which was published in *Speech and Audio Processing for Coding, Enhancement and Recognition* [B], extends the result in [14] by presenting the mathematical proof in general cases without the assumption of Gaussian distributions. As a result of this analysis, it is suggested that normalization techniques should be applied to GLR-based metrics, such as the *information change rate* (ICR) [14] metric. Further investigation into the conventional distance metrics, particularly the likelihood-based techniques mentioned in section 2.4.2, it is observed that only the sum or the mean, the first-order statistics, of the likelihood scores are exploited. These observations prompted for the study of using higher-order statistics in distance metrics. In such study, section 4.2 has shown that the variance, the second-order statistics, of the

likelihood scores is also a useful indicator for speaker similarities. The finding leads to the proposal of the novel distance metric d_s , which incorporates both first and second order statistics and is robust to cluster size variation.

Experiments on various metrics were first conducted on the related speaker verification task to validate the speaker discriminative capability of individual metric. For each metric, the distances among different speaker segments of various lengths were calculated, and the performances in terms of *equal error rate* (EER) and *detection error trade off* (DET) curve were evaluated. The obtained results clearly show the undesired effect of cluster sizes on GLR and BIC: these distances increase as cluster sizes increase. On the other hand, the proposed measure d_s and ICR distances are converging steadily when the cluster sizes are sufficiently large. In overall, d_s performs consistently well, and is substantially better than other metrics when cluster of different sizes are considered.

Experiments were also carried out to assess the performance of various speaker diarization systems using the above mentioned distance metrics for measuring cluster similarities. It is concluded that there are certain correlations between the performance of individual distance metric in terms of EER and the performance of the whole diarization systems in terms of *diarization error rate* (DER). The proposed metric d_s delivered the best performance among the considered metrics.

6.1.2 Stopping criteria

Estimating the number of speakers in an audio recording or determining the number of clusters in an unsupervised clustering problem in general and in diarization task in particular is challenging. Most speaker diarization systems employ a bottom-up hierarchical clustering framework, as a result, determining number of clusters is typically done as detecting the stopping point where the system makes decision of when to stop the merging process. Conventional approaches are either using a threshold learnt from an external development data or applying a model selection technique such as BIC. However, these techniques are shown in section 5.4.1 to be data related and not robust for a variety of tasks. Motivated by these limitations, two criteria are proposed in this work, which are able to estimate the number of clusters in IAHC framework without any predefined thresholds. This technique was published in INTERSPEECH [E]. The technique works

by employing the proposed criteria at each iteration of the clustering steps to measure the cluster partitioning qualities. At the end, the partition with the maximum quality was select, and as a result the number of clusters was deduced. The novelty idea is that the quality given by the proposed criteria is not monotonically increasing or decreasing as in the case of conventional distance metric. The estimated partitioning quality tends to peak at the correct number of cluster and therefore no threshold is needed. The technique was demonstrated to work well on Rich Transcription data sets with competitive results and added benefit of no tuning thresholds.

As spectral subspace methods are well-known for their performance on difficult clustering problem, this study then examined the spectral subspace and it is shown that the speakers are more separated in this subspace (section 5.4.2). The results prompted for an investigation of the proposed criteria in the spectral subspace as it is hypothesized the detection of number of clusters should be better as the clusters form tight groups and are well separated. The experimental results confirmed the hypothesis, as it is shown that the partitioning quality computed using the proposed criteria exhibits more dominant peaks at the correct number of cluster and the detection performance is thus improved. In overall, experiments on meetings data showed that the proposed technique performs comparable or better than conventional methods without relying on any thresholds. This work has been published in IEEE ICASSP 2009 [C].

6.1.3 Adaptive feature stream weighting

The last study considered the technique to improve the performance of multichannel speaker diarization system by analyzing the effectiveness of different feature stream. The problem here is how to combine multiple feature streams effectively as each feature has its own characteristics and whether it is useful towards the task at hand. In the context of multichannel meetings, two streams are considered: the short term spectral acoustic features and the delay features. For each meeting room, the merits of these two stream may vary as the environment changes, the microphone configuration changes, and the participants are different. Moreover, conventional methods either estimate the merit of each stream on development data to derive a fixed set of weights or estimate these weights without making use of the clusters information (for e.g. the inverse entropy technique).

This work proposed an algorithm to estimate the weights of various streams based on Fisher Linear Discriminant Analysis (LDA) or Fisher criterion. The underlying idea is to select the weights such that the ratio of variance of between clusters distances to variance of within clusters distances is maximized. The novelty of this approach includes: the selected weights are to produce more separated clusters, and the weights are adaptive at each iteration of the clustering process with no tuning data required. Experiments were conducted on NIST RT meetings data sets and the approach was shown to perform better than fixed scheme weighting in overall. This technique was presented in the NIST Rich Transcription 2009 Evaluation workshop [D] and was submitted to IEEE Signal Processing Letters [A].

6.2 Future Work

Given multiple feature streams, it is also noteworthy to examine more effective frameworks to combine these individual feature streams. With the recent development in machine learning community, it might be possible to consider multiple feature streams in a multi-view setting. In this setting, the feature streams are considered as different views, each of which is sufficient to learn the target concept. It is then to explore multi-view co-training or co-expectation maximization techniques to learn speaker models in speaker diarization task.

One topic within the meeting domain processing that has received quite some attention recently is speaker overlap detection. It refers to the detection of the segments where more than one speaker is talking at the same time, and the output of an appropriate identity for each participant. Overlapped speech, which is failing to be identified by state-of-the-art diarization systems, produces missed speech errors which can constitute a significant portion of the error of these systems. In addition, this speech can negatively affect speaker modeling in the clustering process, increasing speaker error as a result.

In current state-of-the-art diarization systems, it is assumed that no prior information about the number of speakers is given. Some systems only use this information to determine the number of clusters. It would be interesting to explore particular areas of

application where the number of speakers in a meeting is known. This particular information may change the way the speaker diarization algorithms are designed and some techniques in speaker identification may be applicable.

Finally, one of our objectives is to increase the robustness of the diarization system to accommodate diverse conditions with less tunable parameters. There is still more that can be done in this topic to eliminate as many tuning parameters as possible, letting the algorithms automatically select such parameters solely from the test data.

Publications

- (A) T. H. Nguyen, E. S. Chng, H. Li: “Robust techniques for speaker diarization system”, to be submitted to IEEE Signal Processing Letters.
- (B) T. H. Nguyen, E. S. Chng, H. Li, “Speaker Diarization: An Emerging Research”, in *Speech and Audio Processing for Coding, Enhancement and Recognition*, pp. 229-277, Springer, 2014.
- (C) T. H. Nguyen, H. Li, and E. S. Chng, “Cluster criterion functions in spectral subspace and their application in speaker clustering,” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009 . IEEE International Conference on*, pp. 4085-4088, 2009.
- (D) T. H. Nguyen, H. Sun, S. K. Zhao, Z. K. K. Swe, H. D. Tran, T. L. Nwe, B. Ma, E. S. Chng, and H. Li, “The IIR-NTU speaker diarization systems for RT 2009,” in *RT’09, NIST Rich Transcription Workshop*, Melbourne, Florida, 2009.
- (E) T. H. Nguyen, E. S. Chng, and H. Li, “T-test distance and clustering criterion for speaker diarization.,” in *INTERSPEECH 2008*, pp. 36-39, 2008.
- (F) E. C. W. Koh, H. Sun, T. L. Nwe, T. H. Nguyen, B. Ma, E. S. Chng, H. Li, and S. Rahardja, “Speaker diarization using direction of arrival estimate and acoustic feature information: The I2R-NTU submission for the NIST RT 2007 evaluation,” in *Multimodal Technologies for Perception of Humans* , pp. 484-496, Springer, 2008.
- (G) E. C. W. Koh, H. Sun, T. L. Nwe, T. H. Nguyen, B. Ma, E. S. Chng, H. Li, and S. Rahardja, “Using direction of arrival estimate and acoustic feature information in speaker diarization”, in *INTERSPEECH 2007*, pp. 2149-2152, 2007.

References

- [1] B. Zhou and J. H. L. Hansen, “Efficient audio stream segmentation via the combined T^2 statistic and Bayesian information criterion,” *Speech and Audio Processing, IEEE Transactions on*, vol. 13, pp. 467–474, July 2005.
- [2] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, “Step-by-step and integrated approaches in broadcast news speaker diarization,” *Computer Speech & Language*, vol. 20, no. 23, pp. 303 – 330, 2006.
- [3] A. Stolcke, G. Friedland, and D. Imseng, “Leveraging speaker diarization for meeting recognition from distant microphones,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4390–4393, IEEE, 2010.
- [4] G. Friedland, L. Gottlieb, and A. Janin, “Joke-o-mat: browsing sitcoms punchline by punchline,” in *Proceedings of the 17th ACM international conference on Multimedia*, pp. 1115–1116, ACM, 2009.
- [5] R. Mertens, P.-S. Huang, L. Gottlieb, G. Friedland, and A. Divakaran, “On the applicability of speaker diarization to audio concept detection for multimedia retrieval,” in *Multimedia (ISM), 2011 IEEE International Symposium on*, pp. 446–451, IEEE, 2011.
- [6] M. J. Gales, P. Woodland, H. Y. Chan, D. Mrva, R. Sinha, S. E. Tranter, *et al.*, “Progress in the cu-htk broadcast news transcription system,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1513–1525, 2006.

REFERENCES

- [7] G. Friedland, J. Chong, and A. Janin, “Parallelizing speaker-attributed speech recognition for meeting browsing,” in *Multimedia (ISM), 2010 IEEE International Symposium on*, pp. 121–128, IEEE, 2010.
- [8] NIST, “The 2009 (RT-09) Rich Transcription meeting recognition evaluation plan,” 2009. [Online: <http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf>; accessed 12-December-2014].
- [9] H. Sun, T. L. Nwe, B. Ma, and H. Li, “Speaker diarization for meeting room audio,” in *INTERSPEECH*, pp. 900–903, 2009.
- [10] H. Sun, B. Ma, S. Kalayar Khine, and H. Li, “Speaker diarization system for RT07 and RT09 meeting room audio,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4982–4985, 2010.
- [11] G. Friedland, A. Janin, D. Imseng, X. Anguera Miro, L. Gottlieb, M. Huijbregts, M. T. Knox, and O. Vinyals, “The ICSI RT-09 speaker diarization system,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 371–381, Feb. 2012.
- [12] S. Bozonnet, N. Evans, C. Fredouille, D. Wang, R. Troncy, *et al.*, “An integrated top-down/bottom-up approach to speaker diarization,” *INTERSPEECH 2010, September 26-30, Makuhari, Japan*, 2010.
- [13] K. J. Han and S. S. Narayanan, “Agglomerative hierarchical speaker clustering using incremental gaussian mixture cluster modeling,” in *Interspeech*, pp. 20–23, 2008.
- [14] K. Han, S. Kim, and S. Narayanan, “Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1590–1601, 2008.

REFERENCES

- [15] C. Wooters and M. Huijbregts, “The ICSI RT07s speaker diarization system,” in *Multimodal Technologies for Perception of Humans*, pp. 509–519, Springer, 2008.
- [16] J. Ajot and J. Fiscus, “RT-09 speaker diarization results,” 2009. [Online: <http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/RT09-SPKR-v3.pdf>; accessed 20-December-2013].
- [17] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, and C. Laprun, “The rich transcription 2005 spring meeting recognition evaluation,” in *Machine Learning for Multimodal Interaction*, pp. 369–389, Springer, 2006.
- [18] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, *The rich transcription 2006 spring meeting recognition evaluation*. Springer, 2006.
- [19] J. G. Fiscus, J. Ajot, and J. S. Garofolo, “The rich transcription 2007 meeting recognition evaluation,” in *Multimodal Technologies for Perception of Humans*, pp. 373–389, Springer, 2008.
- [20] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, vol. 8. MIT press, 1964.
- [21] A. G. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. S. Kajarekar, N. Morgan, and S. Sivasdas, “Qualcomm-ICSI-OGI features for ASR,” in *Interspeech*, 2002.
- [22] J. M. Pardo, R. Barra-Chicote, R. San-Segundo, R. de Crdoba, and B. Martnez-Gonzlez, “Speaker diarization features: The UPM contribution to the RT09 evaluation,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 426–435, 2012.
- [23] D. A. van Leeuwen and M. Konečný, “Progress in the AMIDA speaker diarization system for meeting data,” in *Multimodal Technologies for Perception of Humans*, pp. 475–483, Springer, 2008.
- [24] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, vol. 1. Springer, 2008.

- [25] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2011–2022, 2007.
- [26] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.
- [27] X. Anguera, “BeamformIt acoustic beamformer,” 2009. [Online]. Available <http://www.xavieranguera.com/beamformit/>. [Accessed 20-December-2013].
- [28] X. A. Miró, *Robust Speaker Diarization for meetings*. PhD thesis, Universitat Politècnica de Catalunya, Barcelona, oct 2006.
- [29] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.
- [30] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
- [31] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [32] J. Allen, “How do humans process and recognize speech?,” *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 567–577, 1994.
- [33] E. Zwicker and E. Terhardt, “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency,” *The Journal of the Acoustical Society of America*, vol. 68, p. 1523, 1980.
- [34] S. Stevens and J. Volkman, “The relation of pitch to frequency: A revised scale,” *American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.

- [35] S. Shum, N. Dehak, R. Dehak, and J. Glass, “Unsupervised methods for speaker diarization: An integrated and iterative approach,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2015–2028, 2013.
- [36] D. Vijayasenan, F. Valente, and H. Bourlard, “Multistream speaker diarization of meetings recordings beyond MFCC and TDOA features,” *Speech Communication*, vol. 54, no. 1, pp. 55–67, 2012.
- [37] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “A sticky hdp-hmm with application to speaker diarization,” *The Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.
- [38] J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” in *Automatic Speech Recognition and Understanding, 2003. ASRU’03. 2003 IEEE Workshop on*, pp. 411–416, 2003.
- [39] C. Wooters, J. Fung, B. Peskin, and X. Anguera, “Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system,” in *RT-04F Workshop*, vol. 23, 2004.
- [40] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau, “The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*, vol. 2, pp. II–89, 2003.
- [41] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, “The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation,” in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP’04). IEEE International Conference on*, vol. 1, p. I373, 2004.
- [42] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, “Modeling prosodic feature sequences for speaker recognition,” *Speech Communication*, vol. 46, no. 3, pp. 455–472, 2005.

- [43] E. El-Khoury, C. Senac, and R. Andre-Obrecht, “Speaker diarization: Towards a more robust and portable system,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. 489–492, 2007.
- [44] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, “Prosodic and other long-term features for speaker diarization,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, pp. 985–993, 2009.
- [45] D. Imseng and G. Friedland, “An adaptive initialization method for speaker diarization based on prosodic features,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4946–4949, 2010.
- [46] J. M. Pardo, X. Anguera, and C. Wooters, “Speaker diarization for multi-microphone meetings using only between-channel differences,” in *Machine Learning for Multimodal Interaction*, pp. 257–264, Springer, 2006.
- [47] J. M. Pardo, X. Anguera, and C. Wooters, “Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time differences,” in *Interspeech*, 2006.
- [48] P. R. Roth, “Effective measurements using digital signal analysis,” *Spectrum, IEEE*, vol. 8, no. 4, pp. 62–70, 1971.
- [49] G. C. Carter, A. H. Nuttall, and P. G. Cable, “The smoothed coherence transform,” *Proceedings of the IEEE*, vol. 61, no. 10, pp. 1497–1498, 1973.
- [50] C. Eckart, “Optimal rectifier systems for the detection of steady signals,” *Scripps Institution of Oceanography*, 1952.
- [51] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, “RASTA-PLP speech analysis technique,” in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, pp. 121–124, 1992.
- [52] D. A. Reynolds and P. Torres-Carrasquillo, “The MIT Lincoln laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations,” tech. rep., DTIC Document, 2004.

- [53] M. Zamalloa, L. J. Rodriguez-Fuentes, G. Bordel, M. Penagarikano, and J. P. Uribe, “Low-latency online speaker tracking on the AMI corpus of meeting conversations,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4962–4965, 2010.
- [54] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.
- [55] R. Sinha, S. E. Tranter, M. J. Gales, and P. C. Woodland, “The Cambridge University March 2005 speaker diarisation system,” in *Interspeech*, pp. 2437–2440, 2005.
- [56] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, “Speaker diarization: From broadcast news to lectures,” in *Machine Learning for Multimodal Interaction*, pp. 396–406, Springer, 2006.
- [57] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J. F. Bonastre, “NIST RT05S evaluation: pre-processing techniques and speaker diarization on multiple microphone meetings,” in *Machine Learning for Multimodal Interaction*, pp. 428–439, Springer, 2006.
- [58] D. A. van Leeuwen, “The TNO speaker diarization system for NIST RT05s meeting data,” in *Machine Learning for Multimodal Interaction*, pp. 440–449, Springer, 2006.
- [59] S. Cassidy, “The Macquarie speaker diarization system for RT04s,” in *NIST 2004 Spring Rich Transcription Evaluation Workshop, Montreal, Canada*, 2004.
- [60] P. Nguyen, L. Rigazio, Y. Moh, and J. Junqua, “Rich transcription 2002 site report, Panasonic speech technology laboratory (PSTL),” in *Proceedings of the 2002 Rich Transcription Workshop*, 2002.
- [61] J.-L. Gauvain, L. Lamel, and G. Adda, “Partitioning and transcription of broadcast news data.,” in *ICSLP*, vol. 98, pp. 1335–1338, 1998.
- [62] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, “Combining speaker identification and BIC for speaker diarization.,” in *Interspeech*, vol. 5, pp. 2441–2444, 2005.

- [63] T. Hain, S. Johnson, A. Tuerk, P. Woodland, and S. Young, “Segment generation and clustering in the HTK broadcast news transcription system,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, vol. 1998, 1998.
- [64] D. Liu and F. Kubala, “Fast speaker change detection for broadcast news transcription and indexing,” in *Sixth European Conference on Speech Communication and Technology*, 1999.
- [65] X. Anguera, M. Aguilo, C. Wooters, C. Nadeu, and J. Hernando, “Hybrid Speech/non-speech detector applied to Speaker Diarization of Meetings,” in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pp. 1–6, 2006.
- [66] R. Barra-Chicote, J. M. Pardo, J. Ferreiros, and J. M. Montero, “Speaker diarization based on intensity channel contribution,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 754–761, 2011.
- [67] X. Anguera, C. Wooters, and J. Pardo, “Robust Speaker Diarization for Meetings: ICSI RT06s Evaluation System,” in *Ninth International Conference on Spoken Language Processing*, ISCA, 2006.
- [68] S. Wrigley, G. Brown, V. Wan, and S. Renals, “Feature selection for the classification of crosstalk in multi-channel audio,” in *Eighth European Conference on Speech Communication and Technology*, ISCA, 2003.
- [69] S. Wrigley, G. Brown, V. Wan, and S. Renals, “Speech and crosstalk detection in multichannel audio,” *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 1, pp. 84–91, 2005.
- [70] T. Pfau, D. Ellis, and A. Stolcke, “Multispeaker speech activity detection for the ICSI meeting recorder,” in *Proceedings of ASRU*, vol. 1, 2001.
- [71] K. Laskowski, Q. Jin, and T. Schultz, “Crosscorrelation-based multispeaker speech activity detection,” in *Eighth International Conference on Spoken Language Processing*, ISCA, 2004.

- [72] K. Laskowski, C. Fugen, and T. Schultz, “Simultaneous multispeaker segmentation for automatic meeting recognition,” *Proc. EUSIPCO, Poznan, Poland*, pp. 1294–1298, 2007.
- [73] K. Laskowski and T. Schultz, “Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings,” in *Proc. of ICASSP*, pp. 993–996, 2006.
- [74] K. Laskowski, G. Karlsruhe, and T. Schultz, “A Geometric Interpretation of Non-Target-Normalized Maximum Cross-channel Correlation for Vocal Activity Detection in Meetings,” in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 89–92, Association for Computational Linguistics, 2007.
- [75] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.
- [76] A. Dempster, N. Laird, D. Rubin, *et al.*, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.
- [77] S. Bozonnet, N. Evans, and C. Fredouille, “The Lia-Eurecom RT’09 speaker diarization system: Enhancements in speaker modelling and cluster purification,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4958–4961, 2010.
- [78] X. Anguera, C. Wooters, B. Peskin, and M. Aguil, “Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system,” in *Machine Learning for Multimodal Interaction*, pp. 402–414, Springer, 2006.
- [79] X. Anguera, C. Wooters, and J. Hernando, “Automatic cluster complexity and quantity selection: Towards robust speaker diarization,” in *Machine Learning for Multimodal Interaction*, pp. 248–256, Springer, 2006.

- [80] X. Anguera and J. Hernando, “Xbic: Real-time cross probabilities measure for speaker segmentation,” *Univ. California Berkeley, ICSIBerkeley Tech. Rep*, 2005.
- [81] A. Malegaonkar, A. Ariyaeinia, P. Sivakumaran, and J. Fortuna, “Unsupervised speaker change detection using probabilistic pattern matching,” *Signal Processing Letters, IEEE*, vol. 13, no. 8, pp. 509–512, 2006.
- [82] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [83] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967.
- [84] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, “Unknown-multiple speaker clustering using HMM.,” in *INTERSPEECH*, 2002.
- [85] H. Kim, D. Ertelt, and T. Sikora, “Hybrid speaker-based segmentation system using model-level clustering,” in *Proc. 2005 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 745–748, 2005.
- [86] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.
- [87] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass, “Exploiting intra-conversation variability for speaker diarization.,” in *INTERSPEECH*, pp. 945–948, 2011.
- [88] W. Campbell, D. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.
- [89] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital signal processing*, vol. 10, no. 1, p. 1941, 2000.

REFERENCES

- [90] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.
- [91] S. Shum, N. Dehak, and J. Glass, “On the use of spectral and iterative methods for speaker diarization,” *System*, vol. 1, no. w2, p. 2, 2012.
- [92] A. Y. Ng, M. I. Jordan, Y. Weiss, *et al.*, “On spectral clustering: Analysis and an algorithm,” *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.
- [93] J. Silovsky and J. Prazak, “Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4193–4196, 2012.
- [94] M. Rouvier and S. Meignier, “A global optimization framework for speaker diarization,” in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.
- [95] H. Tang, S. Chu, M. Hasegawa-Johnson, and T. Huang, “Partially supervised speaker clustering,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 5, pp. 959–971, 2012.
- [96] W.-H. Tsai, S.-S. Cheng, Y.-H. Chao, and H.-M. Wang, “Clustering speech utterances by speaker using eigenvoice-motivated vector space models,” in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP’05). IEEE International Conference on*, vol. 1, pp. 725–728, 2005.
- [97] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space,” *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 6, pp. 695–707, 2000.
- [98] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, “Automatic segmentation, classification and clustering of broadcast news audio,” in *Proc. DARPA Broadcast News Workshop*, p. 11, 1997.

REFERENCES

- [99] J. Campbell *et al.*, “Speaker recognition: A tutorial,” *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.
- [100] F. Bimbot and L. Mathan, “Text-free speaker recognition using an arithmetic-harmonic sphericity measure,” in *Third European Conference on Speech Communication and Technology*, ISCA, 1993.
- [101] A. Willsky and H. Jones, “A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems,” *IEEE Transactions on Automatic Control*, vol. 21, no. 1, pp. 108–112, 1976.
- [102] S. Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, 1998.
- [103] D. Reynolds, E. Singer, B. Carlson, G. O’Leary, J. McLaughlin, and M. Zissman, “Blind clustering of speech utterances based on speaker and language characteristics,” in *Fifth International Conference on Spoken Language Processing*, ISCA, 1998.
- [104] V. B. Le, O. Mella, D. Fohr, *et al.*, “Speaker diarization using normalized cross likelihood ratio,” in *INTERSPEECH*, vol. 7, pp. 1869–1872, 2007.
- [105] T. Cover and J. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.
- [106] B. Zhou and J. H. Hansen, “Unsupervised audio stream segmentation and clustering via the Bayesian information criterion,” in *Interspeech*, pp. 714–717, 2000.
- [107] M. Ben, M. Betser, F. Bimbot, and G. Gravier, “Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs,” in *Eighth International Conference on Spoken Language Processing*, ISCA, 2004.
- [108] J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, J. Martinez, and F. Rabat, “Fast incremental clustering of Gaussian mixture speaker models for scaling up retrieval in on-line broadcast,” in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, vol. 5, 2006.

REFERENCES

- [109] T. Wu, L. Lu, K. Chen, and H. Zhang, “UBM-based real-time speaker segmentation for broadcasting news,” in *ICME 2003*, vol. 2, pp. 721–724, 2003.
- [110] J.-F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, and C. Wellekens, “A speaker tracking system based on speaker turn detection for NIST evaluation,” in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 2, pp. 1177–1180, 2000.
- [111] A. G. Adam, S. S. Kajarekar, and H. Hermansky, “A new speaker change detection method for two-speaker segmentation,” in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 4, pp. 3908–3911, 2002.
- [112] R. Gangadharaiah, B. Narayanaswamy, and N. Balakrishnan, “A novel method for two-speaker segmentation,” in *Interspeech*, 2004.
- [113] G. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [114] A. Tritschler and R. Gopinath, “Improved speaker segmentation and segments clustering using the Bayesian information criterion,” in *Sixth European Conference on Speech Communication and Technology*, pp. 679–682, ISCA, 1999.
- [115] P. Delacourt and C. Wellekens, “DISTBIC: A speaker-based segmentation for audio data indexing,” *Speech Communication*, vol. 32, no. 1-2, pp. 111–126, 2000.
- [116] K. Mori and S. Nakagawa, “Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition,” in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP'01). 2001 IEEE International Conference on*, vol. 1, 2001.
- [117] J. López and D. Ellis, “Using acoustic condition clustering to improve acoustic change detection on broadcast news,” in *Sixth International Conference on Spoken Language Processing*, ISCA, 2000.
- [118] A. Vandecatseye and J. Martens, “A fast, accurate and stream-based speaker segmentation and clustering algorithm,” in *Eighth European Conference on Speech Communication and Technology*, ISCA, 2003.

REFERENCES

- [119] J. Ajmera, I. McCowan, and H. Bourlard, “Robust speaker change detection,” *Signal Processing Letters, IEEE*, vol. 11, no. 8, pp. 649–651, 2004.
- [120] L. Perez-Freire and C. Garcia-Mateo, “A multimedia approach for audio segmentation in TV broadcast news,” in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP’04). IEEE International Conference on*, vol. 1, 2004.
- [121] J. Gauvain and C. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE transactions on speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.
- [122] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, “Improving speaker diarization,” in *RT-04F workshop*, 2004.
- [123] H. Gish, M.-H. Siu, and R. Rohlicek, “Segregation of speakers for speech recognition and speaker identification,” in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pp. 873–876, 1991.
- [124] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, “Strategies for automatic segmentation of audio data,” in *Acoustics, Speech, and Signal Processing, 2000. ICASSP’00. Proceedings. 2000 IEEE International Conference on*, vol. 3, pp. 1423–1426, 2000.
- [125] H. Jin, F. Kubala, and R. Schwartz, “Automatic speaker clustering,” in *Proceedings of the DARPA speech recognition workshop*, pp. 108–111, 1997.
- [126] M. Nishida and T. Kawahara, “Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*, vol. 1, pp. 172–175, 2003.
- [127] P. Sivakumaran, J. Fortuna, and A. M. Ariyaeinia, “On the use of the Bayesian information criterion in multiple speaker detection,” in *Interspeech*, pp. 795–798, 2001.

REFERENCES

- [128] M. Cettolo and M. Vescovi, “Efficient audio segmentation algorithms based on the BIC,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP’03). 2003 IEEE International Conference on*, vol. 6, 2003.
- [129] M. Roch and Y. Cheng, “Speaker segmentation using the MAP-adapted bayesian information criterion,” in *ODYSSEY04-The Speaker and Language Recognition Workshop*, ISCA, 2004.
- [130] P. Delacourt, D. Kryze, and C. Wellekens, “Detection of speaker changes in an audio document,” in *Sixth European Conference on Speech Communication and Technology*, ISCA, 1999.
- [131] P. Zochova and V. Radova, “Modified DISTBIC algorithm for speaker change detection,” in *Ninth European Conference on Speech Communication and Technology*, ISCA, 2005.
- [132] L. Lu and H. Zhang, “Real-time unsupervised speaker change detection,” in *INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION*, vol. 16, pp. 358–361, 2002.
- [133] H. Wang and S. Cheng, “METRIC-SEQDAC: A hybrid approach for audio segmentation,” in *Eighth International Conference on Spoken Language Processing*, ISCA, 2004.
- [134] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, “Clustering speakers by their voices,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, pp. 757–760, 1998.
- [135] W.-H. Tsai, S.-S. Cheng, and H.-M. Wang, “Speaker clustering of speech utterances using a voice characteristic reference space,” in *Eighth International Conference on Spoken Language Processing*, 2004.
- [136] Q. Jin and T. Schultz, “Speaker segmentation and clustering in meetings,” in *Interspeech*, vol. 4, pp. 597–600, 2004.

- [137] J. Hansen, B. Zhou, M. Akbacak, R. Sarikaya, and B. Pellom, “Audio stream phrase recognition for a National Gallery of the spoken word: ”One Small Step”,” in *Sixth International Conference on Spoken Language Processing*, ISCA, 2000.
- [138] D. Vijayasenan, F. Valente, and H. Bourlard, “Agglomerative information bottleneck for speaker diarization of meetings data,” in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pp. 250–255, 2007.
- [139] D. Vijayasenan, F. Valente, and H. Bourlard, “Combination of agglomerative and sequential clustering for speaker diarization,” in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4361–4364, 2008.
- [140] D. Vijayasenan, F. Valente, and H. Bourlard, “Integration of TDOA features in information bottleneck framework for fast speaker diarization.,” in *INTERSPEECH*, pp. 40–43, 2008.
- [141] D. Vijayasenan, F. Valente, and H. Bourlard, “Mutual information based channel selection for speaker diarization of meetings data,” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4065–4068, 2009.
- [142] D. Vijayasenan, F. Valente, and H. Bourlard, “An information theoretic combination of MFCC and TDOA features for speaker diarization,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 431–438, 2011.
- [143] D. Vijayasenan and F. Valente, “Speaker diarization of meetings based on large TDOA feature vectors,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4173–4176, 2012.
- [144] D. P. Ellis and J. C. Liu, “Speaker turn segmentation based on between-channel differences,” in *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*, pp. 112–117, 2004.

REFERENCES

- [145] J. Luque, C. Segura, and J. Hernando, “Clustering initialization based on spatial information for speaker diarization of meetings,” in *INTERSPEECH*, pp. 383–386, 2008.
- [146] X. Anguera, C. Wooters, J. Pardo, and J. Hernando, “Automatic weighting for the combination of TDOA and acoustic features in speaker diarization for meetings,” in *Proc. ICASSP*, pp. 241–244, 2007.
- [147] O. Vinyals and G. Friedland, “Modulation spectrogram features for improved speaker diarization,” in *Interspeech*, pp. 630–633, 2008.
- [148] B. E. Kingsbury, N. Morgan, and S. Greenberg, “Robust speech recognition using the modulation spectrogram,” *Speech communication*, vol. 25, no. 1, pp. 117–132, 1998.
- [149] A. Friedland, B. Vinyals, C. Huang, and D. Muller, “Fusing short term and long term features for improved speaker diarization,” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4077–4080, 2009.
- [150] J. Huang, E. Marcheret, K. Visweswariah, and G. Potamianos, “The IBM RT07 evaluation systems for speaker diarization on lecture meetings,” in *Multimodal Technologies for Perception of Humans*, pp. 497–508, Springer, 2008.
- [151] S. E. Johnson and J. Woodland, “Speaker clustering using direct maximisation of the MLLR-Adapted likelihood,” in *Proc. ICSLP 98*, pp. 1775–1779, 1998.
- [152] S. Johnson, “Who spoke when?-automatic segmentation and clustering for determining speaker turns,” in *Sixth European Conference on Speech Communication and Technology*, ISCA, 1999.
- [153] S. Meignier, J. Bonastre, and S. Igounet, “E-HMM approach for learning and adapting sound models for speaker indexing,” in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, ISCA, 2001.

REFERENCES

- [154] X. Anguera and J. Hernando, “Evolutive speaker segmentation using a repository system,” in *Proc. International Conference on Speech and Language Processing, Jeju Island, Korea*, 2004.
- [155] I. Lapidot, “SOM as likelihood estimator for speaker clustering,” in *Eighth European Conference on Speech Communication and Technology*, ISCA, 2003.
- [156] W. Tsai and H. Wang, “On maximizing the within-cluster homogeneity of speaker voice characteristics for speech utterance clustering,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France*, 2006.
- [157] F. Valente and C. Wellekens, “Variational bayesian speaker clustering,” in *ODYSSEY04-The Speaker and Language Recognition Workshop*, ISCA, 2004.
- [158] F. Valente and C. Wellekens, “Variational bayesian adaptation for speaker clustering,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP’05)*, vol. 1, 2005.
- [159] F. Valente, “Infinite models for speaker clustering,” in *Ninth International Conference on Spoken Language Processing*, ISCA, 2006.
- [160] T. Ferguson, “A Bayesian analysis of some nonparametric problems,” *The annals of statistics*, pp. 209–230, 1973.
- [161] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, “An HDP-HMM for systems with state persistence,” in *Proceedings of the 25th international conference on Machine learning*, pp. 312–319, ACM New York, NY, USA, 2008.
- [162] Y. Teh, M. Jordan, M. Beal, and D. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.
- [163] S. Tranter, “Two-way cluster voting to improve speaker diarisation performance,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP’05)*, vol. 1, 2005.
- [164] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

REFERENCES

- [165] D. Moore, “The IDIAP smart meeting room,” 2002.
- [166] V. Stanford, J. Garofolo, O. Galibert, M. Michel, and C. Laprun, “The NIST smart space and meeting room projects: signals, acquisition annotation, and metrics,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*, vol. 4, pp. 736–739, IEEE, 2003.
- [167] J. S. Garofolo, C. Laprun, M. Michel, V. M. Stanford, and E. Tabassi, “The NIST meeting room pilot corpus,” in *LREC*, 2004.
- [168] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia, *et al.*, “The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms,” *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 389–407, 2007.
- [169] The ISL Meeting Corpus. [Online:
http://www.is.cs.cmu.edu/meeting_room/corpus/; accessed 20-December-2013.
- [170] International Computer Science Institute, “The ICSI meeting corpus,” 2004. [Online]. Available: <http://www.icsi.berkeley.edu/Speech/mr/>. [Accessed: Dec. 20, 2013].
- [171] NIST Meeting Room Pilot Corpus. [Online:
http://www.itl.nist.gov/iad/mig/test_beds/meeting_corpus_1/index.html ; accessed 20-December-2013.
- [172] The AMI corpus. [Online:
<http://corpus.amiproject.org/documentations/overview/>; accessed 20-December-2013.
- [173] S. N. C. of Competence in Research, 2005. [Online:
<http://www.idiap.ch/scientific-research/projects/IM2>;
accessed 12-December-2014].
- [174] A. Consortium, 2004. [Online:
<http://www.amiproject.org/ami-scientific-portal/ami-overview>;
accessed 12-December-2014].

REFERENCES

- [175] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, and others, “The AMI meeting corpus,” in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.
- [176] C. Fredouille and N. Evans, “The LIA RT07 speaker diarization system,” in *Multimodal Technologies for Perception of Humans*, pp. 520–532, Springer, 2008.
- [177] R. W. Schafer and A. V. Oppenheim, *Discrete-time signal processing*. Prentice Hall, 1989.
- [178] J. Luque, X. Anguera, A. Temko, and J. Hernando, “Speaker diarization for conference room: The UPC RT07s evaluation system,” in *Multimodal Technologies for Perception of Humans*, pp. 543–553, Springer, 2008.
- [179] J. E. Cavanaugh, “Unifying the derivations for the Akaike and corrected Akaike information criteria,” *Statistics & Probability Letters*, vol. 33, no. 2, pp. 201–208, 1997.
- [180] A. Larcher, K.-A. Lee, B. Ma, and H. Li, “RSR2015: Database for text-dependent speaker verification using multiple pass-phrases,” in *INTERSPEECH*, 2012.
- [181] M. Przybocki and A. F. Martin, “NIST speaker recognition evaluation chronicles,” in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.
- [182] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,” *Fifth European Conference on Speech Communication and Technology*, 1997.
- [183] W. J. Conover, “Practical nonparametric statistics,” 1999.
- [184] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
- [185] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.