Enhancing the Front-End of Speaker Recognition Systems

Ahmed Isam Ahmed

The thesis is submitted in partial fulfilment of the requirements for the award of the degree of

Doctor of Philosophy

of the

University of Portsmouth

July 2019

Declaration

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

Ahmed Isam Ahmed
July 2019
$\sim$ 45000 Words

Acknowledgements

I would like to express my profound gratitude to the almighty God who bestowed upon me and my family the gentle care during this journey until the completion of this work.

I would like to express my sincere gratitude to Dr John Chiverton for his excellent supervision and for putting his utmost effort in following my work. I owe him a lot for his support and advice and for his guidance which has shaped my mentality as a researcher. Many thanks to my co-supervisor, Dr David Ndzi, for motivating me to do my research on this interesting topic and for his continued support and advice. I would like to thank Prof. Victor Becerra for supporting my work. Thanks to Mr Gary Burton, Dr Jascha Schewtschenko and Mr Erik Nagy for their technical support.

I am thankful to my friend and colleague , Mahmoud Al-Faris, mostly, for backing me up in difficult times. I would like to thank my friend and colleague, Dr Marta Peña Fernández, for her support and encouragement. Thanks to my friend Anas Abubaker and his family for helping me and my family in difficult times. Thanks to Hayder Murad, Dr Mohanad Alhabo, Ali Malik Al-Bdairi, Katerina Karali, Marco Curto, Dr Ahmed Mohammed, Roxane Bonithon, Saúl Armendáriz Puente, Murtadha Al-Maliki, Tim Wigger and all the friends and colleagues whom their companionship has eased the workload of this journey.

I am very grateful to my wife, Aya Mohammed Najeeb, for her continuous encouragement and support along this study. I owe her a lot for taking care of our children and myself and for bearing many responsibilities so that I can spend a lot of time working on this research. I would like to express my profound appreciation to my father, Dr Isam Ahmed, and my mother, Mrs Sabah Ahmed, for their support, encouragement and continuous prayer for me and my family during this journey. I owe my daughter, Yusur, and my son, Alhasan, a lot for me being very busy with my work. Thanks also go to my sister Zahraa Ahmed, her husband Ahmed Mohammed Najeeb and my brother Abdullah Ahmed.

Abstract

A number of enhancements on the front-end of i-vector based speaker verification and binary key based speaker diarization are introduced. This is achieved by tackling the methods of acoustic feature extraction and feature combination and by proposing a source selection of the speech signal and spatial feature transformation for speaker diarization. A new paradigm for the extraction of the Mel-Frequency Cepstral Coefficients (MFCC) speech features is introduced and it is based on determining the cepstral coefficients from suitably selected subsets of the filters in the filter bank. The extraction of the Linear Predictive Cepstral Coefficients (LPCC) is also tackled by having the required estimation of the autocorrelation function approximated as the inverse of the smoothened multitaper spectral estimates.

A Recurrent Neural Network (RNN) based weighted Principal Component Analysis (PCA) approach is introduced for feature fusion in addition to dimensionality reduction. This RNN based approach provides an eigendecomposition of weighted correlation and covariance matrices. This weighted PCA is found to provide a solution that can be robust to outliers and to be an efficient method for weighted-feature fusion.

Two selection approaches of multiple microphones' signals (channel selection) are proposed for speaker diarization in a meeting scenario. One method selects the most diverse signals based on the spatial diversity of the microphones. The second method selects the best quality signals with reference to a signal obtained by combining all of the signals using the beamforming technique. Additionally, a selection of the least reverberated subbands (of microphones' signals) is proposed and it is based on the estimation of the mean gradient of the spectrum of the speech frames. This is found to provide comparable improvements to the case when features are extracted from selected channels but at a lower feature dimensionality.

An analysis is conducted to identify the reasons preventing the binary key based diarization system from operating on spatial features. Depending on the analysis results, a nonlinear transformation of these features is found to be required to enable their integration into this system which noticeably improves the diarization accuracy. Additionally, as opposed to the uniform initialisation method usually used by this diarization system, six non-uniform initialisation methods are proposed and investigated.

Table of Contents

List of Publications

- Ahmed, A. I., Chiverton, J., Ndzi, D. and Becerra, V. (2017). Channel Variability Synthesis in I-Vector Speaker Recognition. In *IET 3rd International Conference on Intelligent Signal Processing (ISP 2017)* (pp. 1–6). IET.
- Ahmed, A. I., Chiverton, J. P., Ndzi, D. L. and Becerra, V. M. (2019). Speaker Recognition using PCA-Based Feature Transformation. *Speech Communication, 110*, 33-46.
- Boosting the Performance of Binary Key Based Diarization: acoustic feature extraction from selected channels and integrating spatial features following nonlinear transformation. *To be submitted in Expert Systems with Applications.*
- Subband Feature Extraction for Multiple Microphone Diarization. *To be submitted in IEEE Signal Processing Letters.*
- Non-Uniform Initialization of a Binary Key Based Diarization System. *To be submitted in Pattern Recognition Letters.*

List of Abbreviations

AJG	Average Joined Gradient
AHC	Agglomerative Hierarchical Clustering
BIC	Bayesian Information Criterion
BK	Binary Key
CE	Clustering Error
CV	Cumulative Vector
DCT	Discrete Cosine Transform
DER	Diarization Error Rate
EER	Equal Error Rate
FFT	Fast Fourier Transform
GMM-UBM	Gaussian Mixture Models-Universal Background Model
GPHAT	Generalised Cross Correlation with Phase Transform
JFA	Joint Factor Analysis
KBM	Binary Key Background Model
LDA	Linear Discriminant Analysis
LPC	Linear Prediction Coefficients
LPCC	Linear Predictive Cepstral Coefficients
MDM	Multiple Distant Microphones
MFCC	Mel-Frequency Cepstral Coefficients
OE-MFCC	Odd-Even MFCC
PCA	Principal Component Analysis
PLDA	Probabilistic Linear Discriminant Analysis
RNN	Recurrent Neural Network
RT	Real Time
SER	Speaker Error Rate
SG-BM	Single Gaussian-Background Model
SN	Skew Normal Distribution
SNR	Signal to Noise Ratio
SVD	Singular Value Decomposition
TDOA	Time Delay of Arrival
TTDOA	Transformed TDOA
WCSS	Within Class Sum of Squares
WPCA	Weighted Principal Component Analysis

List of Figures

List of Tables

Chapter 1

Introduction

Technological advancements have focused on reducing mundane tasks performed by humans or on overcoming human limitations. One of the earliest attempts to utilise a computer in 'talker recognition' was in (Pruzansky, 1963) by programmatically matching time-frequency energy patterns of speech. In (Atal, 1976), it was anticipated that the advances in digital computing would provide the greatest impetus to research on speaker recognition. Today, voice recognition has become a fundamental component of Artificial Intelligence (AI) where leading technology bodies, like Google (Chiu et al., 2018) and Amazon (Purington et al., 2017), are engaging their resources in speech processing research. Speech processing technologies have even found uses in home robots such as the recently introduced social robots, Jibo (Fan & Wang, 2013) and Anki's Vector (Guizzo, 2018), which have the capability to perform human-like verbal interaction with a person.

The ability to recognise individual speakers can help in personalising speech processing based technologies and to ensure a satisfactory level of individuals' privacy and security. The work of this thesis focuses on this particular field of speech processing which is referred to as speaker recognition. This introductory chapter starts by familiarising the reader with the different forms of speaker recognition with highlights on the ones that are the focus of this study. It describes the context of this research as well as the problem to be tackled. Then, it summarises the objectives, the achievements as well as the contributions of this research. It also presents an outline of the thesis structure.

1.1 Background

Speech recognition technologies translate the words uttered by a speaker into a form that is perceivable by a machine. Since voice is a unique bio-characteristic of the speaker, speaker recognition techniques can, for example, secure human-machine interaction by making the machine accept communications from particular speaker(s) only. Speech recognition and speaker recognition have a lot in common but speech recognition is focused on the content of a speaker's speech while speaker recognition focuses on speaker identification, verification and classification, see Fig. 1.1. Other topics that are usually seen as extensions of speaker recognition include speaker detection, segmentation, clustering and tracking (Beigi, 2011).

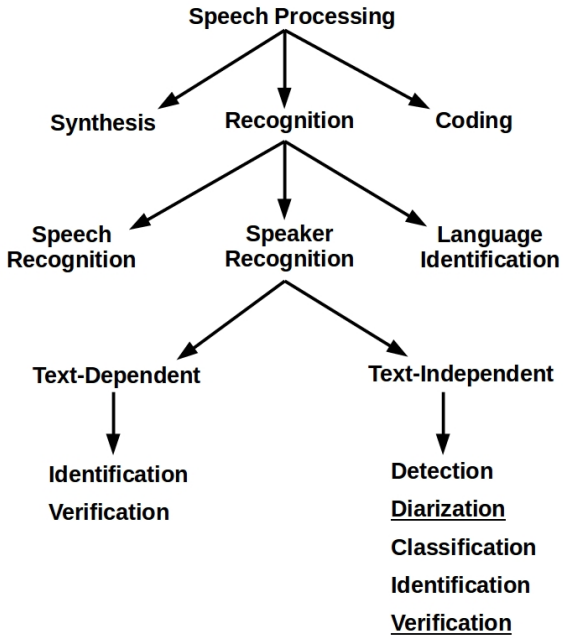


Fig. 1.1 Aspects of Speech Processing. The topics of speaker diarization and verification are the particular applications considered here to demonstrate the benefits of the front-end enhancements. Note that the proposed front-end enhancements could also be used in many other aspects of speaker recognition.

The task of speaker identification aims to decide if a claimed identity is true by comparing the model of the speech presented (to the system) to the speech models of a pre-enrolled group of speakers which are stored in the system. Speaker verification considers the model of the speech provided and contrasts it to both the speech model of the claimed identity and a universal model of speech. Thus it can be used to verify the decision made by a speaker identification module by making sure that the presented speech model did not only happen to

be the closest match to a model in a closed-set of pre-enrolled speakers. On the other hand, speaker classification includes, but is not limited to, the detection of a speaker's age, gender and emotions.

The combination of segmentation and clustering is commonly referred to as speaker diarization. This task attempts to answer the question of 'who spoke when?' in an audio excerpt that involves multiple speakers, see Fig. 1.2. Hence, it also includes speaker detection. Speaker recognition can be generally categorised into text-dependent and text-independent (Campbell, 1997), see Fig. 1.1. A text-dependent system attempts to recognise a speaker who is expected to provide a pre-defined phrase. A text-independent recognition system does not expect the speaker to provide a particular phrase. Speaker diarization strictly falls under this latter category. Text-independent speaker recognition is a challenging research problem for which the National Institute of Standards and Technology (NIST) has been holding a series of yearly evaluations of evolving techniques (Przybocki, 2011 accessed December 13, 2018).

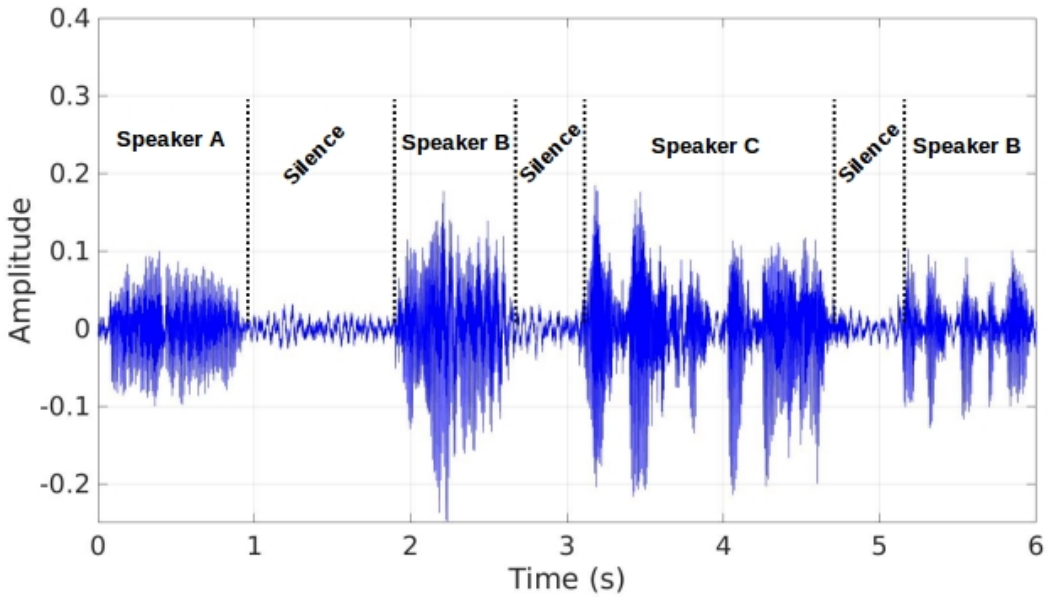


Fig. 1.2 Illustration of the speaker diarization task where the system is supposed to identify speakers' segments within an audio stream.

Text-independent speaker recognition is more versatile in comparison to text-dependent speaker recognition which mostly applies to the task of speaker verification. This is because it is difficult to configure other tasks to work with specific phrases. All of the different tasks require the extraction of suitable features from frames of the speech signal. Text-dependent

systems usually exploit speech recognition technologies like a Hidden Markov Models (HMM) based speech recogniser (Reynolds & Rose, 1995). Since the speaker is expected to provide a specific sentence, an HMM models the temporal sequencing of the speech sounds from one feature frame to another where those were extracted from the sentence. On the contrary, speaker modelling in text-independent speaker recognition attempts to surpass phonetic variations by, for example, averaging across the feature frames as the simplest form of a model.

Text-independent speaker recognition is the scope of this research where the work focuses on speaker verification and diarization. The interest in these particular tasks comes from the fact that they are often desired in many practical applications, see e.g (Kinnunen et al., 2012; Moattar & Homayounpour, 2012; Rosenberg, 1976). Speaker recognition can make use of someone's voice as a biometric measure in access control (security) and forensic applications to name a few. Speaker diarization is particularly useful in audio indexing where it can be used to automatically transcribe an audio recording of mixed speakers. The aim and objectives of this research are set to improve the performance of the widely recognised i-vector based speaker verification system (Dehak et al., 2011) and the fast binary key based speaker diarization system (Anguera & Bonastre, 2011).

1.2 Problem Definition and Context

This work focuses on the problem of the front-end performance in providing a reliable representation of speakers' speech for text-independent speaker recognition in the contexts of speaker verification and speaker diarization. Text-independent speaker-modelling techniques have attracted a considerable amount of research and witnessed a number of advancements as in the i-vector speaker-modelling presented in (Dehak et al., 2011) to compensate for channel variability. On the other hand, it appears that, recently, there have been fewer research works specifically targeting the front-end processes, such as, the feature extraction process that is designed to reflect speech production and perceptual mechanisms.

Although a system's front-end is commonly considered to include feature extraction, this work considers a number of issues that can also be considered to be part of the front-end. Those include: speech signal sourcing, speech feature qualities, the techniques by which they can be combined as well as the statistical condition of the features.

A system's front-end can also influence the effective complexity of the overall system which can sometimes prevent the system from operating within realistic computational bounds. This is to be simultaneously addressed when presenting solutions to enhance the

performance of the front-end. For example, the binary key based diarization system to be studied here is a very fast system that can perform in real-time but it suffers from somewhat limited performance. Such a system can benefit from a well designed front-end that does not considerably affect its appealing speed.

1.3 Aim and Motivation

This research aims at providing robust and efficient enhancements for the front-end of speaker verification and diarization systems. The outcomes of this research could also provide a positive impact on other speaker recognition techniques and applications. Although it should be noted that some aspects are more related to the speaker diarization task, nevertheless, the achievements of this work are not necessarily limited to binary key based diarization.

Speaker recognition systems can have different configurations and modelling techniques depending on the task to be performed but they usually share similar types of front-ends. This has motivated this research to specifically focus on the front-end given the expectation that the outcomes can be useful to a range of tasks in the speaker recognition field.

1.4 Objectives

This section summaries the objectives of this research. They include the following:

- Study existing research in speaker recognition systems and their front-end processes.
- Identify and address the limitations of the extraction methods of speech features that are deemed reliable by most speaker recognition systems.
- Review existing feature fusion techniques and develop a robust and efficient fusion methodology.
- Identify important features for speaker diarization and enable the binary key based diarization system to integrate multiple sources of information as done by other systems.
- Improve acoustic feature sourcing for speaker diarization when a conversation is recorded by multiple microphones.
- Develop suitable methods for the initialisation of the binary key based diarization system.

1.5 Achievements

This section summarises the outcomes of this research. This work:

- Introduced a new paradigm for MFCC extraction based on odd and even subsets of filter banks.
- Fitted the multitaper spectrum estimation method in the extraction of LPCC features.
- Introduced a weighted PCA technique based on a recurrent neural network for feature fusion.
- Presented a data augmentation method based on adding simulated Gaussian channel effect to enable the establishment of the i-vector verification system for the evaluation of feature extraction and fusion methodologies.
- Introduced non-linearly transformed Time Delay of Arrival (TDOA) features using the Box-Cox power transformation which enables the integration of spatial features in the binary keys diarization system.
- Presented two channel selection methods to provide suitable signals for the extraction of acoustic (MFCC) features. This is related to speaker diarization where a concatenation of features extracted from selected channels is used.
- Introduced a new framework of acoustic feature extraction from selected least reverberated channels' subbands.
- Introduced an initial cluster purification method combined with the k-means algorithm for the initialisation of the binary keys diarization systems. Binary keys and cumulative vectors are used together with Jaccard coefficient and cosine similarity metrics.

1.6 Thesis Outline

This section describes the structure of the thesis and directs the reader to the chapters where the pre-described objectives are addressed. For convenience, results and evaluations of the proposed methodologies are reported separately in the relevant chapters.

- **Chapter 2**: this chapter presents a review of previous work on acoustic and spatial features, feature fusion with PCA, channel selection, speaker verification systems as

well as speaker diarization systems and their initialisation. It also includes the technical background about the algorithms and systems that this research builds on.

- **Chapter 3**: the proposed paradigm for odd-even MFCC feature extraction is presented in this chapter. This is followed by describing the methodology for multitaper-fitted LPCC feature extraction. The methodology of data augmentation for the establishment of the i-vector speaker verification system is also described which makes it possible to evaluate its performance with the proposed features.

- **Chapter 4**: this chapter introduces the framework of RNN-based PCA. It highlights a number of critical aspects that should be considered when performing the principal component analysis. Then, it describes the RNN solution for the eigenvalue problem. It also presents the weighting criterion of the feature vectors which aims to down-weight the contribution of noisy and outlying feature vectors to the extraction of the principal components. Weighted RNN-based PCA is used for dimensionality reduction and fusion of the features presented in Chapter 3. This chapter reports the evaluation of the i-vector speaker verification system using the features obtained using weighted PCA.

- **Chapter 5**: the speaker diarization problem is specified in this chapter whose objectives target the performance of binary key based diarization. It focuses on two issues: the selection of suitable channels for the extraction of acoustic features and the statistical condition of TDOA (spatial) features. It presents two channel selection methods one is based on channels' spatial diversity and the other is based on channels' quality. It then presents an analysis of the behaviour of the binary key based system to identify the requirements for integrating TDOA features. Accordingly, it identifies a suitable non-linear transformation of TDOA features. The performance of binary key based diarization is evaluated using acoustic features extracted from selected channels, transformed TDOA features and their integration with acoustic features in a systems' score fusion fashion.

- **Chapter 6**: this chapter comprises two distinct parts. The first is related to channel selection where it presents the methodology for identifying the least reverberated subband across the available channels. Then it describes the feature extraction framework from selected subbands. This is evaluated using binary key based diarization. The second part introduces six initialisation methods specific to the binary key based diarization framework. Then, it performs thorough evaluations to identify the most robust initialisation method.

- **Chapter 7**: the conclusions and future work are introduced in this chapter. The methodologies presented in the previous chapters are discussed in groups according to the technical relations amongst them.

1.7 Thesis Contributions

This work makes the following contributions to the field of speaker recognition:

- Unlike the commonly used overlapped filters bank in MFCC extraction, non-overlapped filters subsets are proposed consisting of the odd and even filters which exhibit a lower residual correlation in their covariance matrices. This is found to enhance MFCC features as indicated by the performance of speaker verification and diarization systems investigated here.

- The estimation of the autocorrelation function in LPCC extraction is achieved by calculating the inverse Fourier transform of the smoothed multitaper spectrum; based on the Wiener-Khinchin theorem. This is found to enhance these features as indicated by the performance of the i-vector based verification system.

- PCA may no longer be simply seen as a dimensionality reduction technique in the field of speaker recognition. Weighted PCA is found to be a more robust solution. It can be very useful in feature fusion as it provides the possibility of assigning different weights to different features.

- Other than for the purpose of data augmentation, added Gaussian channel effect can help in modelling general mismatch between enrolment and test data that can be caused by transmission channels.

- Spatial (TDOA) features have skewed distribution and their normalisation by a non-linear transformation is proposed here which enables their integration in the binary keys system. The normalisation can also equally be used in other diarization systems consisting of modelling techniques that assume normality.

- Although signals combination using beamforming is successful, it makes somewhat limited use of a rich resource. As shown in this work, extracting acoustic features from selected channels and channels' subbands is more efficient and presents higher diarization accuracy.

- Binary key based diarization can efficiently benefit from non-uniform initialisation methods that are compatible with its fast performance.

The following diagram illustrates where each of the contributions are made in the thesis.

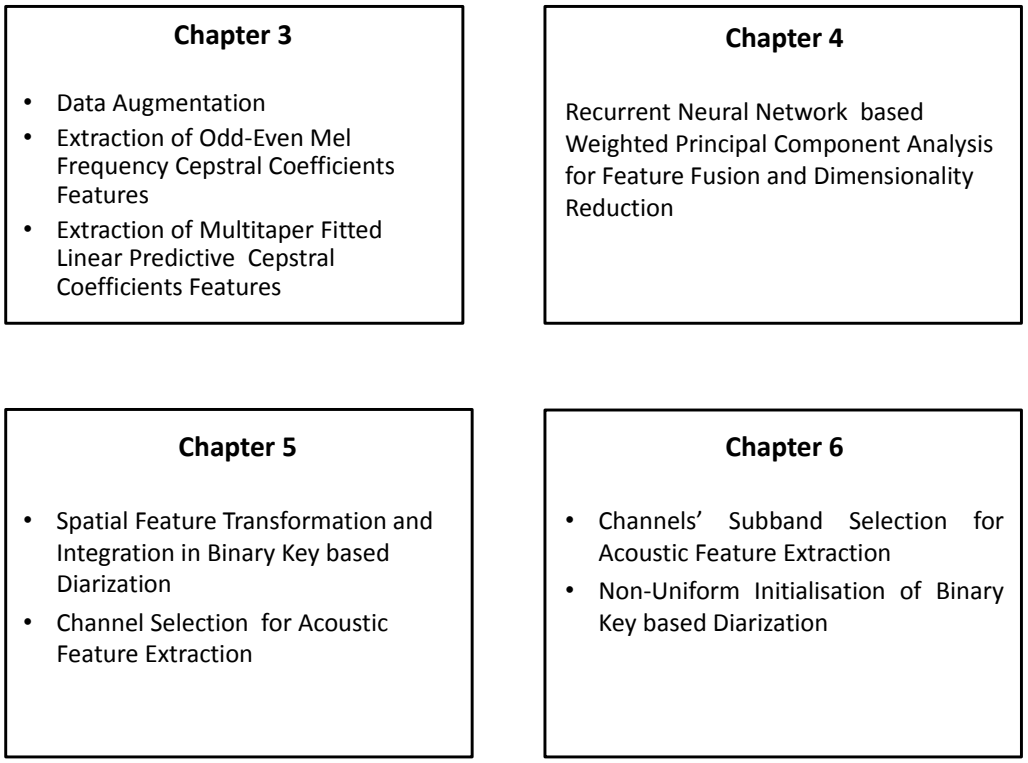


Fig. 1.3 Illustrative diagram of thesis contributions in each chapter.

1.8 Summary

This chapter defined the problem, aim and objectives of this research and summarised the achievements and contributions of the work conducted in this thesis. The description of the thesis structure highlighted the achievements of each chapter. The background given in this chapter identified the scope of this research in the filed of speaker recognition. The next chapter expands on the technical aspects related to this scope and reviews the related works.

Chapter 2

Literature review

This chapter is a review of the literature surrounding the scope of this research. It identifies research questions to be addressed and the enhancements needed. It also attempts to cover the technologies that have found most success to date with a particular focus on the front-end of speaker recognition systems. The front-end is of particular interest as it influences the performance of the later stages of a system.

The review starts with a 'low level' representation of the speaker: the features. This is followed by details of feature fusion techniques. The enhancements that can be achieved in these aspects are transferable to the performance of speaker recognition systems in general. Afterwards, the chapter reviews a higher level representation of speakers, i.e. speaker modelling. It then covers aspects of the research that resulted in the development of the widely recognised i-vector based verification system.

Finally, speaker diarization approaches are reviewed with special focus on binary key based diarization because of its fast performance potentially making it suitable for a large number of applications. However, existing binary key based diarization systems do not possess very competitive diarization accuracy. Therefore, innovative approaches are needed to improve its performance. In the framework of speaker diarization, multiple sources of the speech signal are usually available. The work here shows how these are currently being used and draws the attention to alternatives that are feasible and also possible modifications that can make better use of such resources.

2.1 Feature Representation of the Speaker

The main processing module at the front-end of speaker recognition systems extracts features from appropriate observations made about the speakers. Feature extraction can be described as a number of signal processing procedures, based on some theory or theories, to capture particular aspects of information from a raw measurement. Acoustic features extracted from speech signals are probably used in all speaker recognition systems, see e.g Kinnunen & Li (2010) and Anguera et al. (2012). Spatial features, mainly the difference between the arrival of the speech signal at different acquisition points, are also used in speaker diarization systems as an indicator of speakers' locations in, for example, a meeting room. This section presents a review about these two categories of features.

2.1.1 Acoustic Features

Since the sampled and quantized speech signal is an acoustic measurement, the extracted features are known as acoustic features. In speaker recognition, the main objectives of the extraction algorithms according to Wolf (1972) should be to produce features that

- Provide high discrimination between speakers and low sensitivity to inter-speaker variations;
- Are robust to noise and other distorting effects, for example, channel distortions;
- Are easy and fast to extract;
- Are difficult to synthesise for impersonation purposes;
- Capture unique characteristics about the speaker's voice, especially, in the case of combinations of features.

In acoustic feature extraction, physiological characteristics (such as the shape of the vocal tract) of the speaker can be seen as the major piece of information to be captured from the speech signal as stated in Beigi (2011). Additionally, behavioural aspects (such as the speaking style) of speakers which are delivered in their speech can also be transformed into features. Physiological and behavioural properties of speakers were the basis of categories of features as presented in Tirumala et al. (2017). Kinnunen & Li (2010) divided acoustic features into five categories based on their physical interpretation. For the purposes of the work here, acoustic features can also be divided into three classes according to their discrimination capability and computation complexity as illustrated in Fig. 2.1.

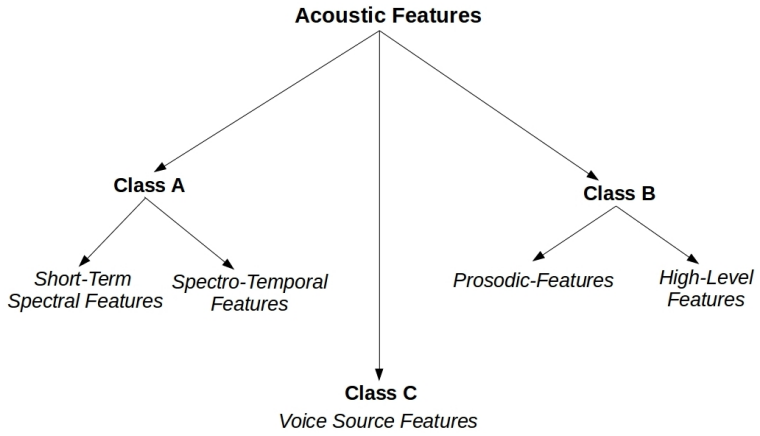


Fig. 2.1 Acoustic feature classes.

- **Class A: low Complexity Features**

  The main downside in the features of this class is their sensitivity to noise Tirumala et al. (2017).

  1. *Short-term spectral features* are a relatively discriminative category of features which are fast to compute. They are extracted from very short segments of speech in the range of 20 ms to 30 ms. In this range, the speech signal is considered stationary because the speed of articulation movements does not change within such a short duration. These features represent the colour of speech in addition to the resonance of the supra-laryngeal vocal tract Benesty et al. (2007), Kinnunen & Li (2010).
  2. *Spectro-temporal features* are another type of features which are not as discriminative as short-term spectral features but are even computationally simpler to achieve. They are almost always used with short-term spectral features. These features are the first and second derivatives of short-term spectral features, hence, they represent formant transitions and can span larger temporal ranges. They are called suprasegmental features as the information they represent exceeds phone and phoneme limits which are the smallest linguistic segments, see e.g Lehiste (1976) .

- **Class B: high Complexity Features**

  The main downside of these features is that they are easy to mimic, see Beigi (2011) and Kinnunen & Li (2010).

1. *Prosodic features* are suprasegmental types of features that involve pitch or loudness and their variations. These features are not very reliable to distinguish between individual speakers because they are easy to impersonate. In this regard, they can be more appropriate for categorical detection as in Kumar et al. (2011) where they were used for gender classification.
2. *High-Level features* are features that attempt to capture information about the attitude of the speaker Doddington (2001). This type of information is called lexicon which can be defined as the type of words used by the speakers. The extraction of such features is computationally demanding where it can involve the use of other systems like a speech recogniser. These features were first introduced in Doddington (2001).

- **Class C: medium Complexity Features**

  Other than the features of classes A & B, *voice source features* is a feature category that characterises the source of voice, for example, the glottal pulse shape. They are more reliable than the features of class B as they carry speaker-specific information. They are, however, less discriminative than short-term spectral features. The methods proposed for the acquisition of these 'glottal features' are more demanding than the extraction methods of short-term spectral features. However, they are less complicated than the extraction of high-level features, such as the so called idiolect introduced in Doddington (2001).

Compared to other feature types, short-term spectral features exhibit appealing properties which made them the focus of a high volume of research, see Tirumala et al. (2017). These features are difficult to control and mimic because their extraction do not involve capturing any information about the attitude of the speaker. The common signal processing condition that is shared between the extraction methodologies of this category of features is the spectral estimate that is forced to be carried over short segments of speech. This is because the speech signal is non-periodic and non-stationary but it is assumed to be periodic and stationary over short temporal ranges.

Short-term spectral features mainly differ in the manner of the spectral decomposition they use. This decomposition could either be adaptive as in linear prediction analysis where the analysis filter poles are distributed on the peaks of the spectrum Dautrich et al. (1983). Alternatively it could be fixed where a pre-designed set of filters (filter bank) are used to perform the spectral decomposition. The sizes and spacing of filters in the filter banks can

differ according to the theory behind the scale used in their design, see Beigi (2011). The mel-scale is one of the most common scales where the filter bank has more emphasis on lower frequencies similar to the human auditory system Makhoul & Cosell (1976). Different shapes of filters can be used in the mel-scale, for example, triangular, rectangular and Gaussian.

The most popular types of short-term spectral features are: Linear Predictive Cepstral Coefficients (LPCC) given by Rabiner & Juang (1990), Mel-Frequency Cepstral Coefficients (MFCC) proposed by Davis & Mermelstein (1980) and Perceptual Linear Prediction (PLP) features introduced in Hermansky (1990). The extraction of LPCC features is based on a theory of the speech production mechanism while the extraction of MFCC features is based on speech perception by the human auditory system. The fundamentals behind the extraction of PLP features can be viewed as a combination of the concepts behind both LPCC and MFCC.

MFCC features and its variations form 97% of the feature extraction methods used in the recent literature as studied by Tirumala et al. (2017). This is because, experimentally, it was found to be successful. Hence, works on its improvement are ongoing and it will also be the focus of a part of this work. One of the recent works on MFCC is the combination of a Gammatone filter bank and multitaper spectrum estimation in its extraction Meriem et al. (2017). The work coupled the advantages of low variance multitaper spectral estimates with the robustness to noise of the auditory Gammatone filter banks. Hence, improved performance was obtained for speaker verification under white, babble and factory noise sources.

Feature combination (fusion) is a common method to improve the front-end of speaker recognition systems, see e.g Neustein & Patil (2012). Features that are different but somehow complementary can be combined so that they provide a richer set of information about the speaker. As stated earlier, PLP shares similarity in its extraction with LPCC and MFCC features. For example, PLP also uses filter banks for spectral decomposition. Accordingly, MFCC and LPCC features are a reasonable choice for combination and an improvement in the extraction of LPCC features is also presented in this work.

2.1.1.1 Mel-Frequency Cepstral Coefficients (MFCC)

The theory of speech perception-based spectral decomposition (using the mel filter bank) is the fundamental concept behind MFCC proposed by Davis & Mermelstein (1980). These features are extracted for short frames of the speech signal. The frame size is usually 25 ms with an overlap of 60%. In conventional MFCC, the frames are smoothed using a Hamming window then the Discrete Fourier Transform (DFT) is used for spectrum estimation. In the

literature of MFCC, DFT is usually referred to as the Fast Fourier Transform (FFT) given the fact that this is how it is implemented. In Kinnunen et al. (2010), the multitaper spectrum estimation method presented in Thomson (1982) was first included in the extraction of MFCC features. In multitaper spectrum estimation, the speech frame is simultaneously windowed by multiple orthogonal windows instead of just a single window and the outputs are averaged resulting in a smooth spectral estimate. After spectrum estimation, the magnitude of the spectrum is calculated. The filter bank is then used to concentrate the spectrum energy into a set of frequency bands (defined by the filters). A bank of triangular filters is defined on a scale referred to as the mel-scale. An illustration of the filter bank is provided in Chapter 3 (Fig 3.3).

The mel-scale is defined as a logarithmic mapping of the frequencies of the linear scale κ_f using the following approximate transformation originally given by Makhoul & Cosell (1976)

$$\zeta_f = 2595 \times \log_{10}\left(1 + \frac{\kappa_f}{700}\right), \tag{2.1}$$

where this transformation and the associated constants are obtained from empirical studies that attempt to measure the psychological sensation of pitch (a perceptual property of sound) Hartmann (2004).

Let M be the total number of overlapping filters. The filters are linearly spaced on the mel-scale and the spacing, Δ, is determined with

$$\Delta = \frac{\zeta_{fmax} - \zeta_{fmin}}{M+1}, \tag{2.2}$$

where ζ_{fmax} and ζ_{fmin} correspond to κ_{fmax} and κ_{fmin} which are the range of the frequency band of interest: 50Hz to 4KHz for telephone speech[1]. Filters' centres on the mel-scale are given by

$$\zeta_{f_c}(m) = m\Delta \quad \text{where} \quad m = 1\ ,\ldots\ ,\ M. \tag{2.3}$$

Let (2.1) be inverse transformed by a function $\hat{\psi}$ such that

$$\kappa_f = \hat{\psi}(\zeta_f) = 700(10^{\zeta_f/2595-1}). \tag{2.4}$$

A bank of overlapping triangular filters $\mathscr{H}(m, \kappa_f)$ with the centres of (2.3) can then be

[1]The lowest audible frequency is 20Hz. However, the low cut-off frequency of wideband transmission systems is around 50Hz, see (Valin & Lefebvre, 2000).

determined in the linear frequency scale as:

$$\mathscr{H}(m,\kappa_f) = \begin{cases} \dfrac{\hat{\psi}(\zeta_f) - \hat{\psi}(\zeta_{f_c}(m-1))}{\hat{\psi}(\zeta_{f_c}(m)) - \hat{\psi}(\zeta_{f_c}(m-1))} & \textbf{for} \quad \hat{\psi}(\zeta_{f_c}(m-1)) \leq \hat{\psi}(\zeta_f) < \hat{\psi}(\zeta_{f_c}(m)); \\ \dfrac{\hat{\psi}(\zeta_f) - \hat{\psi}(\zeta_{f_c}(m+1))}{\hat{\psi}(\zeta_{f_c}(m)) - \hat{\psi}(\zeta_{f_c}(m+1))} & \textbf{for} \quad \hat{\psi}(\zeta_{f_c}(m)) < \hat{\psi}(\zeta_f) \leq \hat{\psi}(\zeta_{f_c}(m+1)); \\ 0 \quad \text{elsewhere.} \end{cases} \tag{2.5}$$

By definition, the cepstrum is obtained by taking the inverse DFT of the log of the speech spectrum, see e.g Benesty et al. (2007). The filter bank log-energies are determined as in the following

$$\mathscr{H}_{\mathscr{E}}(m) = \log_e \left\{ \sum_{\kappa_f=1}^{K_f} \mathscr{H}(m,\kappa_f) |\mathbf{s}_f(\kappa_f)| \right\} \tag{2.6}$$

where $|\mathbf{s}_f(\kappa_f)|$ is the speech spectrum magnitude and κ_f is the higher limit of the speech spectrum.

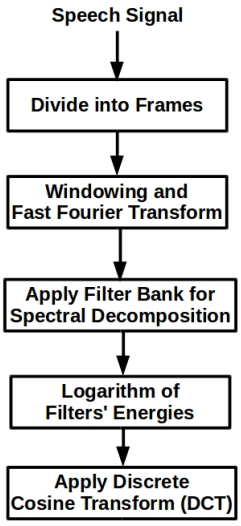


Fig. 2.2 Illustrative diagram of MFCC feature extraction.

The cepstral coefficients (comprising the cepstrum) are calculated by applying the Discrete Cosine Transform (DCT) to the log of the filter bank outputs (for MFCC). Cepstral coefficients are preferred as speech features in speaker recognition because of their inherent invariance to linear spectral distortions, see e.g Beigi (2011). The DCT is especially useful due to its decorrelating properties enabling it to help separate out the important information

contained in the log-energies of filter bank outputs as they are highly correlated. The cepstral coefficients are obtained using the DCT as follows

$$\text{MFCC}_r = \sum_{m=1}^{M} \mathscr{H}_{\mathscr{E}}(m, \kappa_f) \cos\left[r\left(m - \frac{1}{2}\right)\frac{\pi}{M} \right] \quad \text{for} \quad r = 1, 2, \ldots, R \tag{2.7}$$

where R is the number of MFCC cepstral coefficients.

The DCT is applied to all the filter bank output log-energies together. As such, narrow-band noise affects the entire set of DCT coefficients because the log-energy of each filter's output contributes to all of the coefficients, see Sahidullah & Saha (2012). Mostly for this reason, a number of works are found in the literature where speech features are extracted separately from individual subbands of the speech spectrum as discussed below.

Subband feature extraction has been used for noisy speech recognition by e.g Tibrewala & Hermansky (1997). The features were obtained from the power spectrum values of the PLP filter bank followed by cube-root compression and then further processed for loudness equalisation. The recognition output was achieved by merging the results of classifiers acting separately on each subband. The scheme of the system aimed to allow selective de-emphasis of unreliable subbands given the assumption that the speech signal can be partially degraded by frequency-selective noise. The number of subbands of the full spectrum was 2,4 and 7. The frequency range for each subband is given in Table 2.1. This work provided improved system performance for speech corrupted with a variety of different noise sources including destroyer-engine, factory, pink, babble and car engine noise sources. However, for clean speech, the performance was similar to full band feature extraction.

No. of Subbands	Frequency Range for each Subband
2	0-1140 Hz & 1046-4000 Hz.
4	0-765 Hz, 700-1640 Hz, 1515-2700 Hz & 2100-4000 Hz.
7	0-360 Hz, 330-640 Hz, 580-950 Hz, 860-1360 Hz, 1265-1920 Hz, 1800-2700 Hz & 2515-4000 Hz.

Table 2.1 The frequency ranges for the subbands considered by Tibrewala & Hermansky (1997).

Also for noisy speech recognition, Chen et al. (2000) presented a cosine transformation for blocks of the mel filter bank outputs as opposed to applying the DCT to all the outputs at once. The resultant features were referred to as Block Discrete Cosine Transform based MFCC (BMFCC). Subband features were concatenated and used in one recognition system. The test data used were contaminated with several types of noise including voice babble,

factory and car engine noise sources. The training data, however, was kept clean. Two spectrum subbands were chosen with the ranges of 0-1257 Hz and 1104-4000 Hz. BMFCC was found to outperform conventional MFCC under noisy conditions and to provide slight improvements for clean test data.

In speaker recognition, Besacier & Bonastre (2000) also addressed distortions caused by noisy environments that partially affect the speech spectrum. That work was motivated by the success of subband feature extraction in speech recognition. The spectrum considered ranged from 47 Hz to 7597 Hz and it was decomposed using a bank of 24 filters of the mel scale. Twenty-one subbands were chosen, each containing a subset of 4 filters which are highly overlapped such that the first subband had the filters 1-4 and the second had the filters 2-5 and so on. For each subband, the extracted feature vectors of a training sample were modelled by a single Gaussian and the scoring was achieved by determining the log-likelihood value of the test sample's corresponding feature vectors from the Gaussian model. The identification system had 21 sub-systems where the log-likelihood values were combined for a global score. It was centred on the identification task and given the target of subband feature extraction it could be considered difficult to scale to other recognition systems. Also, the approach provided similar performance compared to full band feature extraction for clean and telephone speech. However, it performed better for speech recorded in noisy environments.

Chakroborty et al. (2007) extracted MFCC features separately from two sets of filter banks. One was designed according to the conventional distribution of filters on the mel-scale and the other was an inverted copy of the first one. Hence, while the first put more emphasis on lower frequencies, the latter put more emphasis on higher frequencies. The aim of the idea was to capture complementary information by using the separately extracted features together in a speaker identification system. The methodology improved speaker identification performance when the features of each set were used in separate sub-systems with score fusion.

For speaker identification and verification, Kim et al. (2008) presented a similar approach to Besacier & Bonastre (2000) that additionally shown improvements on clean speech. The features of the subbands were used together or separately in sub-recognition systems. Experiments included full band cepstral coefficients extracted from the outputs of mel filter bank of 33 filters and from 2,3 and 4 subbands that roughly contained equal numbers of filters. The evaluation included clean speech contaminated with eight types of noise such as airport, restaurant and car noise sources. The work also included a feature selection to prevent noisy speech frames from contributing to the recognition scores. The proposed methodology outperformed conventional MFCC under noisy conditions. For clean speech,

the improvement in the performance was given by using sub-systems each dealing with features extracted from one of the subbands.

In Sahidullah & Saha (2012), block based MFCC was proposed for the extraction of the cepstral coefficients. In addition to tackling the problem of narrow band noise, the methodology also aimed to prevent the three peaks associated with the formants of speech from affecting each other when extracting the cepstral coefficients. The work considered two subbands, one ranging from 0 to 883.17 Hz and the other ranging from 745.93 to 4000 Hz. Blocks of subsets of a bank of 20 filters distributed on the mel scale were used to extract subband cepstral coefficients. Experiments included non-overlapping blocks (of filters) and blocks overlapping by no more than two filters. The extracted features were concatenated and used for speaker verification where the performance exhibited improvements for clean and noisy speech. For the purpose of spoofing countermeasures, Paul et al. (2017) extracted cepstral coefficients from overlapping blocks of the inverted mel-scale filter bank introduced by Chakroborty et al. (2007).

It can be observed that previous work in subband feature extraction mostly aimed at tackling the problem of the presence of narrow band noise in the speech signal. The methodologies presented provided slight improvements or similar performance to the case of using full band feature extraction of speech not contaminated with artificial noise. The performance improved further when sub-systems dealt with subband features and the scores were fused, even for non-noisy speech. However, this might be difficult to scale to other or all recognition systems and can be impractical as in the case of Besacier & Bonastre (2000).

The performance of block-transformation for the extraction of MFCC in Chen et al. (2000) and Sahidullah & Saha (2012) was compared to full band application of the DCT in terms of the residual correlation. For a set of cepstral coefficients, the residual correlation [1] in the associated correlation matrix was used in (Sahidullah & Saha, 2012) to indicate how well the DCT transformation compacted the output of the filter bank, where the lower the residual correlation the better transformation.

The work proposed here focuses on the selection of particular subsets of filter banks' in light of their associated correlation matrices. As filter bank based spectral-decomposition is a transformation of the speech spectrum, it is argued in this work, that the lower the residual correlation of the filter bank's correlation matrix the better the performance of this transformation; as strictly related to the concept behind MFCC. Subband DCT works referred to earlier, had subsets of filter banks that caused the residual correlation of their correlation

[1] Residual correlation is the mean of the absolute values of all the off diagonal elements of the correlation matrix, see e.g Sahidullah & Saha (2012).

matrix to increase. This is caused by the following: the filters are overlapped and the subsets have a lower number of adjacent filters than those of the full set. These points are addressed in the methodology presented here in Chapter 3 for extracting MFCC coefficients from subsets of a filter bank.

2.1.1.2 Linear Predictive Cepstral Coefficients (LPCC)

Linear Predictive Cepstral Coefficients (LPCC) features are based on Linear Prediction Coding (LPC) which models speech production mechanism. This makes LPCC a good feature candidate for combination with MFCC since it adds knowledge from a different perspective. Linear prediction models speech production as an autoregressive process where a speech frame can be predicted from past frames (delayed versions of the frame). This process is fitted to an all-pole digital filter model where the coefficients of the filter represent the vocal tract (the spectral envelop). Hence, the goal is to find the filter coefficients that minimise the error between the speech frame and its predicted version. This in turn is realised using autocorrelation, see e.g Broersen (2006).

In the all-pole filter model, a speech sample s_n is assumed to be a linear combination of R past samples and an input u_n Makhoul (1975)

$$s_n = -\sum_{r=1}^{R} a_r \, s_{n-r} \; + G \, u_n, \tag{2.8}$$

where G is the filter gain and a_r is a filter coefficient (LPC coefficient) of order r. The transfer function of the filter is expressed as

$$H(z) = \frac{G}{1 + \sum\limits_{r=1}^{R} a_r \, z^{-r}}. \tag{2.9}$$

The problem is to find the coefficients of this filter model. It is assumed that s_n can be predicted from previous samples. Denote this predicted signal by $\tilde{s}_n = -\sum\limits_{r=1}^{R} a_r \, s_{n-r}$, then the error between this signal and the actual signal is

$$e_n = s_n - \tilde{s}_n = s_n + \sum_{r=1}^{R} a_r \, s_{n-r}, \tag{2.10}$$

and the sum of the squares error is

$$E = \sum_{n=-\infty}^{+\infty} e_n^2. \tag{2.11}$$

In the practical case, the signal length is limited and the sum of squares error is determined as follows

$$E = \sum_{n=0}^{N-1} \left(s_n + \sum_{r=1}^{R} a_r \, s_{n-r} \right)^2. \tag{2.12}$$

This problem can be solved using the method of the least squares where the parameters can be achieved by minimising the error with respect to each parameter a_r by taking the derivative $\frac{\partial E}{\partial a_r}$. This minimisation problem can also be expressed in terms of the autocorrelation function of the signal, for frame l, in the form of R linear equations (see e.g Beigi (2011))

$$\sum_{r=1}^{R} a_r \hat{r}_l(|i-r|) = \hat{r}_l(i), \tag{2.13}$$

where $i = \{1, 2, ..., R\}$. The autoregressive model of LPC uses the assumption that the signal is stationary, hence the autocorrelation function is determined for speech frames of short length (25 - 40 ms). The autocorrelation function of frame l is

$$\hat{r}_l(i) = \sum_{n=0}^{N-1-R} s_{l,n} s_{l,n+i}. \tag{2.14}$$

Equation (2.13) can be expressed in matrix form (known as the Yule-Walker equations Kendall (1949)) as

$$\mathbf{R}_l \mathbf{a}_l = \mathbf{r}_l, \tag{2.15}$$

where $\mathbf{R}_l$ is the autocorrelation matrix and $\mathbf{r}_l$ is the autocorrelation vector. Thus the vector of LPC coefficients is $\mathbf{a}_l = {\mathbf{R}_l}^{-1} \mathbf{r}_l$.

$\mathbf{R}_l$ is a Toeplitz matrix which makes it simpler to solve for $\mathbf{a}_l$ by applying Levinson-Durbin algorithm Durbin (1960) to $\mathbf{r}_l$ without needing to compute $\mathbf{R}_l^{-1}$ (Beigi, 2011). This is because $\mathbf{r}_l$ comprises the same elements as $\mathbf{R}_l$ Makhoul (1975) as illustrated below

$$\begin{bmatrix} r_l(0) & r_l(1) & r_l(2) & \dots & r_l(p-1) \\ r_l(1) & r_l(0) & r_l(1) & \dots & r_l(p-2) \\ \vdots & \vdots & \vdots & & \vdots \\ r_l(p-1) & r_l(p-2) & r_l(p-3) & \dots & r_l(0) \end{bmatrix} \begin{bmatrix} a_l(1) \\ a_l(2) \\ \vdots \\ a_l(p) \end{bmatrix} = \begin{bmatrix} r_l(1) \\ r_l(2) \\ \vdots \\ r_l(p) \end{bmatrix}, \tag{2.16}$$

where p is the order of the linear prediction coefficients.

Rabiner et al. (1993) explained how the Durbin algorithm can be applied to solve (2.16). Furthermore, Beigi (2011) provided a pseudo-code for solving (2.16) to calculate the predictor coefficients (a_l). Once the predictor coefficients are calculated, the following recursion is used to extract the cepstral coefficients (LPCC)

$$LPCC_r = a_r + \sum_{j=1}^{r-1} \left(\frac{j}{r}\right) LPCC_j \; a_{r-j} \; , \tag{2.17}$$

for $1 \leq r \leq R$.

In speaker and speech recognition fields, different efforts have been put to improve linear prediction based feature extraction, like LPCC, as will be now discussed. For example, samples selection for linear prediction (LP) analysis of voiced speech (like vowels) was presented in Ma et al. (1993). A weighted linear prediction framework was proposed where speech samples were selectively weighted based on their match to the speech production model. The method emphasised the contribution of high amplitude samples that are assumed to be less likely affected by noise. The work reported better accuracy in the estimation of the LPC coefficients obtained by the weighted analysis than the accuracy of the ones obtained by the conventional analysis.

An orthogonal framework was presented in Hu (1998) for robust LP analysis. It facilitated the use of a number of error minimisation criteria and it included a weighting as a function of the prediction residual. It was noted that for voiced speech, the prediction residual often comprises of impulsive innovations and random noise. In comparison to analysis criteria that focus on either of those types of residuals, the weighting function adapted the proposed framework for both and it was found to be successful in that regard.

A general formulation of weighed LP methods was introduced in Pohjalainen & Alku (2013). Various temporal weighting functions were included for the optimisation of the all-pole filter coefficients. The work addressed the problem of having a speech spectrum corrupted by effects, like noise which was tackled by proposing a generic spectrum analysis framework which can be adjusted in relation to the corruptions encountered. It was oriented

around speech-based emotion recognition as a classification problem and the methodology outperformed standard LP in that task.

For speaker verification under noisy conditions, Hanilci et al. (2012) investigated the robustness of speech features extracted with spectral estimates based on regularised linear prediction. The work included regularisation of some of the weighted linear prediction methods. For speaker verification under factory and babble noise, regularised weighted linear prediction outperformed conventional and weighted linear prediction methods. However, it provided similar performance for non-noisy speech. Regularisation was assumed to reduce the mismatch between training and test data by providing smooth spectral estimates. The idea was motivated by the regularised linear prediction presented earlier in Ekman et al. (2008).

In the work of Ekman et al. (2008), regularised LP was presented as a parametric spectral modelling method. The methodology tackled the problem of over-sharpening of the formants by penalising rapid changes in the spectral envelop of high-pitch speakers. High pitch frequencies cause standard LP envelop estimation to fail in separating the short-term dependency (the envelop) from the log-term dependency (the pitch), resulting in an envelop estimate that is contaminated with harmonics. The regularisation is based on the inclusion of a penalty measure that increases as the spectral envelop gets more peaky. It was shown that regularised LP provided a smoother spectral envelop than the conventional LP method.

The use of higher-lag autocorrelation coefficients in the autoregressive model was introduced and investigated in Shannon & Paliwal (2006). The method was based on the fact that the autocorrelation function of white random noise is zero everywhere except for zero time lag. Such autocorrelation function values are confined to low-time lags for broadband noise and are very small for higher time lags. The extracted speech features provided higher accuracy for noisy speech recognition.

From the same perspective of Shannon & Paliwal (2006), Alku & Saeidi (2017) used linear predictive spectral estimates based on higher-lag autocorrelation coefficients to extract robust speech features for noise-robust speaker verification. They further introduced a combined higher-lag linear prediction which takes advantage of both zero-lag and higher-lag predictions. The methodology provided the same performance compared to conventional LP for clean speech and better performance in the presence of additive noise.

In summary, the works reviewed above aimed to have LP based spectral estimates that are smooth and robust to noise. A simple method that will be presented here can address those issues together. According to the Wiener-Khinchin theorem, the autocorrelation function can be determined by taking the inverse Fourier transform of an FFT spectral estimate. Using that criterion, one can avoid the use of autocorrelation function estimates which at low-time

lags can be affected by the presence of noise as addressed in Shannon & Paliwal (2006). Also, if the FFT spectral estimate can be smoothed in some way, sharp peaks in the spectrum will be implicitly penalised which was the issue addressed in Ekman et al. (2008).

2.1.2 Spatial Features

In the field of speaker diarization, the speakers normally exist in the same enclosed space (e.g a meeting room) and conversations are sometimes recorded with multiple distant microphones (MDM). A speaker location in the spatial space is a useful property which was utilised for speaker segmentation in Ellis & Liu (2004). Locations of the speakers can be estimated by measuring the difference in the time of arrival of the speech signal at pairs of available microphones, see Fig. 2.3. These measurements, commonly known as Time Delay of Arrival (TDOA), were used as spatial features for speaker diarization alone in Pardo et al. (2006) and combined with acoustic features in Pardo et al. (2007) . They are extracted using the Generalised Cross Correlation with Phase Transform~~ation~~ (GCC-PHAT or GPHAT) algorithm Knapp & Carter (1976) which computes the normalised cross correlation between two signals in the frequency domain.

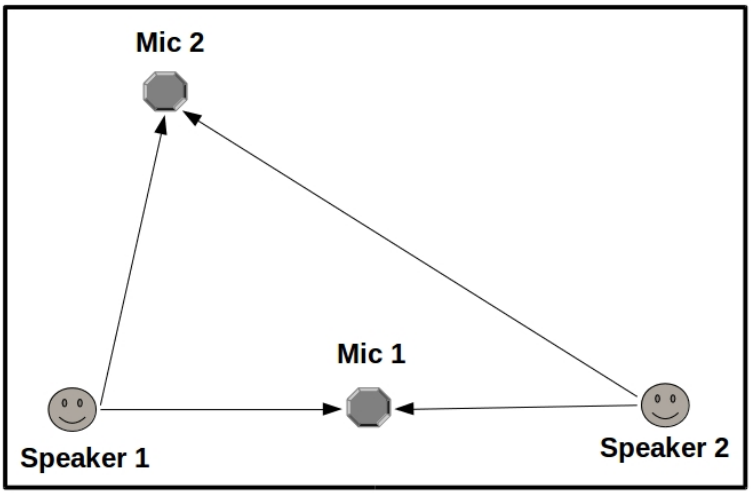


Fig. 2.3 A meeting room layout illustrating how delay features can indicate speakers' location which is helpful in speaker diarization. One can observe that the two speakers have noticeably different distances to Mic 2. When calculating the delays with the reference to a specific microphone, let it be Mic 1, the delay in speech signals arrival to Mic 2 would vary depending on speakers locations.

Denote the Fourier transforms of the speech signal arriving at microphones i and j by $s_{f,i}$ and $s_{f,j}$, respectively, then TDOA features are determined as follows

$$\text{GPHAT}_{i,j}(f) = \frac{s_{f,i}[s_{f,j}]^*}{|s_{f,i}[s_{f,j}]^*|} \tag{2.18}$$

$$\text{TDOA}_{i,j} = d(i,j) = \arg\max_d(R_{\text{PHAT}}(d)), \tag{2.19}$$

where $d(i,j)$ is the delay between channels i and j, $[\,]^*$ denotes the complex conjugate and $R_{\text{PHAT}}(d)$ is the inverse Fourier transform of $\text{GPHAT}_{i,j}(f)$. In (2.18). the complex spectrum of one of the signals is multiplied by the conjugate of the other signal. This corresponds to correlation in the time domain. The whitening function, $1/|s_{f,i}[s_{f,j}]^*|$, normalises the numerator of (2.18) so that the correlation peak is not confused with frequency components of high magnitudes. After taking the inverse Fourier transform, in (2.19), the peak in the time domain indicates the delay between the two signals.

TDOA features are calculated in segments of the speech signal for some segment rate. The size of segment should not be too large as otherwise it will degrade the resolution of the estimation or if it is too small, it will affect the robustness of the features Anguera et al. (2007). These features are usually estimated in segments of 250 ms Vijayasenan et al. (2011b) while the segment rate may vary.

Delays (TDOA features) are usually estimated between a microphone selected to be the reference microphone and the rest of the microphones. Alternatively, the channel (microphone) with the highest SNR is sometimes used as the reference one Pardo et al. (2007). However, the central microphone is more often used as a reference. The central microphone is selected as the one that has the maximum average cross-correlation with the rest of the microphones Anguera et al. (2007).

Vijayasenan & Valente (2012) proposed the extraction of high dimensional TDOA features where the delays were estimated between every possible pair of microphones such that no information was missed. This method was found to be superior to estimating the delays in relation to the reference channel only. TDOA features were also estimated from microphone pairs selected with methods that are based on dynamic margin and cross correlation to name a few González et al. (2012); Martínez-González et al. (2017). Those methods also provided better performance, compared to TDOA features estimated in reference to a single channel; in addition, to a lower computational complexity compared to using high dimensional TDOA features. The work of Martínez-González et al. (2017) also proposed the estimation of high dimensional TDOA features in combination with PCA to reduce the dimensionality. The use

of PCA was necessary because in cases of high number of microphones, for example 16, the feature dimension will be 120 when the delays are estimated between all possible pairs of microphones.

A different prospective was taken by Anguera et al. (2007), who introduced an improvement in the estimation of TDOA features. Instead of choosing the TDOA as the maximum value of $R_{\text{PHAT}}(d)$ as in (2.19), a number of maximum values are kept for each segment. Then the most reliable delay value for each segment is selected in two steps of the Viterbi decoding algorithm. That should let the estimated values of TDOA to follow the talking speaker and not to be disturbed by sudden noisy events like a door closing.

Another spatial feature that can be estimated when multiple microphones are available is the Direction of Arrival (DOA) of the speech signal Brandstein & Silverman (1997). These features were also used in speaker diarization as in Koh et al. (2008). For a microphone array, DOA features can be obtained by mapping the TDOA features after taking the array geometry into account Dmochowski et al. (2007). The estimation of the DOA degrades as the effect of noise and reverberation increases as addressed and tackled in Dmochowski et al. (2007); Evers et al. (2017).

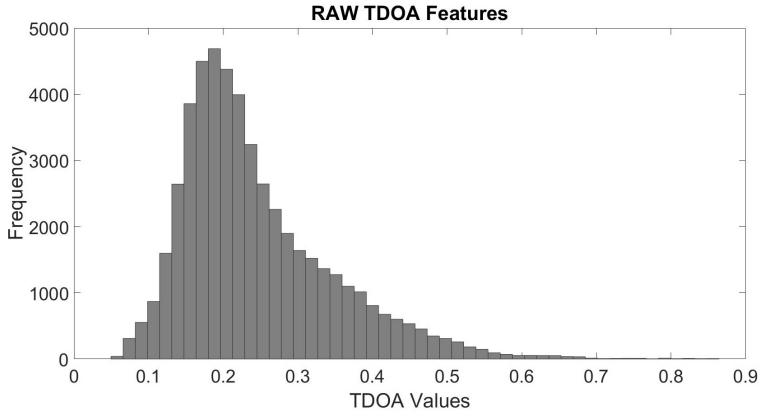


Fig. 2.4 Distribution of raw TDOA features.

TDOA features are more popular and are relatively quick to estimate. The improvement in their extraction that was introduced in Anguera et al. (2007) also addressed the reverberation problem. The work in this thesis addresses an issue of TDOA features that comes after their estimation. During an inference in the Binary Keys diarization system conducted here, it was found that these features have a skewed distribution, Fig. 2.4. This indicates that they may not be appropriately modelled by a Gaussian model or a Gaussian Mixture Model (GMM) as in Martínez-González et al. (2017); Pardo et al. (2007). Part of this work will focus on

identifying a suitable normalisation of the distribution of these features and the impact on the binary key based diarization approach.

2.1.3 Feature Fusion

As mentioned earlier, using more than one type of feature in a recognition system can improve the performance. The reason is that when some aspect of information about the speaker is not appropriately or completely covered by a particular feature type, it can be captured by the other feature(s). In most cases multiple features are fused at feature level (commonly by feature concatenation) or at the score level.

- **Feature-level fusion**

  Features extracted from speech signals at the same frame-rate can be concatenated and used in speaker recognition systems as if they were one feature. MFCC and LPCC features were concatenated to improve speaker identification in Omar & El-Hawary (2017). In Zeinali et al. (2017a), i-vectors were calculated for text-dependent speaker verification using a concatenation of MFCC and bottleneck features. Bottleneck features is the term used to describe features that are extracted from acoustic features using a DNN as first done by Yu & Deng (2014).

- **Score-level fusion**

  In other cases, separate systems have dealt with a combination of different features independently and the recognition scores were then fused. This method can present better performance compared to feature-level fusion but it increases the overall system complexity because it comprises as many sub-systems as the number of feature types. Bottleneck, short-term spectral and modulation spectral features were fused at the score-level for speaker verification in Sarria-Paja & Falk (2018). MFCC and TDOA features were also fused at score-level for speaker diarization by Martínez-González et al. (2017).

Using either of these fusion criteria come at the cost of increasing the number of computations. Score-level fusion is particularly rigid in the sense that an increase in the number of computations is not possible to avoid because of the need to establish more than one system. For feature concatenation, the possible gain in the accuracy obtained by combining different features may not be as considerable in relation to the increased complexity as a result of the growth in dimensionality.

The increase in dimensionality can also present other issues. For example, if the features of interest are used to fit to a Gaussian Mixture Model (GMM), the growth in dimensionality causes the required amount of data to increase exponentially for reliable density estimates of the GMM, see Kinnunen & Li (2010). Dimensionality reduction techniques, such as Principal Component Analysis (PCA), can be used to overcome such shortcomings. When PCA is performed on concatenated features it can be regarded as feature-level fusion since the principal components can be seen as a linear combination of the input features, see e.g Jolliffe (2002). In Chibelushi et al. (1997), PCA was used to fuse audio and visual information in speaker identification. MFCC features were used for the audio information while outer lip-margin features represented the visual aspect. Lee & Narayanan (2005) used covariance matrix-based PCA to fuse a large number (up to 15) of different features in emotion recognition from spoken dialogues.

For speaker verification using Support Vector Machines (SVM)s, Kajarekar (2005) used covariance PCA to reduce the dimensionality of polynomial coefficients (of 11479 dimensions) which were a transformation of MFCC features. In synthetic speech detection, Wu et al. (2013) used PCA to fuse phase modulation features and phase features as well as phase and MFCC features. However, despite that it was conducted on a combination of different features, PCA was viewed as a dimensionality reduction tool and the work did not report results that might help in distinguishing the effect of PCA. In speech recognition using Deep Neural Networks (DNN)s as a feature extractor, a bottleneck layer (a narrow hidden layer) is placed in the middle of a network trained for phoneme classification, then bottleneck features are extracted from that layer (Zhang et al., 2014). However, it was stated that such a narrow layer degrades the efficiency of the DNN training. Among other modifications, a relatively large bottleneck layer was incorporated and the dimensionality was reduced using PCA. This was then found to outperform the conventional bottleneck features for speech recognition.

In speaker identification using a probabilistic neural network, Ahmad et al. (2015) used PCA to reduce the dimensionality of MFCC features separately for each speaker and it was based on the covariance matrix. McLaren & Lei (2015), introduced 2D-DCT coefficients as speech features and used PCA to reduce their dimension in a speaker verification system. The system was based on the i-vectors of Dehak et al. (2011) (to be described shortly) and the lower dimensional features showed improvements over the original features in a number of cases. In Liu et al. (2015), outputs from hidden layers of various network models were used to provide high dimensional deep features for text-dependent speaker verification in a number of systems. That work used PCA to reduce the dimensionality of those deep features

and also used a concatenation of the reduced dimensional deep features and spectral features. The work did not report results where original deep features were used, hence, the effect of using PCA on the performance cannot be inferred. In i-vector based speaker verification with whispered and normal speech, Sarria-Paja et al. (2016) used PCA for the fusion of MFCC and Weighted Instantaneous Frequencies (WIF) features; but it also did not report results without PCA fusion.

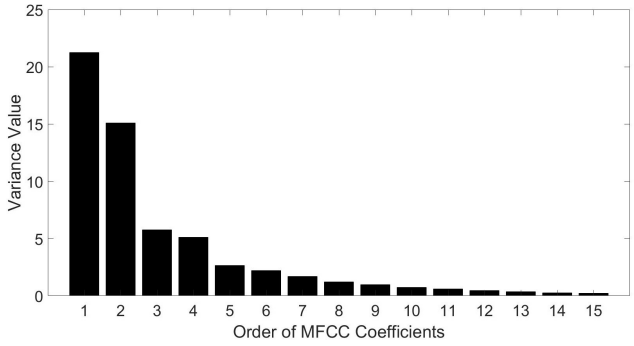
As the principal components are orthogonal, the new attributes of speech are uncorrelated which is important for GMMs with diagonal covariance matrices. Another advantage of PCA-based feature fusion is the reduction in system complexity as an effect of the reduced feature dimensionality. However, performing PCA separately for each speech sample, for example in (Ahmad et al., 2015), in the training or testing phase is undesirable: it adds another level of computations to the system. In addition, it will result in having speech features in different spaces which can be more appropriate for individual speaker modelling Kwok et al. (2004).

An alternative methodology works by defining one set of 'universal' principal components such that the analysis of PCA is performed once and all speech samples' features are projected to a unified reduced dimensional space as in (Sarria-Paja et al., 2016). This method of defining global principal components was proposed for dimensionality reduction in speaker identification using the Gaussian Mixture Model-Universal Background Model (GMM-UBM) recognition system in Seo et al. (2009). A global covariance matrix was estimated from the features of speech samples for a relatively large number of speakers. This method was found to outperform concatenation of features by Sarkar et al. (2014), where it was used to combine cepstral features and phonetically discriminant features for speaker verification. A similar technique in Zhang et al. (2016) also used global covariance PCA for feature fusion in an i-vector system.

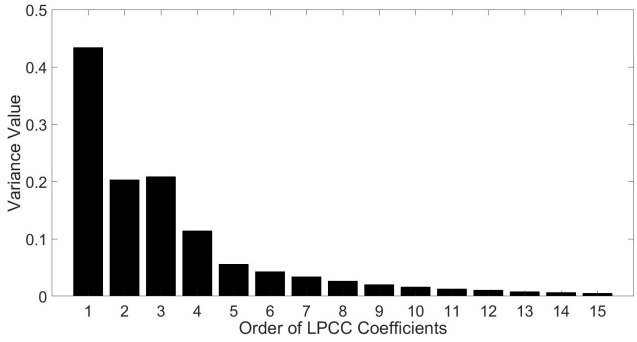
From the works reviewed above, one can notice that PCA is not carefully tuned when used to describe speech features. For example, many works did not mention the technique used to extract the principal components. It is therefore assumed to be the classical eigen-decomposition or the singular value decomposition (SVD). Except for a few works which mentioned that PCA was based on the covariance matrix, others did not provide information regarding this aspect which implies that such a factor was not deemed important.

In So & Paliwal (2008), it was reported that the variances of MFCC coefficients largely differ from each other. Fig. 2.5 includes the plots of the variances of different orders of the static cepstral coefficients of MFCC as well as LPCC. The training samples of the NIST 2002 SRE telephone data (Martin & Mark, 2004) were used to produce these plots. The relatively

high differences in the coefficients' variances strongly suggests that the analysis for PCA must be based on the correlation matrix; or equivalently a normalisation of variances must be made beforehand. Otherwise, the attributes with higher variances will dominate the first few principal components (Rencher, 1992). These high variance features may not have superior importance over the others. For example, lower order coefficients of MFCC are considered to be more sensitive to undesirable effects caused by factors such as the transmission channel So & Paliwal (2008), yet they have relatively high variances.



(a) MFCC



(b) LPCC

Fig. 2.5 Variances of MFCC and LPCC cepstral coefficients.

This work will extract the principal components from speech features of many speakers' utterances. This is time saving as the analysis for PCA will be performed once. It is also more compatible with the concept of a universal background model (to be described shortly in Section 2.2.1) where all utterances are transformed by the same global principal components. Hence, PCA will be based on a global correlation matrix (or a global covariance matrix). In such a case, there will be a plethora of feature vectors and they must not be allowed to equally contribute in the analysis that will eventually produce the principal components. That is because there might be outliers or underlying noise variance which PCA cannot distinguish from the variance of the features Bailey (2012); Delchambre (2014). This suggests the

necessity of adopting PCA methods that are more robust than the traditional techniques as will be discussed now.

For dimensionality reduction or feature fusion, robust methods of PCA do not appear to have attracted much attention in the literature. A study in the field of voice conversion, (Rao et al., 2016), used robust PCA of (Hubert et al., 2005) to remove the effect of outliers when performing dimensionality reduction. The method combines projection pursuit techniques (Jones & Sibson, 1987) and minimum covariance determinant estimators (Rousseeuw, 1984). That robust PCA starts by obtaining the projections of the original feature vectors on a particular number of principal components extracted using the SVD technique. Then a subset of the new feature vectors are selected and the determinant of their overall covariance matrix is determined. For various subsets of the new feature vectors, an iterative process is followed to find the subset with the lowest determinant of its covariance matrix which is a slow process as reported in (Hubert et al., 2005). A similar robust PCA approach based on minimum determinant covariance estimators was used in (Lee et al., 2002) for robustness to outliers in feature dimensionality reduction for speaker identification.

These methods are not suitable for extracting global principal components for two reasons. First, the projection pursuit technique is clearly meant to optimise the extraction of the principal components using the projections of the feature vectors on those components. The framework of PCA in this work requires the principal components to be retained before any projection is made. Second, they are based on iterative estimations of covariance matrix determinants which was reported to be slow and is therefore not feasible to deal with a large number of feature vectors.

Additionally, in comparison to classical eigendecomposition and SVD, there are alternative methods that can retain the principal components more precisely and efficiently. This was indicated in Roweis (1998) where an Expectation-Maximisation (EM) based method was introduced for PCA. A number of extensions of the EM method were introduced for PCA in the case of noisy or missing data, see e.g Bailey (2012). In Delchambre (2014), a power iteration method was introduced as an improvement over the EM algorithm of Bailey (2012), and it was faster and superior in finding the principal components in the order of variance they represent. However, this method suffers from a low convergence rate under particular conditions as shown by Delchambre (2014).

To tackle those observations about PCA, a weighted correlation PCA is introduced in this work where the principal components are iteratively estimated using a Recurrent Neural Network (RNN). One of the earliest works that used neural networks for PCA can be found in Oja (1982), where a class of unconstrained Hebbian-type learning rules were derived for this

purpose. In that work, the dominant eigenvector is directly estimated from the input sequence not from a correlation or covariance matrix. In Oja (1992), a Stochastic Gradient Ascent (SGA) neural network was proposed to extract the less dominant eigenvectors. Since these methods, of Oja (1982) and Oja (1992), can only retain the principal components directly from a sequence of feature vectors, it is not straight forward to modify them to deal with a weighted correlation matrix (or weighted covariance matrix).

In Rajasekaran & Pai (2002), a Recurrent Neural Network (RNN) was introduced to find the largest eigenvalue and the associated eigenvector of a real symmetric matrix. Yi et al. (2004) proposed a similar method to additionally find the smallest eigenvalue and the associated eigenvector. This latter work also provided a comprehensive analysis of the dynamic behaviours of the RNN model which justified its use in the solution of the eigendecomposition problem. The framework of Rajasekaran & Pai (2002) is only constrained by the condition that the matrix to be decomposed must be real and symmetric. Therefore, it is extended here to solve the eigendecomposition problem of a weighted correlation matrix (and a weighted covariance matrix) and retain the entire set of principal components.

2.2 Speaker Modelling and Verification

The past decade and a half witnessed the development of sophisticated modelling techniques in the field of speaker recognition. The achievements in this regard have helped in developing successful speaker recognition systems. This section describes speaker modelling, the i-vector speaker verification system and its development.

2.2.1 Overview

Recall that short-term spectral features are estimated from short speech frames usually at a frame rate of 10 ms as in (Dehak et al., 2011). This results in having a hundred feature vectors for only one second of speech. The simplest method to assess the similarity between two utterances could be to determine, for example, the Euclidean distance between each feature vector of one utterance and all the feature vectors of the other utterance. Feature modelling (i.e. speaker modelling) was first focused on having the feature vectors of a speakers' utterances represented by a lower number of vectors. Template models were first used in this case to build a speaker model from enrolment (reference) utterances. Vector quantization was one of the approaches used to build a template model of the speaker see e.g Soong et al. (1985). To test if the utterance of an unknown speaker matches the template

model of a known speaker, deterministic measures like Euclidean distance or Mahalanobis distance were used see e.g Campbell (1997).

Later, stochastic models were used for speaker modelling such as Gaussian Mixture Models (GMM) introduced for speaker identification by Reynolds & Rose (1995). The enrolment feature vectors were fitted to a GMM. Probabilistic measures, usually the log-likelihood value, were used in this case to determine if a test utterance matches a speaker's GMM model. According to the enrolment (training) criterion, template and stochastic models are seen as generative models since they characterise the distribution of speech features. Artificial Neural Networks (ANN) (e.g Yegnanarayana & Kishore (2002)) and Support Vector Machines (SVM) (e.g Campbell et al. (2006a)), have also been used for speaker recognition except by modelling the boundaries between speakers and so they are regarded as discriminative models (see e.g Kinnunen & Li (2010)).

A suitable speaker model is important to have an efficient speaker recognition system. The following few paragraphs may briefly summarise the main issues addressed in the literature towards meeting that goal.

Instead of modelling each speaker independently, Reynolds et al. (2000) proposed coupled speaker modelling for speaker verification. In that method, a GMM was fitted to feature vectors of a relatively large number of speakers and is referred to as the Universal Background Model (UBM) usually abbreviated to GMM-UBM. The GMM-UBM represents a broad acoustic space of speech sounds. A particular speaker GMM is then obtained by adapting the GMM-UBM parameters to that speaker's enrolment feature vectors using Maximum a Posteriori (MAP) optimisation. Hence, the speaker's GMM is assumed to retain the GMM-UBM acoustics for speech sounds not seen in the speaker's feature vectors. This is important because it helps to indirectly address the issue that a reliable speaker model requires collecting as much speech as possible from individual speakers. Adaptive vector quantization presented in Zhou & Mikhael (2006) is another form of coupled modelling that was used for speaker identification. It presented better performance than conventional (not coupled) vector quantization based identification. Nonetheless, GMM based models have attracted more attention and have been adopted in further developments in speaker modelling.

It is feasible to have a simple model of the speaker so that the use of a number of techniques can be facilitated. The simplest speaker model is probably the one presented in Markel et al. (1977) which is a vector obtained by time-averaging an utterance's feature vectors. However, it provided poor recognition performance. The speaker GMM, on the other hand, presents a good performance but it comprises a mean vector, a covariance matrix and a weight for each mixture component. In contrast, another robust and relatively simple

model called the supervector can be formed by stacking the means of the speaker's GMMs. This relatively high dimensional vector was introduced in Campbell et al. (2006b) where it enabled effective usage of the SVM for speaker verification. Supervectors were widely used in speaker recognition as in the supervector-based SVM classifier presented for age and gender recognition in Li et al. (2013). Binary keys by Anguera & Bonastre (2010), identity vectors (i-vectors) from Dehak et al. (2011) and x-vectors proposed by Snyder et al. (2017) and more are other developments of simple vector models.

In practice, it is very likely that the enrolment and detection utterances are recorded over different channels. This channel mismatch negatively affects the recognition performance Beigi (2011). To address this problem, Kenny (2006) presented a theory and proposed Joint Factor Analysis (JFA) based algorithms to model channel variability in addition to speaker variability and considered the supervectors as speakers' models. The development that followed in Kenny et al. (2007) namely, eigenchannels, reduced the computational resources required to perform the modelling. Further studies in Kenny et al. (2008) and Dehak (2009) eventually led to the introduction of the i-vectors in Dehak et al. (2011).

The i-vector is a simple low dimensional representation of a speaker's utterance which enabled the use of a number of techniques such as Linear Discriminant Analysis (LDA). The estimation of i-vectors, which will be described shortly, accounts for speaker-and-session variability and coupled modelling, using a GMM-UBM, is a fundamental element of the process.

In Dehak et al. (2011), the similarity between the i-vectors was determined using the cosine similarity metric or an SVM classifier. Later, a Probabilistic LDA (PLDA) model (presented in Garcia-Romero & Espy-Wilson (2011)) became the standard scoring criterion. In the recent literature, i-vector based speaker recognition systems are found to provide state-of-the-art performance in many related applications and further enhancements have been presented mostly for the i-vector/PLDA framework as discussed now.

Kenny et al. (2013) addressed the issue that i-vectors extracted from long utterances are more reliable than those extracted from shorter utterances. Considering an arbitrary utterance length, a methodology was proposed to quantify this uncertainty by propagating it to the PLDA model. Rajan et al. (2014) presented the idea of using the average of multiple i-vectors to enrol a speaker. Novoselov et al. (2015) used a Deep Neural Network to estimate two different nonlinear PLDA models that outperformed the linear PLDA model (introduced in Garcia-Romero & Espy-Wilson (2011)), especially, when both of the nonlinear PLDA models were combined.

Kheder et al. (2016) introduced a joint probabilistic model of short and long utterances i-vectors. The Stereo Stochastic Mapping (SSM) algorithm was used to map short utterances i-vectors to, supposedly, their long utterance i-vectors which provided noticeable improvement for short utterances. Cumani & Laface (2017) proposed a non-linear transformation of the i-vectors to normalise their distribution as assumed by the PLDA model (as will be explained shortly). Most recently, Khosravani & Homayounpour (2018) proposed a non-parametric training of the PLDA model.

Speech modelling using i-vectors have also found use in many speech processing related tasks see Table 2.2. Pal & Saha (2017) proposed a new voice conversion (VC) method using i-vectors. Zeinali et al. (2017b) presented a state-of-the-art i-vector based approach for text-dependent speaker verification. Safavi et al. (2018) studied speaker recognition, gender and age-group classification of children for a number of systems where the i-vector system presented the best recognition performance.

Task	Example
Text-Independent Speaker Verification	Dehak et al. (2011)
Language Identification	Song et al. (2013)
Voice Conversion	Pal & Saha (2017)
Text-Dependent Speaker Verification	Zeinali et al. (2017b)
Age and Gender Classification	Safavi et al. (2018)

Table 2.2 A number of different speech processing related tasks that use the i-vector modelling.

In order to establish the i-vector system, a set of development data is required. Development data is a large amount of speakers' utterances recorded over different channels and are used to learn model parameters. Regarding the evaluation set, a speaker's utterance(s) that is used as a reference is called the enrolment utterance and the one used at the recognition (verification, etc.) phase is called the test utterance. These utterances are not used in the system establishment (development).

The i-vector speaker recognition system has become a standard in speaker recognition and it is used here in this work to evaluate the performance of the proposed methodologies for acoustic feature extraction and fusion. The x-vectors recently introduced in Snyder et al. (2017) are similar to the i-vectors and their performance was better for short utterances but comparable for long utterances. x-vectors are Deep Neural Network (DNN) embeddings and they also represent variable length speech samples by a fixed length vector. However, they do

not include channel variability modelling and also there is a requirement for a relatively large amount of development data as will be clarified at the end of the following section (2.2.2).

2.2.2 Development of the i-vector Based Verification System

The Joint Factor Analysis (JFA) model presented by Kenny (2006) is expressed, according to the definition of Rubin & Thayer (1982), as

$$\mathbf{m}_u = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}, \tag{2.20}$$

where $\mathbf{m}_u$ is a speaker-and-channel dependent supervector of a particular utterance (u) comprised of the components of speaker and channel subspaces combined. These components are: $\mathbf{m}$ is a global speaker-and-channel independent supervector (the Universal Background Model (UBM) supervector); $\mathbf{V}$ and $\mathbf{D}$ define the speaker subspace, where $\mathbf{V}$ is the eigenvoice matrix and $\mathbf{D}$ is a diagonal residual term which represents inter-speaker variability not captured in $\mathbf{V}$, and $\mathbf{U}$ defines a session subspace (eigenchannel matrix). The vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are the speaker-and-channel dependent factors in their respective subspaces; each is assumed to be a normally distributed random variable.

It was observed by Dehak (2009) that the channel factors in (2.20) which are only expected to model channel effects also contain information about the speaker. That motivated the definition of the total variability space which simultaneously contains speaker and channel variabilities. Hence, the Joint Factor Analysis (JFA) of (2.20) became a simple factor analysis expressed as (Dehak et al., 2011)

$$\mathbf{m}_u = \mathbf{m} + \mathbf{T}\mathbf{w}_u, \tag{2.21}$$

where $\mathbf{T}$ is a rectangular low-rank total variability matrix of the eigenvectors with the highest eigenvalues of total variability covariance matrix and $\mathbf{w}_u$ is the i-vector.

In order to extract the i-vectors, the system initially requires the estimation of a GMM-UBM (Λ) mainly to obtain the supervector $\mathbf{m}$ of (2.21). This speaker-and-channel independent supervector is obtained by concatenating the means of all mixture components C of Λ. For a set of feature vectors, $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_L)$, the parameters of Λ (means, covariance matrices and weights) are estimated using the Expectation-Maximisation (EM) algorithm see e.g Reynolds & Rose (1995). $\mathbf{Y}$ can vectors set of, for example, MFCC features. Each component $\mathscr{G}_c$, where $1 \leq c \leq C$ and C is the total number of mixture components, has an associated probability that is expressed by the following multivariate Gaussian density

function

$$\mathscr{G}_c(\mathbf{y}_l) = \frac{w_c}{(2\pi)^{D/2}|\mathbf{\Sigma}_c|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}_l - \boldsymbol{\mu}_c)'\mathbf{\Sigma}_c{}^{-1}(\mathbf{y}_l - \boldsymbol{\mu}_c)\right], \tag{2.22}$$

where w_c, $\boldsymbol{\mu}_c$ and $\mathbf{\Sigma}_c$ are the weight, mean and covariance matrix of component c, respectively. Note that this is a D variate function, where D is the feature dimension.

The speaker-and-channel dependent supervector $\mathbf{m}_u$ of (2.21) is an 'adaptation' of $\mathbf{m}$ for the utterance feature vectors of a particular session of a particular speaker. In the factor analysis model of the i-vector, a statistical alignment for an utterance's feature vectors is made by estimating the Baum-Welch[1] statistics instead of a supervector adaptation with *Maximum a Posteriori* (MAP) estimation which was suggested by Kenny et al. (2004). This alignment is the posterior probability of a mixture component c for the feature vector $\mathbf{y}_t$. It provides latent information on how the feature vectors react to each mixture component. In i-vector based systems, the alignment is met by the determination of the Baum-Welch statistics for the utterance feature vectors given the GMM, Λ see Dehak et al. (2011). For component c of Λ and utterance feature vectors for T frames, with $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T)$, the 0^{th} order Baum-Welch statistics are calculated as follows (Dehak et al., 2011)

$$\mathbf{n}_c = \sum_{t=1}^{T} \mathscr{P}(c|\mathbf{y}_t, \Lambda), \tag{2.23}$$

and the 1^{st} order Baum-Welch statistics are calculated as

$$\mathbf{f}_c = \sum_{t=1}^{T} \mathscr{P}(c|\mathbf{y}_t, \Lambda)\mathbf{y}_t, \tag{2.24}$$

where $\mathscr{P}(c|\mathbf{y}_t, \Lambda)$ is the posterior probability of the mixture component c generating the feature vector $\mathbf{y}_t$.

Now the extraction of the i-vector requires the 1^{st} order Baum-Welch statistics of a speech utterance u centralised based on Λ, such that

$$\tilde{\mathbf{f}}_c = \sum_{t=1}^{T} \mathscr{P}(c|\mathbf{y}_t, \Lambda)(\mathbf{y}_t - \boldsymbol{\mu}_c), \tag{2.25}$$

where $\boldsymbol{\mu}_c$ is the mean of mixture component c of the GMM-UBM (Λ). It can be noticed from (2.25) that the size of $\tilde{\mathbf{f}}_c$ depends on the feature dimensionality. The supervector $\tilde{\mathbf{f}}_u$ of a speech utterance u has the size $C \times D$ as it is a concatenation of the centralised 1^{st} order

[1] Refer to Appendix A.1 for the definition of Baum-Welch Statistics.

Baum-Welch statistics of the feature vectors of the speech utterance for all the components $1 \leq c \leq C$.

The total variability subspace $\mathbf{T}$ is learned from the supervectors of many development utterances using the EM algorithm. The process is the same as the one used for learning the eigenvoice matrix as with Kenny et al. (2005) except that all the supervectors are pooled together without speakers' labels. The complexity of the process depends on the size and number of the supervectors. The resultant $\mathbf{T}$ is a low-rank matrix where the number of rows corresponds to the number of total factors (the dimension of the i-vector) and the number of columns is equal to the size of the supervectors.

The extraction of the i-vector, $\mathbf{w}_u$, is based on the computation of the posterior distribution of a speaker's supervector, $\mathbf{m}_u$ of (2.21), which was achieved by calculating Baum-Welch statistics for the feature vectors of the speaker see Dehak (2009). This posterior distribution is assumed to be Gaussian, its mean, also the latent variable in the JFA model of (2.21), is determined as[1]

$$\mathbf{w}_u = (\mathbf{I} + \mathbf{T}^T \mathbf{\Sigma}^{-1} \mathbf{N}_u \mathbf{T})^{-1} . \mathbf{T}^T \mathbf{\Sigma}^{-1} \tilde{\mathbf{f}}_u, \tag{2.26}$$

where $(\mathbf{I} + \mathbf{T}^T \mathbf{\Sigma}^{-1} \mathbf{N}_u \mathbf{T})^{-1}$ is the covariance matrix of the i-vector ($\mathbf{w}_u$). $\mathbf{I}$ is an identity matrix with the size of the total factors of $\mathbf{T}$, $\mathbf{N}_u$ is a diagonal matrix of dimension $C \times D$ by $C \times D$ with diagonal blocks $\mathbf{n}_c \mathbf{I}$, and $c = 1, 2, ..., C$. $\tilde{\mathbf{f}}_u$ is a supervector of dimension $C \times D$ by 1 obtained by concatenating all 1^{st} order Baum-Welch statistics ($\tilde{\mathbf{f}}_c$). The residual variability not captured in $\mathbf{T}$ is modelled by the diagonal covariance $\mathbf{\Sigma}$ of size $C \times D$ by $C \times D$. The fixed length of the i-vectors enabled the application of LDA to minimise within class variance caused by channel effects as well as reducing the i-vectors' dimensionality, typically from 400 to 150. The development data required to perform LDA must contain more than one utterance for each speaker.

The Probabilistic LDA (PLDA) model was first presented in Prince & Elder (2007) in order to address the problem of different pose and lighting of test and enrolment data in face recognition, thus it assumes the data is resulting from a generative model which incorporates within and between class variance. It was later introduced for speaker recognition to perform i-vector scoring by Garcia-Romero & Espy-Wilson (2011). According to the PLDA generative model, the i-vector of a speaker utterance u over a particular channel can be decomposed with

$$\mathbf{w}_u = \mathbf{w} + \mathbf{\Phi}\boldsymbol{\beta} + \mathbf{\Gamma}\boldsymbol{\alpha}_u + \mathscr{E}_u \tag{2.27}$$

[1] The derivation of (2.26) is provided in Appendix A.2.

where $\mathbf{w}+\mathbf{\Phi}\beta$ is a speaker term and, $\mathbf{\Gamma}\alpha_u+\mathscr{E}_u$ is a channel term which depends on the utterance u . These terms describe the between-speaker variability Φ (eigenvoices) and within-speaker variability $\mathbf{\Gamma}$ (eigenchannels). The statistically independent latent vectors $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_u$ have standard normal distribution. The global offset (the mean of the development i-vectors) is $\mathbf{w}$ and $\mathscr{E}_u$ is a residual term assumed to be Gaussian with zero mean and diagonal covariance. A full covariance matrix $\mathbf{\Sigma}$ of $\mathscr{E}_u$ can compensate for $\mathbf{\Gamma}\boldsymbol{\alpha}_u+\mathscr{E}_u$ as proposed in Kenny (2010), hence the PLDA generative model of (2.27) was modified to $\mathbf{w}_u=\mathbf{w}+\mathbf{\Phi}\boldsymbol{\beta}+\mathscr{E}_u$.

The PLDA model training is simple and computationally efficient, however, it assumes that the input observations (i-vectors) are Gaussian distributed. It was reported in Garcia-Romero & Espy-Wilson (2011) that Gaussian PLDA gives inferior performance compared to Heavy-Tailed PLDA Kenny (2010) unless a transformation is applied to the i-vectors, where the Radial Gaussianisation (RG) technique was used for this purpose. The model parameters $\{\mathbf{w},\mathbf{\Phi},\mathbf{\Sigma}\}$ are obtained using the EM algorithm as described in Prince & Elder (2007) with a large collection of development i-vectors that are associated with the corresponding speakers' labels.

The scoring criterion is based on the log-likelihood ratio of the same $\mathscr{H}_s$ versus different $\mathscr{H}_d$ speaker hypotheses which aims to determine if the i-vectors of two utterances (test and enrolment) belong to the same speaker or to different speakers

$$score=\log\frac{\mathscr{P}(\mathbf{w}_t,\mathbf{w}_e|\mathscr{H}_s)}{\mathscr{P}(\mathbf{w}_t|\mathscr{H}_d)\mathscr{P}(\mathbf{w}_e|\mathscr{H}_d)}, \tag{2.28}$$

where $\mathbf{w}_e$ is the i-vector of an enrolment utterance and $\mathbf{w}_t$ is the i-vector of a test utterance.

This log-likelihood ratio is easily computed in a closed-form since the marginal likelihoods (i.e., the evidence) are Gaussian. According to Garcia-Romero & Espy-Wilson (2011), each i-vector length is normalised to unity and the scores are determined as

$$\begin{aligned} score=&\log\mathscr{N}\left(\begin{bmatrix}\mathbf{w}_t\\ \mathbf{w}_e\end{bmatrix};\begin{bmatrix}\mathbf{w}\\ \mathbf{w}\end{bmatrix},\begin{bmatrix}\mathbf{\Phi}\mathbf{\Phi}^T+\mathbf{\Sigma} & \mathbf{\Phi}\mathbf{\Phi}^T\\ \mathbf{\Phi}\mathbf{\Phi}^T & \mathbf{\Phi}\mathbf{\Phi}^T+\mathbf{\Sigma}\end{bmatrix}\right)\\ &-\log\mathscr{N}\left(\begin{bmatrix}\mathbf{w}_t\\ \mathbf{w}_e\end{bmatrix};\begin{bmatrix}\mathbf{w}\\ \mathbf{w}\end{bmatrix},\begin{bmatrix}\mathbf{\Phi}\mathbf{\Phi}^T+\mathbf{\Sigma} & 0\\ 0 & \mathbf{\Phi}\mathbf{\Phi}^T+\mathbf{\Sigma}\end{bmatrix}\right). \end{aligned} \tag{2.29}$$

The development phase of the i-vector based verification system is illustrated in the diagram of Fig. 3.1 following the description provided here.

The establishment of the i-vector/PLDA framework requires large development data which can be difficult to obtain. This motivated the introduction of a suitable data augmenta-

tion method here to overcome such a problem by adding noise to copies of the available data. A similar strategy has been used before to adapt the system parameters (during development) to specific conditions related to the test utterances. It works by either of the following: incurring particular effects on the development data that are related to those ones embedded in the test utterances or using a development data that is already contaminated with such effects. It is usually referred to this strategy as multi-condition training and it has mostly been considered in training the PLDA model.

Garcia-Romero et al. (2012) proposed multi-condition training for the PLDA model where a number of effects were added to the development speech signals to match similar noise embedded in the test samples. The added effects comprised one of reverberation plus babble, car and helicopter noise. The effect of multi-condition training of PLDA has been studied in Rajan et al. (2013). It was found to be important for the system performance under noisy conditions. For the development and evaluation data, the study included ventilation, air-condition and crowd noise sources. Then it tested the cases of using the original development data with original evaluation data and noisy development data with noisy evaluation data. It was shown that the performance improved when the noise added to the development data had similar power to the noise added to the evaluation data. However, the performance was found to be degraded when the development data was contaminated with noise and original evaluation data was used in the testing.

A number of works appear to have been built on the idea of (Garcia-Romero et al., 2012). In Villalba & Lleida (2013), a mixture of channel-dependent PLDA models were trained to take into account the channel conditions of each test utterance presented at the detection phase. In Mak et al. (2016), another mixture of PLDA models was trained and the presented test speech was directed to the PLDA model that best matched the test sample's signal-to-noise ratio. The work in Martinez et al. (2014) investigated one channel feature-domain noise compensation combined with multi-condition training. A full multi-condition training approach was presented in Ribas et al. (2015) where all the development stages of the i-vector based system included various types of noise added to otherwise clean speech samples.

These aforementioned systems all sought to attempt to model different sources of background noise. Unfortunately, the positive effects in terms of improved performance that can be brought by the frameworks of those systems do not generalise to both non-noisy and noisy evaluation data as found by (Rajan et al., 2013).

The Deep Neural Network (DNN) based x-vector system requires even larger amounts of development data than the ones required for i-vector based systems. The work in Snyder et al. (2018) used data augmentation to increase the amount of development data in order to

improve the recognition performance using DNN based x-vectors over the system described by Snyder et al. (2017). Reverberation effect as well as some types of noise like bubble noise and music were randomly added to clean speech samples to produce condition-variable samples which increased the amount of data used to train the DNN. However, that method was not as helpful for the i-vector system (as reported in the same study) possibly because of the types of effects added. Also, the power of noise added is somehow arbitrary in the sense that it was not fine tuned by observing the system performance.

2.3 Speaker Diarization

The configuration of a speaker diarization system is fundamentally different and more complicated than that of a verification system. Diarization systems generally use similar speaker modelling techniques as other recognition systems. However, speaker diarization aims to determine 'who spoke when?' in an audio stream where the number of speakers is one of the desired outcomes see Tranter & Reynolds (2006).

Usually, in an unsupervised manner, a diarization system performs speech segmentation and clustering with one for each speaker. By doing that, the system automatically attributes spoken words (or their representations) to individual speakers and delivers the outcome for further processes. Those processes may include speech reconstruction from the MFCC feature vectors as in the method introduced in Milner & Shao (2007). The method estimated the fundamental frequency and voicing information from the MFCC feature vectors and used those parameters together with MFCC feature vectors to reconstruct the time-domain speech signal. This implies that it is important that a diarization system is fast enough to allow the time required for other necessary processes.

There are a number of modalities in speaker diarization which can be classified into two groups of categories. In the first group, the modalities can be categorised according to the recording method of a conversation see Fig. 2.6. In the second group, the modalities can be categorised according to the number of conversations a diarization system is concerned with see Fig. 2.7. For example, in cross-show diarization, the system looks up the existance of a speaker in more than one recording. The different diarization modalities share the same fundamentals. The work here focuses on single show diarization of meetings recorded with multiple distant microphones. Nonetheless, some of the achievements can be applied to IHM and SDM diarization modalities.

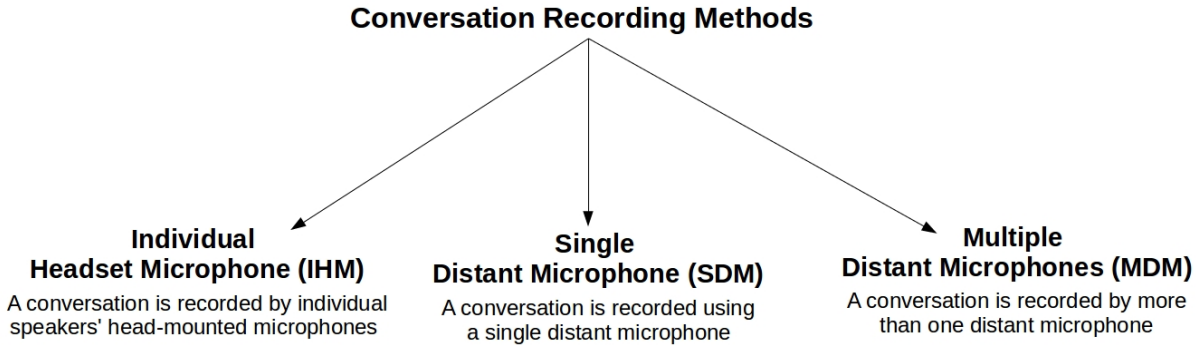


Fig. 2.6 Speaker diarization modalities categorised according to the method used to record a conversation.

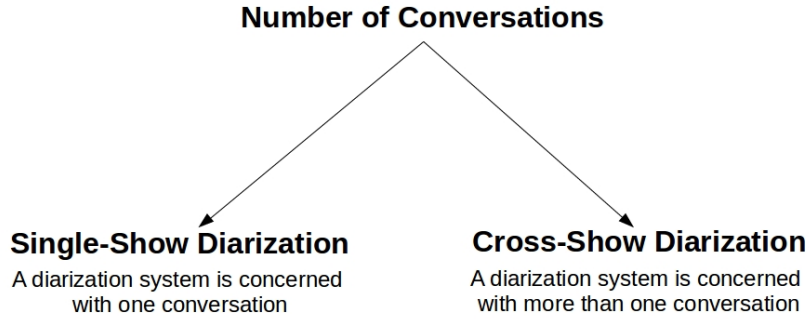


Fig. 2.7 Speaker diarization modalities categorised according to the number of conversations that the diarization system is concerned with.

The performance of diarization systems is mainly evaluated based on the Diarization Error Rate (DER) introduced in Fiscus et al. (2006), for example, as done by Martínez-González et al. (2017). DER is defined as the fraction of speaker time that is not attributed correctly to a speaker. DER comprises of Speaker Error Rate (SER), False Alarm speech E_{FA} and Missed Speech E_{MISS}. All of these measures, are usually determined, as in Anguera (2006) and also in this work, using a script named *MD-eval-v21.pl* developed by NIST see Fiscus et al. (2006). The DER is expressed as (Anguera, 2006)

$$\mathrm{DER}(\%) = \frac{1}{T_{score}} \sum_{s=1}^{S} \zeta(s)(\max(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s)), \tag{2.30}$$

where $T_{score} = \sum_{s=1}^{S} \zeta(s) N_{ref}$ is the amount of speech time scored. N_{ref} is the number of reference speakers, S is the total number of speech segments and $\zeta(s)$ is the duration of a

speech segment. Within a segment s, N_{ref} is the total number of speakers, N_{hyp} is the number of hypothesised speakers detected by the system and $N_{correct}$ is the number of speakers correctly matched between the N_{ref} and N_{hyp}.

SER is the speaker time attributed to a wrong speaker and it is determined as

$$\text{SER}(\%) = \frac{1}{T_{score}} \sum_{s=1}^{S} \zeta(s)((\min(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s)). \tag{2.31}$$

E_{FA} is the percentage of scored time that a hypothesised speaker is labelled as a non-speech in the reference. It is calculated, only over segments where the reference segment is labelled as non-speech, as in the following

$$\text{E}_{\text{FA}}(\%) = \frac{1}{T_{score}} \sum_{s=1}^{S} \zeta(s)(N_{hyp}(s) - N_{ref}(s)) \qquad \forall(N_{hyp}(s) - N_{ref}(s)) > 0. \tag{2.32}$$

E_{MISS} is percentage of scored time that a hypothesised non-speech segment correponds to a reference speaker segment. It is determined, only over segments where the hypothesis segment is labelled as non-speech, as in the following

$$\text{E}_{\text{MISS}}(\%) = \frac{1}{T_{score}} \sum_{s=1}^{S} \zeta(s)(N_{ref}(s) - N_{hyp}(s)) \qquad \forall(N_{ref}(s) - N_{hyp}(s)) > 0. \tag{2.33}$$

E_{FA} and E_{MISS} ~~mainly~~ indicate the performance of the clustering phase in detecting the correct number of speakers. Equation (2.30) can now be re-written as

$$\text{DER} = \text{SER} + \text{E}_{\text{FA}} + \text{E}_{\text{MISS}}. \tag{2.34}$$

2.3.1 Diarization Approaches and Systems

There are two main approaches in speaker diarization according to Anguera et al. (2012) and Moattar & Homayounpour (2012). The bottom-up approach, also known as Agglomerative Hierarchical Clustering (AHC), starts with a relatively high number of clusters and iteratively merges similar clusters until, in the ideal case, the correct number of clusters have been reached see e.g Siegler et al. (1997). The top-down approach, on the other hand, usually starts with one cluster and iteratively partitions until the correct number of clusters have been reached see e.g Fredouille & Senay (2006).

There are a number of other approaches that , in some sense, differ from the aforementioned approaches. The information-theoretic approach introduced in Vijayasenan et al. (2007) aims to minimise the loss in mutual information between subsequent clustering whilst preserving the mutual information in terms of a relevance variable. Another approach is binary key based diarization proposed by Anguera & Bonastre (2011) where segments of feature vectors are converted to single binary vectors and the clustering and re-segmentation are performed in the binary domain. An approach introduced by Rouvier & Meignier (2012) formulated the clustering part of diarization as an Integer Linear Programming (ILP) problem with the aim of minimising the number of clusters in addition to the dispersion within them.

The bottom-up approach is widely used mostly because of the algorithm presented in Ajmera & Wooters (2003) that became a standard system. It is based on an ergodic Hidden Markov Model (HMM) formalism where the number of states is equal to the initial number of clusters. The probability density function (PDF) of each state is assumed to be a GMM. An Information Bayesian Criterion (BIC) is used to assess the similarity between clusters and also as stopping criteria. For any two clusters, one GMM is fitted to the feature vectors of each and a third GMM is fitted to the feature vectors of both. The BIC based method depends on the log-likelihoods values for cluster merging. After each merging, a re-segmentation (refinement) step occurs where short segments (1-2 seconds) of speech feature vectors are reassigned to the clusters (GMMs) using the log-likelihood value in combination with the Viterbi algorithm to determine the best segmentation path. Cluster merging stops when the change in the BIC values becomes less than zero where at that stage each cluster is assumed to represent one speaker. The diarization system based on this algorithm is referred to here as BIC based.

The BIC based diarization system presents good performance at the cost of computation duration that exceeds Real Time (RT) in the standard form of the system as in (Anguera & Bonastre, 2011) where it was 1.19 $\times$RT. Probably the best achievement in speeding this system up is by Gonina et al. (2011) which provided (0.004-0.02) $\times$RT performance. This was achieved by parallelising the training of the GMMs using a GPU thus incurs an additional hardware cost. Some of the alternative approaches presented in the literature achieved comparable performance to the BIC based system but also with a cost-effective reduction in computational complexity.

The information-theoretic approach (Vijayasenan et al., 2007) gave 3-6 times faster performance than BIC based diarization and the binary keys approach performed at 0.103 $\times$RT (see Anguera & Bonastre (2011)). These alternative approaches have attracted attention because of their appealing performance in terms of speed. They have also been the target

of research that aimed to increase their diarization accuracy. Vijayasenan et al. (2011a) integrated acoustic (MFCC) and TDOA features in the information-theoretic approach to improve the system performance with a minimum of 0.34 $\times$RT speed Vijayasenan et al. (2008). Delgado et al. (2015a) introduced several improvements on the binary keys diarization system including a speed boost of up to 0.0354 $\times$RT making it the fastest diarization system as found by Joshi et al. (2016).

Integer Linear Programming (ILP) uses i-vectors as inputs where the i-vector extraction requires a large amount of external development data as explained earlier. In i-vector extraction, total variability modelling was made possible, so that this method could be suitably applied to the problem of cross-show speaker diarization as investigated by Dupuy et al. (2012). It was noticed in this latter study that the computation time for the ILP approach increases when speech duration increases. For 5, 10 and 15 hours of speech the computation time was 0.06 $\times$RT, 0.19 $\times$RT and 0.26 $\times$RT, respectively. Further improvement on the ILP approach was introduced by Dupuy et al. (2014) that enabled competitive performance to the BIC based system in the diarization of Broadcast News. Before the ILP approach, i-vectors have also been used for telephone speech diarization with the cosine similarity metric in Shum et al. (2011). Sell & Garcia-Romero (2014) used the Probabilistic LDA (PLDA) model for i-vector scoring in the diarization system.

Recent research efforts in speaker diarization include: microphone pair selection for the extraction of TDOA features in Multiple Distant Microphone (MDM) based diarization using the BIC based system Martínez-González et al. (2017). Garcia-Romero et al. (2017) presented a method similar to the one of Sell & Garcia-Romero (2014) but it replaced the i-vectors with the DNN embeddings (x-vectors) introduced by Snyder et al. (2017) which provided comparable performance for telephone speech diarization. A GMM modelling of TDOA features with the expectation-conditional maximisation algorithm and minorisation-maximisation approach was introduced in Parada et al. (2017). Given prior knowledge of the correct number of speakers, that method achieved comparable performance to the BIC based system in MDM diarization. A recent review of practical challenges in speaker diarization Church et al. (2017) included the computational complexity as one of the issues where a low complexity can deliver a real-time performance.

The binary keys diarization system is probably the most efficient in terms of speed and it does not need any prior modelling (e.g extensive external data as in the case for i-vector and x-vector based diarization systems). However, the performance of this system is not satisfactory. The work in Delgado et al. (2015b) improved the system performance for cross-show diarization by introducing intra-session and intra-speaker variability compensation. It

was then expanded in Delgado et al. (2015a) where further improvements were introduced for the cases of single and cross show diarization. Nonetheless, this latter work mostly improved the speed of single show diarization but its accuracy remained somewhat limited. The work here focuses on improving the system performance for the Multiple Distant Microphones (MDM) single show diarization. The existing approaches to binary key based diarization (see e.g Delgado et al. (2015a) and Anguera & Bonastre (2011)) make limited use of the availability of multiple microphones and the main challenge here is to integrate spatial (TDOA) features. Using these features in addition to acoustic can improve the performance of binary key based diarization as was the case with other systems, see e.g (Martínez-González et al., 2017).

2.3.2 Binary Key Based Diarization

This section reviews the work that lead to the development of the binary key based diarization system that will be investigated in this research. It describes the underlying concept and the system operation.

Binary keys were first introduced for speaker modelling in Anguera & Bonastre (2010). A binary key is a relatively low dimensional vector of binary values and it is derived from anchor models. The basic concept of anchor modelling is to represent a speaker's utterance with information gained from a set of models (anchor models) pre-trained from a defined set of speakers (anchor speakers), see Sturim et al. (2001). A binary key also models a speaker based on the concept of anchor modelling. For speaker identification in Anguera & Bonastre (2010), the anchor models were GMMs fitted to speech features of selected speakers. The speech features (in the enrolment or test phase) of the speakers were projected onto these models to yield the speaker representation. Although the selection of particular speakers' models as anchor models is important, the size of the anchor models is also important and it was the focus of Anguera & Bonastre (2010) in the process of deriving the binary keys. A model that represents the global acoustic space, similar to the Universal Background Model (UBM), is required to be obtained and it is termed the binary Key Background Model (KBM). The KBM consists of a collection of anchor speaker models and the overall number of mixture components determines the size of the binary keys. Anchor speakers' GMMs can be obtained by the Expectation-Maximisation (EM) algorithm or by Maximum a Posteriori adaptation (MAP) of the UBM to speakers' feature vectors as suggested by Anguera & Bonastre (2010).

In order to obtain a binary key $\mathbf{v} = (v_1, v_2, ..., v_B)$, $v_i \in \{0, 1\}$ for $1 \leq i \leq B$, to represent an utterance, a cumulative vector, $\boldsymbol{\upsilon} = (\upsilon_1, \upsilon_2, ..., \upsilon_B)$, $\upsilon_i \in \mathbb{N}$ for $1 \leq i \leq B$, with the same size of the binary key is required to be initialised with zeros. B indicates the size of the KBM which is also the size of the binary key and cumulative vector. Given the feature vectors of an utterance, the log-likelihood of each feature vector is determined for each Gaussian component of the KBM. In the cumulative vector, a ratio of Ω_1 of the positions of top Gaussian components with the highest log-likelihood is incremented by one. Then, to derive the binary key, the positions of a different ratio of Ω_2 of top Gaussians component with highest accumulated scores in the cumulative vector are set to one. The rest of the positions are set to zeros. This process is also illustrated in Fig. 2.8.

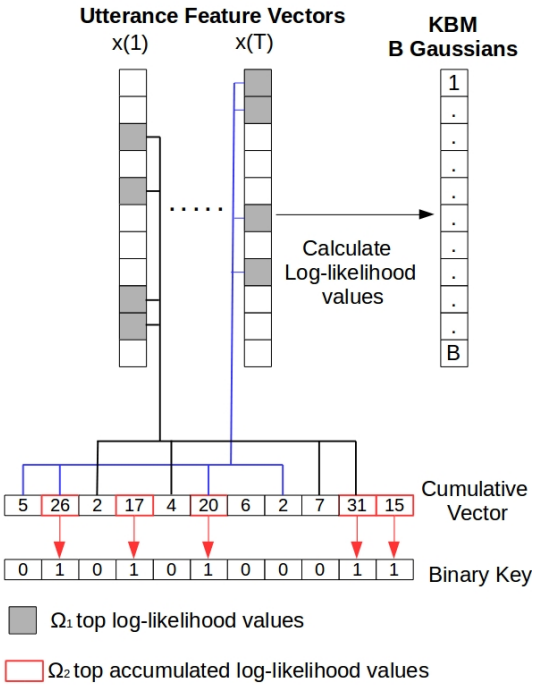


Fig. 2.8 This figure illustrates how the cumulative vector and then the binary key are derived from a speech utterance.

A value of 1 in the binary key indicates that the speech utterance coexists in the same acoustic region as the KBM's Gaussian component of that position. Speakers' speech features are represented according to their relative occupancy in the acoustic space represented by the KBM. Hence, theoretically speaking, two utterances of the same speaker should result in similar binary keys for the same KBM.

For the task of speaker diarization, the KBM was obtained from the meeting speech itself by Anguera & Bonastre (2011). Thus, the KBM training spared the need for a separate development data which also adds the advantage of avoiding mismatch between possible development data and the speech of the conversation of interest.

For diarization, the KBM training also includes the selection of discriminant anchor models which is one of the underlying concepts for binary keys. A meeting speech is divided into segments of 2 seconds with 75% overlap ratio. Then, a single Gaussian is trained for each segment where the segment size and the overlapping ratio guarantee that each Gaussian would be acoustically centred in a speaker (not on an uttered sound). Then, a process is conducted to select the subset of Gaussians that form the KBM. The first Gaussian is selected as the one that best models the segment it was trained on. The log-likelihood value was used for this purpose. Then, symmetrised Kullback-Leibler divergence was used in Anguera & Bonastre (2011) to select the most dissimilar Gaussian with the one selected first. This process proceeds until the desired size of the KBM is reached.

Some of the components of the system were later improved in Delgado et al. (2015a) and the work here follows this latest advancement. As the goal of the system was to present fast diarization, the symmetrised Kullback-Leibler divergence was replaced with the cosine similarity for the Gaussians means in the anchor models selection for obtaining the KBM.

The diagram in Fig 2.9 illustrates the system operation described here. Although it might not be apparent from Fig. 2.9, the system is considered to be composed of two main stages: a clustering stage and a re-segmentation stage. In a preliminary step of the clustering stage, all feature vectors are projected onto the KBM and the top Gaussians (specified number of top Gaussians) are determined for each feature vector. Let the projected feature vectors be called frames hereafter. The system commonly uses a uniform initialisation where the meeting frames stream is divided into uniform clusters (with some initial number of clusters, commonly 16). Then the binary keys of these clusters are derived from the frames initially assigned to them. The same meeting frames stream is divided into relatively small segments of 1 seconds size. As in Anguera & Bonastre (2011), these segments are extended by 1 second of frames on both sides such that the segment size becomes 3 seconds.

Binary keys are also derived for these segments using the relevant frames. After that, an agglomerative process is used to merge homogeneous clusters and segment re-assignment is performed given the new clusters. These processes are described as follows. A segment binary key is scored against all clusters binary keys. Each cluster is assigned particular segments that show more similarity to that cluster. Then, cluster binary keys are re-estimated using the newly assigned segments' frames. Any two clusters with the highest similarity

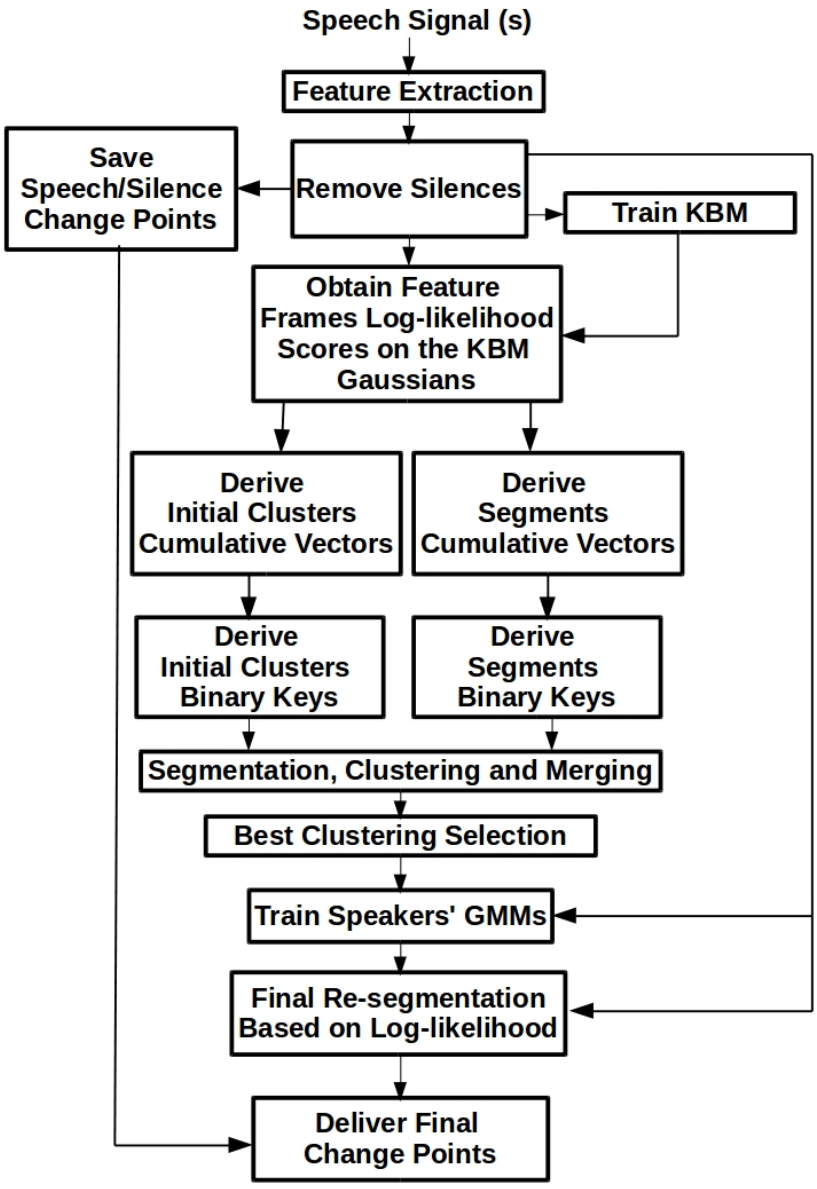


Fig. 2.9 Descriptive diagram of the binary key based diarization system.

between them are merged and the clusters number decreases by one. The process is continued until the number of clusters becomes one. However, the clustering structure at each iteration is saved for best clustering selection. The metric used for measuring the similarity between two binary keys is the Jaccard similarity coefficient expressed as follows

$$\mathscr{J}(\mathbf{v},\mathbf{v}') = \frac{\sum_{i=1}^{B}(\mathbf{v}(i) \wedge \mathbf{v}'(i))}{\sum_{i=1}^{B}(\mathbf{v}(i) \vee \mathbf{v}'(i))}. \tag{2.35}$$

where $\wedge$ indicates the boolean AND operator and $\vee$ indicates the boolean OR operator. The Jaccard coefficient is known as the intersection over union ratio and it is a more suitable metric for binary values (Boesch et al., 1977).

In Delgado et al. (2015a), the cumulative vectors $\boldsymbol{\upsilon}$ themselves were also used instead of the binary keys with the cosine similarity as a metric. At each iteration of the agglomerative clustering process, there is a different number of clusters and different segments assigned to those clusters. The selection of the best clustering structure among others is what distinguishes this approach. In Anguera & Bonastre (2011), the best clustering was indicated by a maximum T-test value determined for the distributions of within cluster and between cluster similarities. Delgado et al. (2015a) improved clustering selection by introducing a technique based on Within Cluster Sum of Squares (WCSS). For any clustering structure $\mathbb{C}$ of Θ clusters, $\theta_1, \theta_2, ..., \theta_\Theta$, the WCSS is given by

$$\mathscr{W}(\mathbb{C}) = \sum_{i=1}^{\Theta} \sum_{\mathbf{g}\in\theta_i} ||\mathbf{g} - \tilde{\mathbf{g}}_i||^2 \tag{2.36}$$

where $\tilde{\mathbf{g}}_i$ is the mean (centroid) of cluster θ_i. $||\ \ ||$ indicate vector normalisation. Good clustering structures result in low values of $\mathscr{W}$. Clustered structures with a relatively low number of clusters compared to the correct number of speakers have a high value of $\mathscr{W}$ with the highest value resulting at the cluster structure of $\Theta = 1$. The best clustered structure is selected using a graphical approach as illustrated in Fig. 2.10. The lowest and highest values of $\mathscr{W}$ are connected by a straight line. Then, all the values of $\mathscr{W}$ are plotted which forms a curve under that straight line. The clustering structure with the $\mathscr{W}$ value that forms the so called 'elbow' of the curve is selected as the best one. The elbow point is the one with the highest distance from the straight line.

The final step in this binary key based system is the re-segmentation process. The best cluster structure provides segments' labels in relation to the clusters. The feature vectors for those segments are used to train a GMM for each cluster. Finally, the log-likelihood of the

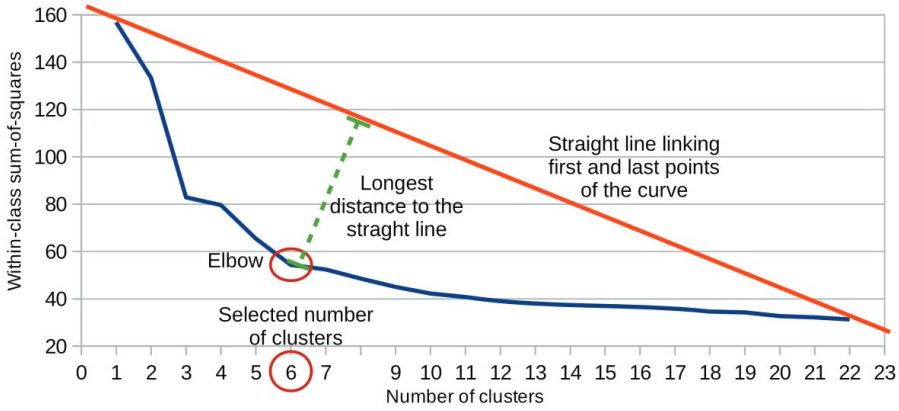


Fig. 2.10 Best clustering selection based on Within Cluster Sum of Squares (WCSS) (Delgado et al., 2015a).

feature vectors to the GMMs, with a log-likelihood smoothing window, is used to finely label corresponding speech for each speaker.

2.3.3 Acoustic Feature Extraction in MDM Diarization

If Multiple Distant Microphones (MDM) are available in a meeting then acoustic features can be extracted from a random or central microphone (van Leeuwen, 2006; van Leeuwen & Huijbregts, 2006). Alternatively, an independent diarization system can be set for the features extracted from each channel and then the results are combined Anguera et al. (2005). A more practical solution was presented in Anguera et al. (2005) which combines all the channels' signals in a weighted delay and sum fashion to produce an enhanced signal. This was later expanded and an acoustic beamforming algorithm was introduced in Anguera et al. (2007) and it became common practice in MDM diarization and used until recently in Martínez-González et al. (2017).

This beamforming algorithm works as follows. A Wiener filter is first applied to each channel's signal in order to reduce noise. Next, the central microphone is identified as the one that has the maximum cross-correlation with the rest of the channels and it is used as a reference (Anguera et al., 2007). As the channels' amplitudes are not consistent, an overall channels weighting factor is estimated for all channels in order to normalise the amplitudes. This normalising factor is estimated by finding the average of the absolute maximum amplitude over the segments of the speech signal. Then, the delays are calculated in segments of 250 ms between each channel and the reference channel. The best delays are selected as explained in Section 2.1.2. The delays are used to align the relevant segments

from all the channels to be summed. In the summation step, the channels are given weights according to their qualities and a triangular window is used to eliminate discontinuities in the resulting beamformed signal. Similar to delay estimation and channels summation, the weights are also estimated for each segment of the signals. The weight for segment j of channel i is given by (see Anguera et al. (2007))

$$\omega_i(j) = \begin{cases} \frac{1}{\tilde{M}} & j = 0 \\ (1-\alpha_a)\omega_i(j-1) + \alpha_a \tilde{R}_i(j) & \text{otherwise,} \end{cases} \tag{2.37}$$

where $\tilde{R}_i(j)$ is the average cross-correlation between segment j for channel i and the relevant aligned (based on the pre-estimated delays) segments of the rest of the channels. $\tilde{M}$ is the total number of channels and α_a is an adaptation ratio empirically set to 0.05.

This algorithm clearly has a number of dependencies such as the normalising factor, the selection of the best delays and the channel weights. Recently in Tu et al. (2017), alternative solutions were presented to overcome the imperfection of some beamforming techniques such as the problem of direction of arrival mismatch. One of the solutions proposed was the concatenation of features extracted from multiple signals which was found to introduce better recognition accuracy. Subsets of channels were used to obtain several beamformed signals and a concatenation of the speech features extracted from these beamformed signals was used.

Despite the enhanced performance in speech recognition, this method did not avoid beamforming. However, it motivated an idea presented here to use a concatenation of features extracted from individual channels instead of a beamformed signal. On the other hand, extracting features from all available channels is computationally inefficient and some channels are less likely to be of good quality. Therefore, a channel selection is established that aims to meet the plausible objective of identifying good quality channels.

In general, speech signal quality measures can be divided into intrusive and non-intrusive (see Falk et al. (2010)). Intrusive measures require a reference (such as a clean signal) while non-intrusive measures do not require a reference. The SNR parameter was used in Pardo et al. (2007) to identify a good quality channel to be used as a reference in delay estimation. Although estimating the SNR is non-intrusive, it may not be feasible since it is difficult to estimate (Bosworth et al., 2008).

Speech recognition research appears to have more interest in developing signal quality measures than speaker recognition research. In Distant Speech Recognition (DSR), Wolf & Nadeu (2014) introduced a number of decoder-based measures, like the variance of

the speech intensity envelope, for channel selection. Those measures are demanding as they require a classification of the recognised speech then the selection is made and the recognition is repeated using the selected channels. The modulation spectrum ratio introduced in Himawan et al. (2015) was also used for channel selection in distant speech recognition within a somewhat complicated framework. Original speech was convolved with different rooms' impulse responses. Then the correlation between contaminated speech and the Word Error Rate (WER) was used to predict the recognition performance. By assuming an exact knowledge of a real room impulse response, that measure was used to select the best channel.

Cepstral distance is an efficient signal quality measure. This intrusive measure was initially introduced by Kitawaki et al. (1988) to assess the distortion presented by speech coding techniques in reference to the original speech signal. Cepstral distance was long known for its flexibility and effectiveness in different applications (Guerrero et al., 2016). It was recently used for the selection of the least distorted channel by Flores et al. (2018) for distant speech recognition. As an intrusive measure, the use of the cepstral distance requires a reference channel which is assumed to provide a clean speech signal in some sense. In Flores et al. (2018), the authors proposed to compute a reference signal as the logarithm of the geometric mean of the signals from the available microphones calculated in the magnitude spectrum domain.

This method makes no distinction between the quality of the signals used in the computation of the reference signal. As a result, good and bad quality signals similarly contribute in the computation because of the unweighted mean element of the method. It would be more robust to assign preliminary quality-based weights in such an averaging process. A more reliable reference signal is used in this work for selecting good quality channels.

In comparison to the case of detecting the quality of a channel in general, the selection of the least reverberated channel in a non-intrusive way is further addressed in this work. The reverberation problem has been the focus of considerable research efforts. One way to tackle this problem is to de-reverberate the speech signal or features as in Feng et al. (2014) for speech recognition where a deep auto-encoders was used for this purpose. However, de-reverberation is difficult and non-reliable since it can introduce objectionable artefacts to the processed speech Falk et al. (2010). Alternatively, in Giri et al. (2015), a feature vector that characterises reverberation was extracted from the speech signal and input to a DNN in a room-aware DNN training for speech recognition. A similar concept was presented in Oo et al. (2018) in a reverberation-aware DNN training.

As of the channel selection target here, a method that characterises the degree of reverberation is required. The concept of modulation transfer function (MTF) is one of the earliest

approaches applied to evaluate the quality of speech transmission (against reverberation and other effects) between the speaker and the listener in an auditorium Houtgast & Steeneken (1985). In Malik & Farid (2010), reverberation is detected by estimating a decay parameter that embodies the extent of reverberation. That parameter is estimated from the speech signal using a maximum likelihood estimation. Falk et al. (2010) introduced a measure termed speech-to-reverberation modulation energy ratio for the diagnosis of de-reverberated speech to test for the feasibility of de-reverberation algorithms. In Jiang et al. (2014), binary classification using a DNN was introduced for reverberant speech segregation. That required the extraction of binaural features of the intraural time differences and intraural level differences that were used as the main auditory features.

Depending on the room characteristics, the degree of reverberation varies between subbands of the speech spectrum Ismail (2013). This will be tackled here hence time-domain reverberation measures, such as that of Malik & Farid (2010), are not applicable. Reverberation variability between subbands will be accounted for here by the selection of the least reverberated channel-subband. This selection does not require precise estimation of the degree of reverberation as expected from the measures presented in (Falk et al., 2010) and Ismail (2013). The method proposed here characterises the degree of reverberation in relation to the rest of the channels. While reverberation can be observed over the pitch period, at the frame level or long segments Wolf & Nadeu (2010), the new method characterises reverberation by considering spectrum subbands over the entire speech signal.

2.3.4 Diarization Systems Initialisation

Initialisation of diarization systems refers to the manner by which the process starts. More specifically, the way in which the conversation segments are obtained to produce initial models of the clusters. A carefully designed initialisation method that can improve the performance of the binary key diarization approach will be very useful. Unsupervised diarization systems, including the binary key based, commonly start with uniform clusters obtained by dividing the underlying conversation into large equal segments Tranter & Reynolds (2006). However, tuning of parameters such as the number of initial clusters and the number of Gaussian mixtures in Agglomerative Hierarchical Clustering (AHC) systems, like the BIC based, is often important Moattar & Homayounpour (2012).

A suitable initialisation method for binary key based diarization would be one that improves the performance and only adds a small computational complexity such that the appealing system's speed is approximately maintained. As discussed below, the initialisation

methods proposed in the literature may not be particularly suitable for the binary key based system since they can add a considerable computational load. This is because they depend on the acquisition of additional information or their application would require additional modelling.

In Anguera et al. (2006b), within the GMM-BIC framework, a preliminary speaker change points estimation is carried out, then the segments are classified into 'friend' and 'enemy' groups to finally create an initial set of clusters. Anguera (2006) presented a Cluster Complexity Ratio (CCR) to adapt the number of initial clusters and Gaussian mixtures. The CCR was optimised on a development dataset and then used for the evaluation set. A similar complexity measure that relates the number of feature vectors to the number of Gaussian mixtures was used in Woubie et al. (2015). A parameter called Constant Seconds Per Gaussian (CSPG) was used in van Leeuwen & Konečnỳ (2008) also to adapt the number of Gaussian mixtures and it was later further developed to Adaptive Seconds Per Gaussian (ASPG) by Imseng & Friedland (2009).

An initialisation method introduced in Luque et al. (2008) was based on clustering of TDOA features. A type of pre-clustering and a technique to estimate the number of initial clusters based on prosodic features was presented in Imseng & Friedland (2010). The work in Garau & Bourlard (2010) integrated visual cues in the initialisation process by using Visual Focus of Attention (VFoA) features and motion intensities.

The K-means algorithm has also been used for initialisation. However, in BIC based diarization, it was reported in Ajmera & Wooters (2003) that K-means operating on feature vectors did not have a significant impact on the performance compared to using uniform clusters. In Shum et al. (2011), K-means was used to perform first pass clustering (initial clustering) on i-vectors as well as the final segmentation refinement. That final refinement was, however, not precisely k-means based as the cluster 'means' were estimated using new clusters' i-vectors estimated from speech feature vectors assigned to them according to Dehak et al. (2011). Also, that work focused on telephone conversations where the number of speakers is normally two which is relatively low.

For the binary key based system, suitable methods, such as K-means, are proposed and investigated in this work. These methods are more appropriate since they largely depend on parameters that are already estimated within the system's framework.

2.4 Summary

This review has covered many of the important techniques for the field of speaker recognition. It has been found that there are a number of gaps that should be investigated. In the extraction of the most widely used speech feature, MFCC, the use of a subset of a bank of overlapping filters increases the residual correlation in the correlation matrix of the filters' outputs. That can affect the efficiency of the mel-scale filter bank analysis of the speech spectrum, especially considering that the analysis is based on the human perception of sound unlike the DCT which is data independent. In the extraction of LPCC, determining the autocorrelation function as the inverse Fourier transform of an FFT based smooth spectral estimate can address two problems together. On the one hand, it can avoid the occurrences of corrupted autocorrelation estimates under noisy conditions. On the other hand, it can address the case of having the conventional LP-based spectrum containing sharp peaks for speakers with a higher voice pitch.

PCA was shown to be a commonly used technique for dimensionality reduction and feature fusion. The estimation of the principal components can be influenced by differences in the variances of the speech features unless it is based on the correlation matrix or proceeded by variance normalisation. Also, when global principal components are estimated from feature frames of many speakers, those frames must be suitably weighted to decrease the contribution of the undesired ones and the outliers. Also, non-classical, iterative approaches for PCA can be more efficient. The recently developed techniques may, however, be slow or suffer from a low convergence rate.

It is necessary to demonstrate that any positive impact, that would result from the methodologies proposed here for acoustic feature extraction and fusion, can generalise to other speaker recognition tasks (frameworks). In addition to the diarization system, evaluations will be carried on the well recognised i-vector based verification system which required relatively large amount of development data. When data augmentation (by adding effects like noise) is used to increase the amount of available data, effects that simulate speech transmission channels could be useful. The use of such an effect has not been investigated. Also, in the literature, the amount of power for any added effects can be seen to be somewhat arbitrary given that it was not calibrated in any way or relation to observations of system performance.

Another important speaker recognition problem is speaker diarization. Diarization systems can benefit from the integration of TDOA features. Those systems usually use modelling approaches that assume normality in the distribution of the underlying features.

However, GCC-PHAT delay estimates (TDOA features) has a positively skewed distribution. Especially for the binary key based system (as will be analysed in Chapter 5), the distribution of these features must be normalised by a suitable transformation.

Binary key based diarization is fast but its accuracy is not particularly competitive. Besides the integration of TDOA features, when a conversation is recorded over multiple microphones, alternative techniques that can outperform beamforming are necessary for the performance. A concatenation of acoustic features extracted from all of the channels (microphones) can be advantageous as opposed to features extracted from a single beamformed signal. However, it is expensive when many channels are available, hence, a selection of channels is mandatory. Most diverse or best quality channels can be selected. Additionally, the selection of the best quality channels requires an appropriate choice of a reference channel.

Another issue, that may be overlooked by channel selection and certainly by beamforming, will also be addressed in this work. Due to a meeting room's impulse response, reverberation effect may vary across the speech spectrum. Selection of the least reverberated channel's subbands can tackle this issue. In such a case, acoustic features will only be extracted from the selected subbands.

Finally, it was observed that non-uniform initialisation methods have been the focus of a number of works. Non-uniform approaches proposed in the literature can outperform uniform initialisation. However, they can consume processing time that may exceed the overall diarization time of the binary key based approach. This does not appear to have been addressed before and so this work proposes a number of fast non-uniform initialisation methods that can be seen to be more appropriate.

Chapter 3

Data Augmentation and Acoustic Feature Extraction

The methodology presented in this chapter addresses two distinct aspects of the front-end of speaker recognition systems. The first is specific to the i-vector based verification system where a data augmentation method is presented to tackle the problem of insufficient development data for the establishment of the system. The second is a modification in the extraction of MFCC features and an extension to the extraction of LPCC features. This latter part tackles the quality of features generally used in speaker recognition systems. However, speaker recognition performance using these new features is evaluated in the framework of i-vector based verification. The modification of MFCC feature extraction essentially lies in the calculation of the cepstral coefficients separately from subsets of a filter bank. In LPCC, the extension is based on the idea of fitting a multitaper spectrum estimation in the extraction of LPCC features.

The effect of using data augmentation on each component of the i-vector system is separately demonstrated. In the new MFCC, the performance is evaluated for a number of parameter variations including the number of filters in the filter bank and cepstral coefficients. Also, the effect of using a Hamming window and multitaper spectrum smoothing in the new MFCC is investigated. The appropriate type of multitaper and number of tapers for the new LPCC is identified empirically based on the experimental results presented.

3.1 Data Augmentation

The i-vector based speaker recognition system, first introduced by Dehak et al. (2011), models inter-speakers and intra-speaker variability (total variability) simultaneously. As explained in the literature review, the goal of intra-speaker variability (session/channel variability) modelling is to reduce the effect of session or channel mismatch between enrolment and test speech in speaker recognition. For that purpose, the establishment of the system requires special development data that contains speech samples recorded over different channels for the development speakers. This is specifically required for learning the total variability subspace, conducting Linear Discriminant Analysis (LDA) and for training the Probabilistic Linear Discriminant Analysis (PLDA) model used for scoring in the recognition phase.

When the development data is not sufficient, the i-vector system cannot perform appropriately because it will not be able to model session and channel variability for speakers (intra-speaker variability). Such development data is not widely available to researchers hence a data augmentation technique is introduced here to tackle the problem. The goal of this technique is to produce additional channel-variable recording by incurring simulated channel effect on a recording in hand.

3.1.1 Theory Behind Data Augmentation

The data augmentation in this work is theoretically based on a speaker model synthesis suggested by Teunen et al. (2000) which is flipped here as will be explained shortly. The model of Teunen et al. (2000) tackled the problem of channel mismatch between enrolment and test samples. According to that work, when there exists two (enrolment and test) utterances of the same speaker with 'speaker and channel'-dependent supervectors $\mathbf{m}_u$ and $\tilde{\mathbf{m}}_u$ (expressed by (2.21)), the model uses the assumption that $\tilde{\mathbf{m}}_u$ was synthesised from $\mathbf{m}_u$ by adding a supervector $\tilde{\mathbf{c}}$ that depends only on the channel conditions of the two utterances

$$\tilde{\mathbf{m}}_u = \mathbf{m}_u + \tilde{\mathbf{c}}, \tag{3.1}$$

where $\tilde{\mathbf{c}}$ is assumed to be a channel compensation supervector. This assumption is flipped here by passing the speech signal through a Gaussian channel in order to incur a different channel effect on the signal, see equation (3.6). Hence, any available recording becomes two recordings, one of them is the original and the other is with an added Gaussian channel effect.

The reason for using a Gaussian channel is that, in information theory, Gaussian noise is a basic statistical model used to mimic the effect of random processes that occur in nature (see e.g Houdré et al. (2016)). It is also used to model many practical channels such as wired and wireless telephone channels. The additive noise in such channels are due to a combination of causes. By the central limit theorem, the cumulative effect of a number of random effects will be approximately normal thus the Gaussian assumption becomes valid (see e.g Cover & Thomas (2012)).

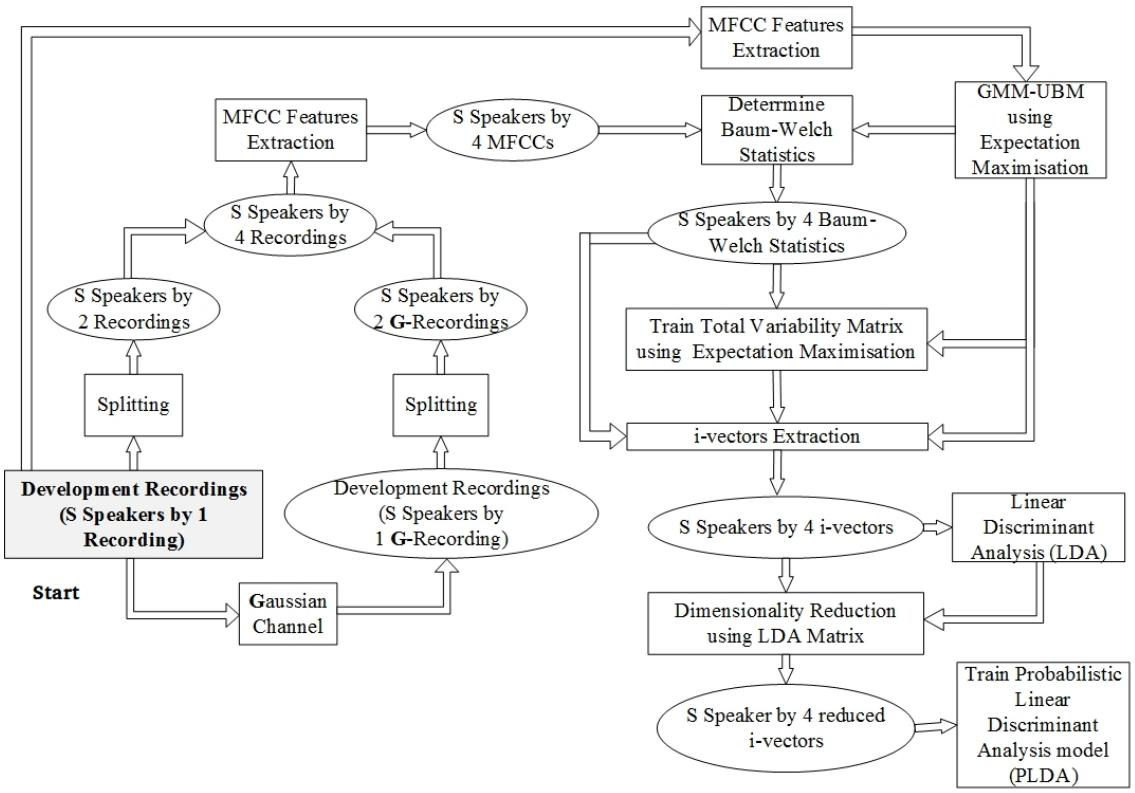


Fig. 3.1 Diagram of the i-vector based system with data augmentation. MFCC features are used as an example.

The development data that was available for this work has one utterance for each speaker. The pre-described data augmentation can increase this number by one. This is because only one type of noise (Gaussian) is used here for the purpose of data agumentation by noise addition. However, two utterances per speaker are still not sufficient for the development of the i-vector system. In Rao & Mak (2013), it was observed that in LDA and Within-Class Covariance Normalisation (WCCN), the system performance is more influenced (degraded) by utterance length of less than 1 minute. Hence, it was suggested that sub-utterances of a

long utterance can help produce more i-vectors for each speaker which was found to enhance speaker verification performance. The conclusion of that study is used here in this work by splitting the previously achieved two utterances to make them four. This is found to enable reasonable performance of the i-vector system as used here in this work.

The diagram in Fig. 3.1 illustrates how data augmentation is performed and the number of utterances delivered to the system per each development speaker. Note that data augmentation is not used for the enrolment and test samples.

3.1.2 Gaussian Noise Power Determination

The power of the additive Gaussian noise is controlled such that the produced speech signals maintain a fixed signal-to-noise ratio (SNR). In other words, the methodology takes into account the signal power to prevent the added noise power from becoming destructive. Note that, since the speech signals in the development and test data used here are not strictly clean, Gaussian noise is added on top of the noise embedded in the signals. Transmission channel and environmental noise types with different SNRs may be present in the signals.

The speech signal is a random continuous-time signal that becomes a discrete-time signal after sampling. Suppose a speech signal $\mathbf{s}$ with finite length of N samples is expressed as

$$\mathbf{s} = [s_1, s_2, ..., s_N]. \tag{3.2}$$

The power of the signal is defined by

$$\varphi_{\mathbf{s}} = \frac{1}{N} \sum_{n=1}^{N} s_n^2. \tag{3.3}$$

Then the power of the additive Gaussian noise that maintains a desired linear signal-to-noise ratio η is

$$\varphi_{\mathbf{r}} = \frac{\varphi_{\mathbf{s}}}{\eta}. \tag{3.4}$$

The added Gaussian noise is defined by a normally distributed random vector $\mathbf{r} = [r_1, r_2, ..., r_N]$. The elements of this vector r_n are thus Gaussian distributed with a constant mean of zero and constant variance $\varphi_{\mathbf{r}}$, i.e. $r_n \sim \mathcal{N}(0, \varphi_{\mathbf{r}})$; the distribution of which can be

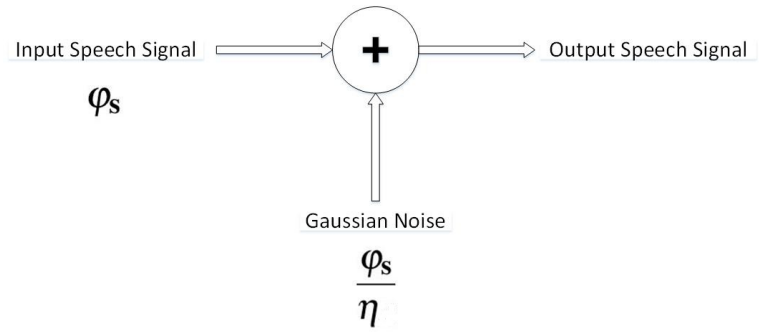


Fig. 3.2 Addition of Gaussian noise with SNR controlled power.

expressed by the probability density function:

$$\mathscr{G}(r_n;0,\varphi_{\mathbf{r}}) = \frac{1}{\sqrt{2\pi\varphi_{\mathbf{r}}}} \exp\left(-\frac{r_n^2}{2\varphi_{\mathbf{r}}}\right). \tag{3.5}$$

As noise normally has zero mean, this makes the power equal to the variance. Hence, the desirable Gaussian noise is a normally distributed random vector with zero mean and $\varphi_{\mathbf{r}}$ variance, where $\varphi_{\mathbf{r}}$ is determined in (3.4). The new speech signal with added Gaussian noise will be

$$\mathbf{s}_{\mathscr{G}} = \mathbf{s} + \mathbf{r}. \tag{3.6}$$

The value of η_d is empirically decided based on the performance of the system indicated by the equal error rate (EER) as will be shown in the results section (3.3). Such calibration is necessary but was not present in the previous work by e.g Snyder et al. (2017).

3.2 Acoustic Feature Extraction

3.2.1 Odd-Even MFCC (OE-MFCC)

Conventional mel filter banks (FB) comprise of filters that are overlapped (by 50%) in order not to lose the speech spectrum attenuated by the edges of each filter. Due to this overlap, the log energy of a particular filter somewhat resembles that of the adjacent ones especially if they (all three) capture a slowly varying section of the spectrum. Hence, overlapping filters bank may present relatively high residual correlation in the covariance matrix of the filter bank's output log-energies. The proposed methodology is to use subsets in the form of the odd indexed and even indexed filters, as illustrated in Fig. 3.3; and to extract cepstral coefficients separately from each subset. This can present the following advantages:

- It decreases the residual correlation for each subset (as assumed to be desired);
- No spectrum is lost compared to overlapped filter banks, see Fig. 3.3;
- The effect of narrow-band noise on the cepstral coefficients is reduced, see e.g (Besacier & Bonastre, 2000);
- The computation complexity in DCT application is minimised, see (Sahidullah & Saha, 2012).

The proposed methodology can also compensate for the limitation of extracting higher order cepstral coefficients in standard MFCC. It should be noted that higher order coefficients of MFCC are more susceptible to noise Reyes-Galaviz & García (2009).

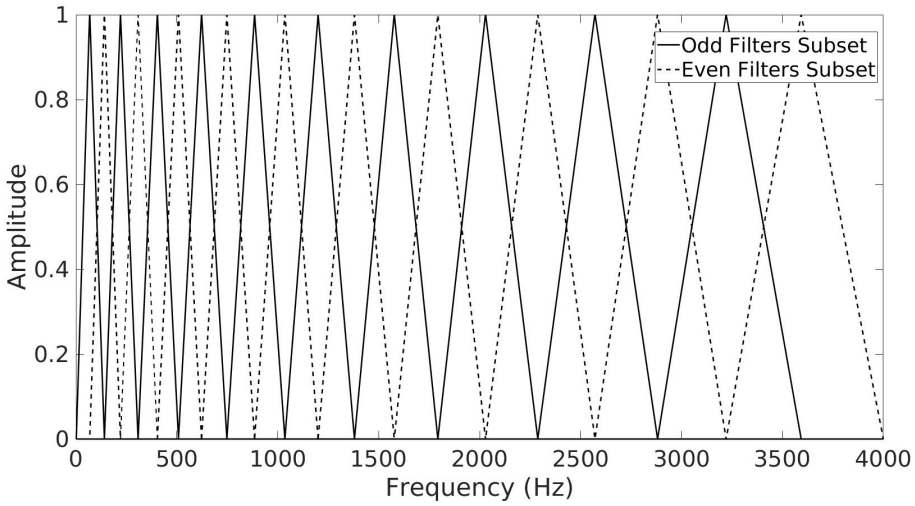


Fig. 3.3 Odd and even subsets of a filter bank that consists of overlapping filters. Each subset is applied separately to the output of FFT and cepstral coefficients are extracted separately for the output of each of them.

Calculating the energy of odd and even indexed filters separately has been used before. Previously it was used to achieve computational efficiency in hardware implementation of MFCC as in Jo et al. (2016). In Vu et al. (2010), only the odd filters subset was determined and the points of the even filters subset were estimated by subtracting each odd indexed filter from 1. However, in both of these cases, the log-energies of all odd and even filters were pooled together and the DCT was then applied unlike the methodology proposed here in this work.

3.2.1.1 Construction of Odd and Even Filters Subsets

Leading from equations (2.1), (2.2), (2.3) and (2.4) and given that M is the total number of an overlapping set of filters in a filter bank, the centres of the odd indexed filters on the mel-scale are defined as

$$\zeta_{f_{c,1}}(n) = (2n-1)\Delta \quad \text{where} \quad 1 \leq n \leq \lceil M/2 \rceil, \tag{3.7}$$

where Δ is the spacing between the filters on the mel-scale which was previously defined in (2.2). n is an integer and $2n-1$ are the indices of the odd filters.

The centres of the even indexed filters on the mel-scale are defined as

$$\zeta_{f_{c,2}(n)} = 2n\Delta \quad \text{where} \quad 1 \leq n \leq \lfloor M/2 \rfloor, \tag{3.8}$$

where $2n$ are the indices of the even filters. The odd subset of the triangular filters $\mathscr{H}_1(n, \kappa_f)$ with the centres of (3.7) are constructed in the linear frequency scale according to the following formulation:

$$\mathscr{H}_1(n,\kappa_f) = \begin{cases} \dfrac{\hat{\psi}(\zeta_f) - \hat{\psi}\left(\zeta_{f_{c,1}}(n) - \dfrac{\Delta}{2}\right)}{\hat{\psi}(\zeta_{f_{c,1}}(n)) - \hat{\psi}\left(\zeta_{f_{c,1}}(n) - \dfrac{\Delta}{2}\right)} \textbf{ for } \hat{\psi}\left(\zeta_{f_{c,1}}(n) - \dfrac{\Delta}{2}\right) \leq \hat{\psi}(\zeta_f) < \hat{\psi}(\zeta_{f_{c,1}}(n)); \\ \dfrac{\hat{\psi}(\zeta_f) - \hat{\psi}\left(\zeta_{f_{c,1}}(n) + \dfrac{\Delta}{2}\right)}{\hat{\psi}(\zeta_{f_{c,1}}(n)) - \hat{\psi}\left(\zeta_{f_{c,1}}(n) + \dfrac{\Delta}{2}\right)} \textbf{ for } \hat{\psi}(\zeta_{f_{c,1}}(n)) < \hat{\psi}(\zeta_f) \leq \hat{\psi}\left(\zeta_{f_{c,1}}(n) + \dfrac{\Delta}{2}\right); \\ 0 \quad \text{elsewhere,} \end{cases} \tag{3.9}$$

where ζ_f is the mel-scale nonlinear frequency calculated in (2.1) and $\hat{\psi}$ is the transformation to the linear frequency scale as expressed by (2.4).

The filters of the even subset $\mathscr{H}_2(n,\kappa_f)$ can be similarly achieved as in the following

$$\mathscr{H}_2(n,\kappa_f)=\begin{cases}\dfrac{\hat{\psi}(\zeta_f)-\hat{\psi}\Big(\zeta_{f_{c,2}}(n)-\dfrac{\Delta}{2}\Big)}{\hat{\psi}(\zeta_{f_{c,2}}(n))-\hat{\psi}\Big(\zeta_{f_{c,2}}(n)-\dfrac{\Delta}{2}\Big)} \textbf{ for } \hat{\psi}\Big(\zeta_{f_{c,2}}(n)-\dfrac{\Delta}{2}\Big)\leq\hat{\psi}(\zeta_f)<\hat{\psi}(\zeta_{f_{c,2}}(n));\\ \dfrac{\hat{\psi}(\zeta_f)-\hat{\psi}\Big(\zeta_{f_{c,2}}(n)+\dfrac{\Delta}{2}\Big)}{\hat{\psi}(\zeta_{f_{c,2}}(n))-\hat{\psi}\Big(\zeta_{f_{c,2}}(n)+\dfrac{\Delta}{2}\Big)} \textbf{ for } \hat{\psi}(\zeta_{f_{c,2}}(n))<\hat{\psi}(\zeta_f)\leq\hat{\psi}\Big(\zeta_{f_{c,2}}(n)+\dfrac{\Delta}{2}\Big);\\ 0 \quad \text{elsewhere.}\end{cases} \tag{3.10}$$

These subsets are used to decompose the spectrum of each frame into two complementary subsets of filters energies. Finally, the DCT is applied to the log of the energies of each subset separately. This process is described in Chapter 2 by equations (2.6) and (2.7). The cepstral coefficients obtained from $\mathscr{H}_1(n,\kappa_f)$ and $\mathscr{H}_2(n,\kappa_f)$ can be referred to as $MFCC^{odd}$ and $MFCC^{even}$, respectively. Let $\mathscr{H}_{\mathscr{E},1}(z)$, $1\leq z\leq Z$, be the log of the odd filters energies. $\mathscr{H}_{\mathscr{E},1}(z)$ is obtained by using $\mathscr{H}_1(n,\kappa_f)$ to decompose the speech spectrum as in equation (2.6). $MFCC^{odd}$ can then be calculated as

$$\text{MFCC}_r^{odd}=\sum_{z=1}^{Z}\mathscr{H}_{\mathscr{E},1}(z)\cos\left[r\Big(z-\frac{1}{2}\Big)\frac{\pi}{Z}\right] \quad \text{for} \quad r=1,\,2,\,\ldots,\,R \tag{3.11}$$

where R is the number of $MFCC^{odd}$ cepstral coefficients. $MFCC^{even}$ is calculated in the same way. The resultant coefficients are then augmented to form one feature vector.

3.2.1.2 Residual Correlation of the Covariance Matrix of the Filters Output

The residual correlation is the mean of the absolute values of the off-diagonal elements of a correlation matrix, see Sahidullah & Saha (2012). The residual correlation (ε) of the filter bank function (overlapped and non-overlapped) is evaluated for a first order Markov process covariance matrix with different correlation coefficients (ρ). Let $\mathbf{A}$ be the covariance matrix

of a first order Markov process expressed as

$$\mathbf{A} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots \\ \rho & 1 & \rho & \rho^2 & \dots \\ \rho^2 & \rho & 1 & \rho & \dots \\ \rho^3 & \rho^2 & \rho & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3.12}$$

where ρ is the correlation coefficient of the Markov process. Let the size of $\mathbf{A}$ be $a \times a$. Now let $\mathbf{H}$ be the matrix of filterbank filters with size $h \times a$ where h is the number of filters. The transformation of $\mathbf{A}$ using $\mathbf{H}$ can be expressed as

$$\hat{\mathbf{A}} = \mathbf{H}\mathbf{A}\mathbf{H}^T \tag{3.13}$$

The nonzero off-diagonal elements in $\hat{\mathbf{A}}$ form a measure of the residual correlation (Poularikas, 2010).

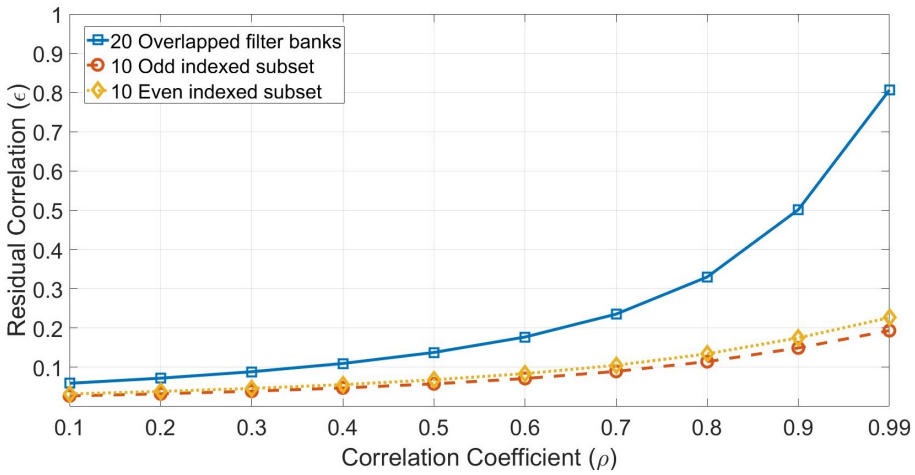


Fig. 3.4 The residual correlation of the filter banks function for different values of the correlation coefficient of a Markov-1 process covariance matrix.

Fig. 3.4 shows that the residual correlation of overlapped filters bank is higher than any of the odd-indexed or even-indexed subsets. It can also be noticed that this difference increases for higher values of ρ. Fig 3.5 shows how subsets of overlapping filters bank exhibit higher residual correlation which increases (except at very high values of ρ) as the number of filters decreases.

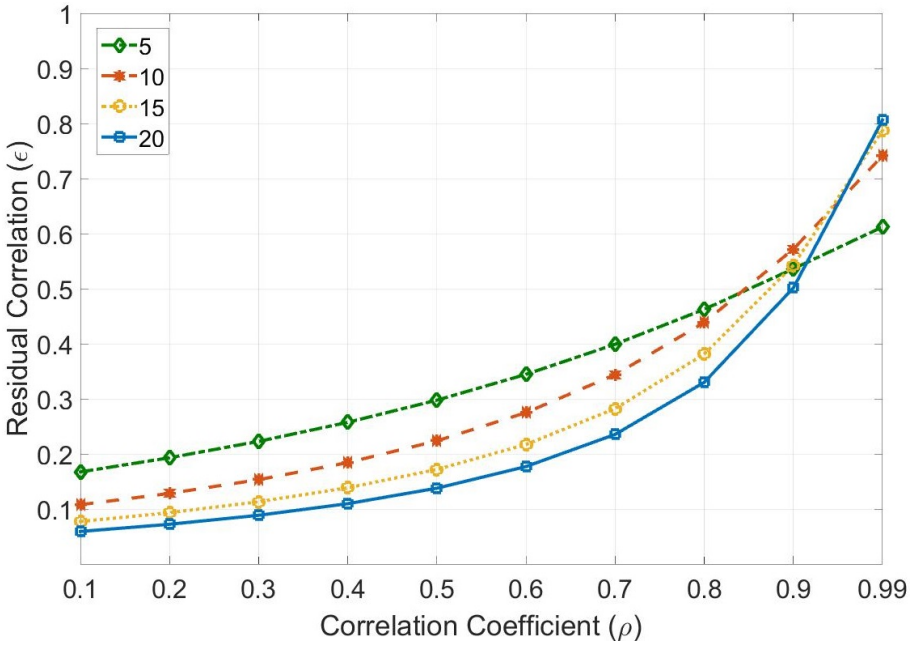


Fig. 3.5 The residual correlation for variable number of overlapping filters. One can notice that for most values of ρ. the residual correlation increases as the number of overlapping filters increases.

No. of Filters	Full Set (ε)	Odd Subset (ε)	Even Subset (ε)
20	0.6547	0.6339	0.6416
24	0.6486	0.6314	0.6367
28	0.6415	0.6258	0.6321

Table 3.1 Residual correlation of the correlation matrix of the filter bank log-energies.

Table 3.1 reports the residual correlation in the correlation matrix of the filter bank output log-energies for speech data. The speech data used is the training samples of the 2002 NIST SRE dataset (described in Section 3.3). It can be observed that for the three cases (in terms of the filters number) of filter bank, the odd and even subsets exhibit lower residual correlation than that of the full set.

Cepstral coefficients extracted from both odd and even subsets are concatenated for use in the speaker recognition system. Both subsets interchangeably cover the full band of the speech spectrum. This relatively increases the residual correlation in the correlation matrix of their cepstral coefficients. However, the performance of speaker verification does not appear to be specifically sensitive to this. In the method of block MFCC proposed by Sahidullah & Saha (2012), the cepstral coefficients of overlapping blocks exhibited relatively higher

residual correlation in their correlation matrix. Nonetheless, they generally result in better performance than some of the other forms of block MFCC presented in that work.

3.2.1.3 Correlation of Cepstral Coefficients of Odd Even Subsets

Apart from the residual correlation, an experiment is conducted to assess the likeliness of peer (from an order perspective) cepstral coefficients of odd and even subsets. The experiment shows that, except for a few, there exists relatively high diversity between peer cepstral coefficients of the odd and even subsets. In the experiment, for a particular speech utterance, cepstral coefficients are extracted with two subsets of 14 odd and 14 even filters. Let $\mathbf{M}_1$ and $\mathbf{M}_2$ represent the cepstral coefficients of the odd and even subsets, respectively. For the sake of comparison, 13 cepstral coefficients are extracted for the same utterance with a set of 14 overlapping filters. Denote this set of cepstral coefficients with $\mathbf{M}$.

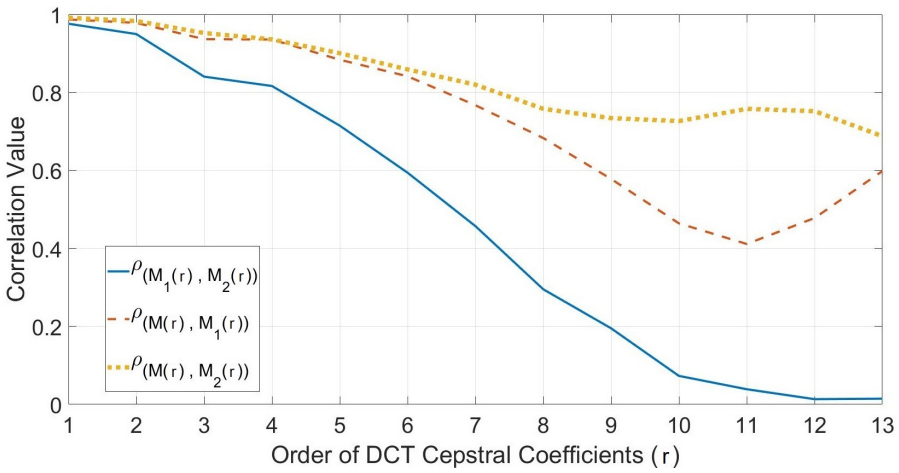


Fig. 3.6 Correlation among cepstral coefficients of overlapped filters bank and the non-overlapped filters subsets.

Afterwards, the degree of correlation between all these sets of cepstral coefficients is measured as in the following Sharma (2005)

$$\rho(\mathbf{M}_1(r),\mathbf{M}_2(r)) = \frac{\mathbf{M}_1(r)\mathbf{M}_2^T(r)}{\sqrt{\sum_{\forall i}\mathbf{M}_{1,i}^2(r) \times \sum_{\forall i}\mathbf{M}_{2,i}^2(r)}} \tag{3.14}$$

where $\rho(\mathbf{M}_1(r),\mathbf{M}_2(r))$ is Pearson's correlation coefficient between the r^{th} order cepstral coefficients of set $\mathbf{M}_1$ and set $\mathbf{M}_2$. The coefficients $\rho(\mathbf{M}(r),\mathbf{M}_1(r))$ and $\rho(\mathbf{M}(r),\mathbf{M}_2(r))$

are also calculated using equation (3.14). Note that the mean of $\mathbf{M}$, $\mathbf{M}_1$ and $\mathbf{M}_2$ must be normalised in order to calculate Pearson's correlation coefficient using (3.14).

Fig. 3.6 shows the correlation between all of these sets of cepstral coefficients for 13 orders of DCT coefficients. Compared to the $\rho_{(\mathbf{M}(r),\mathbf{M}_1(r))}$ and $\rho_{(\mathbf{M}(r),\mathbf{M}_2(r))}$ cases, one can notice that there is low (peer) correlation between the cepstral coefficients of the odd and even subsets as indicated by $\rho_{(\mathbf{M}_1(r),\mathbf{M}_2(r))}$. These findings can be used as an indicator to remove cepstral coefficients from OE-MFCC if they consist of redundant information which may harm the system performance.

3.2.2 Multitaper-Fitted LPCC

The method presented here aims to avoid direct estimation of the autocorrelation function and to have an averaged spectrum which can penalise spectral sharp peaks as discussed in Section 2.1.1.2. One of the objectives in spectrum estimation is to achieve minimal bias that is mostly caused by spectral leakage and is usually reduced by using a window function, commonly, a Hamming window as stated in Neustein & Patil (2012). A single particular window is not an optimal choice as it down-weights the speech frame values at the edges of the window causing loss of information. Furthermore, spectral leakage may also not be minimised compared to the multitaper method proposed by Thomson (1982). When spectral leakage is not minimised, the chance of spectral bias persists Prieto et al. (2007). In Kinnunen et al. (2010), the multitaper method was used for spectral estimation in the extraction of MFCC features as an alternative to the Hamming window and was shown to enhance the performance of speaker recognition.

Multitaper spectrum estimation first introduced by Thomson (1982) results in an averaged spectral estimate from different orthogonal tapers (windows). The use of more than one window also allows those parts of the signal which are attenuated by one window to be captured by some other window in the taper set. The tapers are assigned weights that sum to 1. The first taper has the lowest spectral side lobes (and the highest weight) and the side lobes increase for higher order tapers. The resulting spectrum is smoothed and has less variance so that the spectral leakage is minimised giving a reduced bias, see e.g Prieto et al. (2007). The estimated multitaper power spectrum is a weighted sum of these tapers given by

$$\hat{s}[k] = \sum_{m'=1}^{M'} w_{m'} \left| \sum_{n=1}^{N-1} \lambda_{m'}[n] s[n] \exp\left(-j2\pi \frac{nk}{N}\right) \right|^2, \tag{3.15}$$

where M' is the number of tapers, $\lambda_{m'}$ is a taper associated with a weight $w_{m'}$, $s[n]$ is a speech sample and N is the number of samples.

Accordingly, the multitaper method is integrated into the extraction of the Linear Prediction Cepstral Coefficients (LPCC) features here. The extraction of these features requires the determination of the Linear Prediction Coefficients (LPC) which starts by having the speech signal framed. Then each frame is commonly passed to a Hamming window before estimating the autocorrelation function. One of the methods of determining that function is based on the Wiener-Khinchin theorem which states that the Fourier transform of the autocorrelation function is equal to the power spectrum, hence, the inverse Fourier transform of the power spectrum is the autocorrelation function, see Kantz & Schreiber (2004).

To incorporate the multitaper method, the autocorrelation function for each speech frame is determined by having the inverse Fourier transform of the multitaper power spectrum calculated as follows

$$\hat{r}_{ss}[n] = \sum_{k=1}^{K-1} \hat{s}[k] \exp\left(j2\pi \frac{nk}{K}\right). \tag{3.16}$$

Multipeak, Thomson and Sine tapers are tested and the results are presented in Section 3.3.4.

3.3 Experimental Results

This sections presents the results of speaker verification in light of the front-end proposed in this chapter. The impact of data augmentation on the performance of i-vector based verification is first evaluated. Then, the system performance is evaluated using OE-MFCC features. Finally, the performance is evaluated using the proposed Multitaper-Fitted LPCC features. The evaluation parameter used here is the Equal Error Rate (EER) which is an operation point on the Detection Error Trade-off curve (DET). This curve is produced by plotting the detection (verification) False Negative Rate (FNR) against the False Positive Rate (FPR). The point on the curve where the FNR is equal to the FPR is the EER.

3.3.1 Corpora and i-vector Based System Setup

The development data of the i-vector system includes the NIST 2002 SRE telephone training data (English) Martin & Mark (2004), the NCHLT Speech Recognition microphone corpus (English) De Vries et al. (2014) and the LWAZI Speech Recognition telephone corpus (English, Afrikaans, Sesotho and Zulu) de Vries et al. (2014). The system is gender-independent and in order to balance the analyses, the number of development speakers is

639 males and 639 females speakers (1278 speakers). Speech recordings with average length of 2 minutes can be obtained from these datasets for each development speaker. Table 3.2 summarises the number of utterances available, the number after splitting and after adding simulated Gaussian channel effect.

	No. of Speakers ♂, ♀	**No. of Utterances**	**No. of Utterances after Splitting**	**Splitting plus Gaussian Noise**
NIST 2002 SRE (English)	139 , 132	271	542	1084
NCHLT (English)	110 , 100	210	420	840
LWAZI (English)	92 , 104	196	392	784
LWAZI (Afrikans)	101 , 99	200	400	800
LWAZI (Sesotho)	96 , 106	202	404	808
LWAZI (Zulu)	101 , 98	199	398	796
Total	639 , 639	1278	2556	5112

Table 3.2 Summary of Development data and number of utterances obtained from the NIST 2002 SRE data Martin & Mark (2004), the NCHLT data De Vries et al. (2014) and the LWAZI data de Vries et al. (2014).

The data used to evaluate the data augmentation process is the NIST 2002 SRE telephone set Martin & Mark (2004). For each of the 139 male and 191 female speaker of this set (total of 330 speakers), the training data is used as speakers' enrolment. Then one test sample for each speaker is used to evaluate the system. Each test sample is scored against all enrolments samples which makes a total of 108900 gender-independent verification trials. This evaluation set was used for detailed evaluation of the methodology presented for data augmentation.

A subset of the 2010 NIST speaker recognition evaluation dataset, see Martin & Greenberg (2010) is also used for evaluation. This subset is the core-core evaluation condition (commonly referred to by Det5) trials which contains telephone speech for enrolment and test data. Det5 includes a total of 30373 trials, 708 of which are target trials and the rest are non-target trials. This subset is used to evaluate the methodology presented for acoustic feature extraction. Some results on the effect of data augmentation are also presented using this dataset.

Only the original recordings of NIST 2002 SRE telephone training data (139 males and 139 females) are used to estimate a gender-independent GMM-UBM with 2048 mixtures. The dimensionality of the total variability matrix is 400 resulting in i-vectors with 400 dimensions. These are then reduced to 150 using LDA. The PLDA model is trained using

all of the development i-vectors and is used for scoring the i-vectors of the evaluation sets[1]. These parameters are fixed for the system except for the features which are varied in this chapter and in Chapter 4.

3.3.2 Gaussian Noise Level and Impact on i-vector Based System Components

The idea presented for data augmentation is centred on the inclusion of simulated Gaussian channel effect. As described in Section 3.1.1, utterance splitting presented in previous work by Rao & Mak (2013) is used here for the same purpose of increasing the amount of development data. Hence, results are produced for the case when development utterances were only split and a total of 2556 utterances were used to establish the system with some initial performance. Afterwards, Gaussian noise is used to increases the number of utterances to 5112 for more appropriate development of the system. This helps in evaluating, purely, the effect of adding Gaussian noise according to the pre-described methodology.

In this part of the evaluation, the speech features used were 13 MFCC coefficients (excluding the 0^{th} coefficient) calculated from Hamming windowed 25 ms speech frames with 40% overlap (10 ms shift), appended with their first and second derivatives. The filter bank used consists of 24 triangular filters.

Equal Error Rate (EER) and Minimum Detection Cost Functions (minDCF) of the 2008 and 2010 NIST evaluations are used here as performance measures.

The power of the Gaussian noise added is different for each speech sample but the output signals exhibit the same SNR. The appropriate SNR is empirically determined based on the performance of the system. Four values of SNR are assessed as shown in Fig. 3.7. The best system performance is achieved at a SNR of 30 dB. For 10 dB SNR, the noise power is relatively high. However, some performance improvement is achieved compared to the case when the development data is only split. The performance at 20 dB SNR is better compared to that of 10 dB SNR, because the noise power is decreased. Following the best performance accomplished at 30 dB SNR, it can be noticed that the performance at 40 dB SNR is comparable to that at 20 dB although the noise power is less. This is because at such low power, the noise did not have a remarkable effect on the output signals, hence, they were not much variable from the the original signals.

[1]The total variability matrix (**T**) and the Gaussian PLDA model are estimated using the Microsoft Research (MSR) Identity Toolbox (Sadjadi et al., 2013).

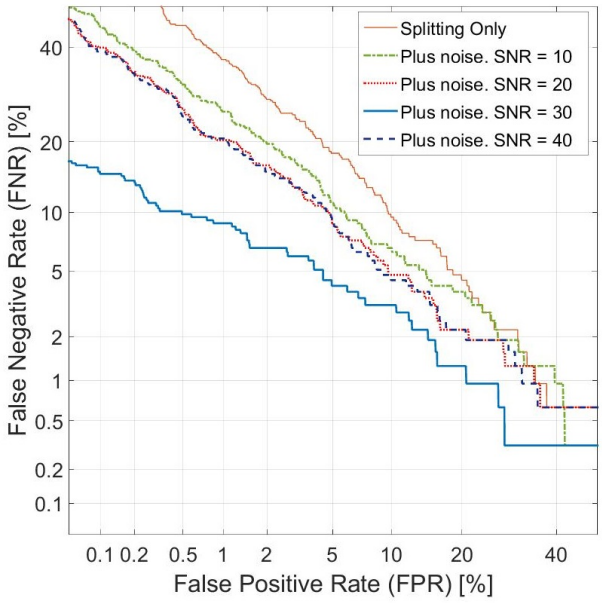


Fig. 3.7 Detection Error Tradeoff curves of system performance at different SNRs of the resulting speech signal with added Gaussian noise.

The effect of using utterances with added Gaussian noise is investigated separately in the development of each system component and the results are reported in Fig. 3.8 and Table 3.3. In LDA, the use of these utterances produced an improvement of 1.35% in EER. For the case of PLDA training, the EER is further reduced by 3.4%. When the utterances with Gaussian noise are involved in the total variability matrix training alone, a slight degredation in the performance was experienced. However, when they are used in all of the system components, the overall reduction achieved in EER was 5.38%. This amount of reduction in EER is higher than that of the methodology presented in Snyder et al. (2018). This is because of the type of noise used here and the methodology for controlling an appropriate level of noise. More importantly, this is attributed to the fact that the performance of the i-vector speaker verification framework degrades when the development data is insufficient.

Methodology in System Components	**EER%**	**minDCF 2010**	**minDCF 2008**
Splitting in (T+LDA+PLDA)	9.81	4.65	0.088
Plus noise in LDA	8.46	3.76	0.085
Plus noise in PLDA	5.06	1.52	0.029
Plus noise in (T+LDA+PLDA)	**4.43**	**1.33**	**0.025**

Table 3.3 System Performance in terms of EER and DCF. It shows the effect of including utterances with added Gaussian noise in different components of the i-vector based system.

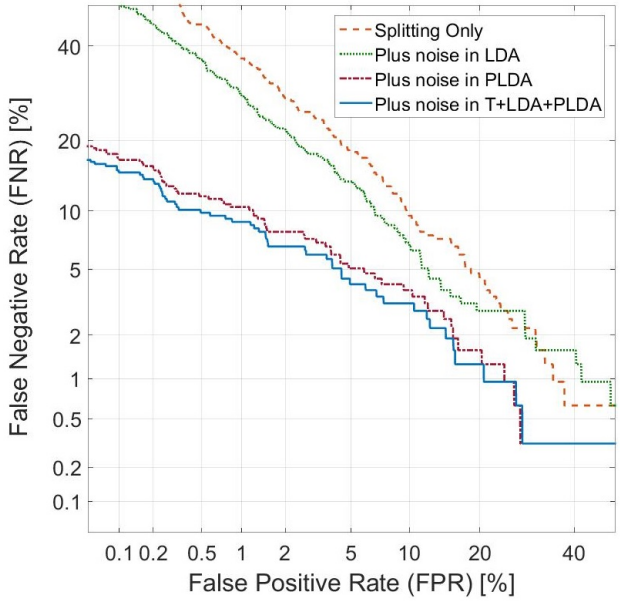


Fig. 3.8 Detection Error Tradeoff curves of system performance. Illustrates the effect of using utterances with added Gaussian noise on the system components [in LDA, in PLDA and in (T+LDA+PLDA)].

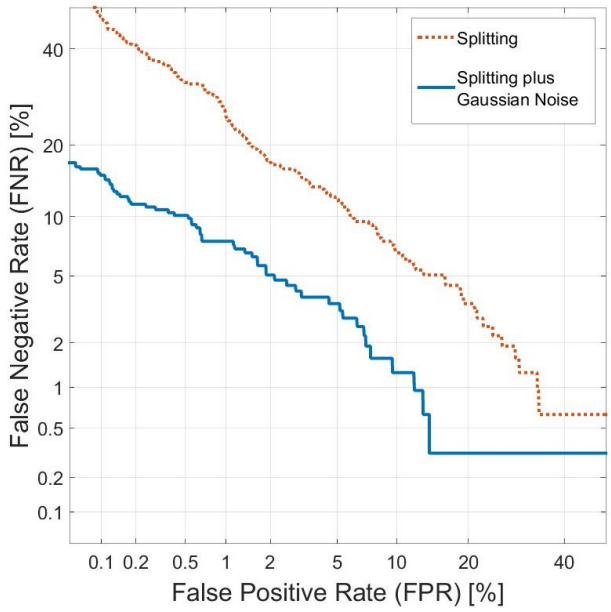


Fig. 3.9 Effect of using Gaussian noise in data augmentation using the Det5 subset of the 2010 NIST SRE set.

Among the types of noise that can be added to the development data, Gaussian noise in particular can be helpful in modelling general mismatch between enrolment and test utterances, especially if it is caused by the transmission channels.

Fig. 3.9 shows another evaluation of the presented methodology on the Det5 subset of the 2010 NIST SRE set Martin & Greenberg (2010). MFCC features are also used but their spectrum is this time estimated using the multitaper method with four multipeak tapers[1].

3.3.3 Effect of Parameters Variations on OE-MFCC

This subsection includes a study of speaker verification performance in light of a number of parameter variations in OE-MFCC. It also presents a comparison to MFCC and block based MFCC. Features are extracted from speech frames of 25 ms length with 10 ms frame shift. For MFCC and OE-MFCC features, the power spectrum of the speech frames is estimated using the multitaper method with four multipeak tapers.

In this part, 13 cepstral coefficients (excluding the 0^{th} coefficient) are obtained by applying the DCT to each set and subsets of filters bank log-energies. The cepstral coefficients are appended with their first and second derivatives. For OE-MFCC, it was experimentally found that the first two coefficients of the 13 basic cepstral coefficients degrade the performance if they are kept together from both the odd and even subsets. This is possibly because keeping these two coefficients only presents redundant information as they are highly correlated between odd and even subsets (see Fig. 3.6). Hence, these two coefficients are removed for the even subset through all the experiments presented but their first and second derivatives are kept.

Figures 3.10 and 3.11 help to illustrate the system performance for MFCC and OE-MFCC comparing the Hamming window results to the multitaper (four multipeak tapers) spectrum smoothing and also the number of filters used. The feature dimension is not varying, 39 and 76 for MFCC and OE-MFCC, respectively. It is notable that OE-MFCC greatly benefits from multitaper spectrum estimation where it results in lower EER compared to MFCC for all amounts of filters investigated. For the Hamming window case, although OE-MFCC is not always superior, it presents better performance in a number of cases and the lowest EER (compared to Hamming window MFCC) with 35 filters bank. OE-MFCC also has a stable low EER operation point in the range of 32 to 34 filters bank.It can be noticed that in this part of the experiment, the number of filters in the filter bank is constrained to the range of 28-35 filters. The lower limit (i.e. 28) is set to allow the extraction of at least 13 cepstral

[1]The basis for the multipeak multitaper is provided in Appendix A.3.

from the odd and even filters subsets. Beyond the specified upper limit (i.e. 35), the spectral decomposition can become poor because of the limited (and relatively low) number of FFT bins. In case of higher number of filters, one filter can have the value of one FFT bin at the low frequency region.

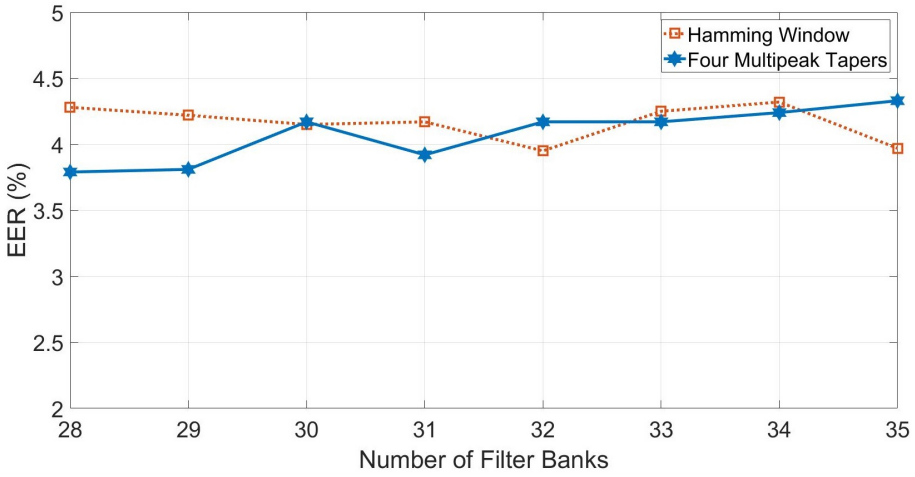


Fig. 3.10 System performance using MFCC features with variable number of filters and fixed feature dimension of 39.

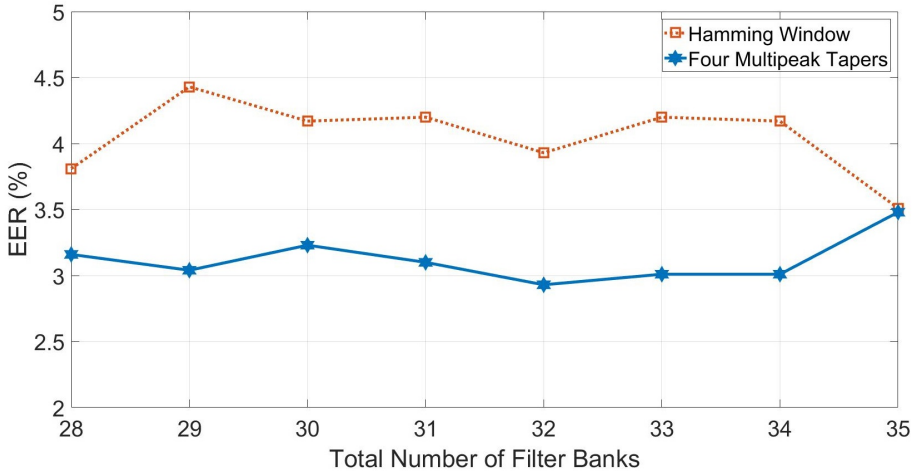


Fig. 3.11 System performance using OE-MFCC features with variable number of filters bank and fixed feature dimension of 76.

With spectrum estimated using the multitaper method, Fig. 3.12 illustrates the performance of OE-MFCC for a lower number of filters with comparison to MFCC. The dimension of MFCC features is the number of filters minus one plus delta and delta delta. OE-MFCC

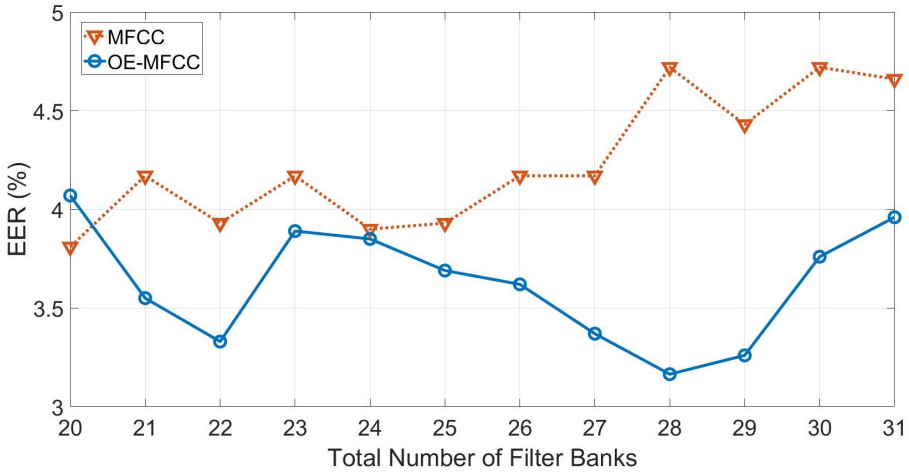


Fig. 3.12 System performance, OE-MFCC and MFCC, with variable number of filters bank and feature dimension.

features dimension is the MFCC's total feature number minus five. For example, with 28 filters, MFCC dimension is 81 and OE-MFCC features dimension is 76, including delta and double delta and after removing two coefficients from OE-MFCC as explained earlier in this subsection. Fig. 3.12 indicates that, for the majority of cases, the performance of OE-MFCC is superior to MFCC even for a lower number of filters (lower than those addressed in figures 3.10 and 3.11) and with varying feature dimensionality as well.

Filter Bank	MFCC Dim.	MFCC **EER**	OE Dim.	OE **EER**	NOBT Blocks	NOBT Dim.	NOBT **EER**	OBT Blocks	OBT Dim.	OBT **EER**
20	57	4.25	52	**4.07**	1-8 , 9-20	54	4.16	1-9 , 8-20	60	4.10
24	69	4.74	64	4.30	1-9 , 10-24	66	**3.88**	1-10 , 9-24	72	4.19
28	81	4.97	76	**3.80**	1-11 , 12-28	78	4.67	1-12 , 11-28	84	4.21

Table 3.4 Performance comparison of OE-MFCC, block MFCC and MFCC using Hamming window spectrum smoothing. EER is in percentage.

Filter Bank	MFCC Dim.	MFCC **EER**	OE Dim.	OE **EER**	NOBT Blocks	NOBT Dim.	NOBT **EER**	OBT Blocks	OBT Dim.	OBT **EER**
20	57	3.81	52	4.06	1-8 , 9-20	54	4.32	1-9 , 8-20	60	**3.79**
24	69	3.90	64	**3.85**	1-9 , 10-24	66	3.99	1-10 , 9-24	72	3.92
28	81	4.72	76	**3.16**	1-11 , 12-28	78	3.84	1-12 , 11-28	84	3.99

Table 3.5 Performance comparison of OE-MFCC, block MFCC and MFCC using multitaper spectrum estimation. EER is in percentage.

Tables 3.4 and 3.5 report a comparative performance of OE-MFCC, block MFCC and MFCC. This is investigated for multitaper and Hamming window spectrum smoothing. In block MFCC, when the blocks are not overlapping the case is referred to as Non-Overlapped Block Transformation (NOBT). Alternatively, they may be overlapped and are referred to as Overlapped Block Transformation (OBT). In NOBT, it is found here that two blocks present good performance (over MFCC) where the first block covers the frequency band 0-883.17 Hz and the second block covers the band 745.93-4000 Hz. Accordingly, for 20 filters bank, the first block includes the 1^{st} to the 8^{th} filter and the second block includes the 9^{th} to the 20^{th} filter. This frequency band coverage is accounted for when the number of filters is higher than 20 filters. For OBT, the blocks are allowed to overlap by one filter, where higher overlap was previously found to harm the performance by Sahidullah & Saha (2012).

One can see from tables 3.4 and 3.5 that OE-MFCC and block MFCC result in better performance than MFCC especially when the number of filters are increased. OE-MFCC is better than block MFCC for most cases. Especially with a relatively high number of filters bank, OE-MFCC presents superior performance and the lowest EER. For all OE-MFCC, block MFCC and MFCC, better features are extracted in the case of multitaper spectrum estimation.

3.3.4 Effect of Multitapers Type and Numbers on Multitaper-Fitted LPCC

Speaker verification performance for the i-vector system is evaluated here using the proposed multitaper-fitted LPCC features. First, the autocorrelation function is determined as the inverse of a multitaper power spectrum. Then 12 LPC coefficients are calculated and used to extract 13 LPCC coefficients appended with their first and second derivatives for a total of 39 coefficients. The performance is then investigated for the taper types previously used for MFCC extraction in Kinnunen et al. (2010). These are the *Thomson*, *Multipeak* and *sine* tapers[1].

From Fig. 3.13, one can observe that the optimum performance indicated by the minimum EER of 4.11% is given when four multipeak tapers are used. The performance is also compared to the baseline of the commonly used Hamming window. It can also be noticed that a higher number of tapers (over 5) appears to degrade the performance due to an increase in the side-lobes of higher order tapers. The verification performance using multitaper-fitted LPCC shows that multipeak tapers is the best taper type.

[1]The bases for these tapers' types are provided in Appendix A.3.

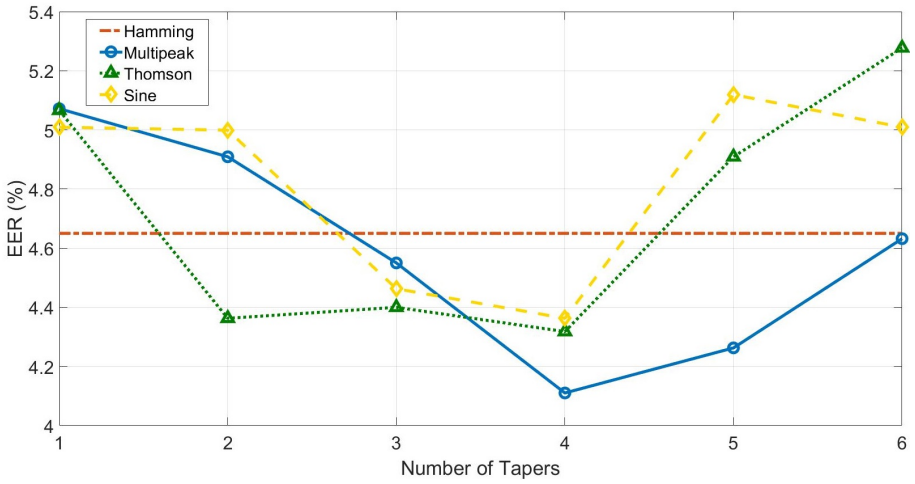


Fig. 3.13 Effect of tapers type and number on EER using LPCC features.

The results presented here by multitaper-fitted LPCC and the results from Kinnunen et al. (2010) and Alam et al. (2013) on MFCC features, confirm together that multipeak tapers are the best multitaper type for speech processing in speaker recognition. In this work, using four multipeak tapers in MFCC and OE-MFCC features extraction provided verification EERs of 3.79% and 3.16%, respectively. Despite that, in comparison to Hamming window based LPCC, the verification performance is improved with multitaper-fitted LPCC, MFCC and OE-MFCC features presents lower verification EER.

3.4 Summary

The main focus and interest of this chapter had two important aspects. First, it tackled the problem of the lack of development data required to establish an i-vector based speaker verification system. It was shown that Gaussian noise is an appropriate simulated channel effect for data augmentation. The data augmentation procedure helped in developing the speaker verification system with reasonable performance which was useful to evaluate system performance aspects related to factors such as the type of features used.

The second aspect focused on improving the extraction of two popular speech features of speaker recognition systems. In comparison to other subband based MFCC, OE-MFCC helped place the emphasis on improving the performance of the filter bank 'transformation' of the speech spectrum. The experiments on speaker verification appear to confirm the potential usefulness of OE-MFCC. The experiments also provided comparisons to traditional

MFCC and block MFCC for both cases of Hamming window and multitaper based spectrum estimations.

The latter multitaper based spectral estimation was fitted in the extraction of LPCC features. In comparison to traditional LPCC, multitaper fitted LPCC appeared to show improvements to the performance of speaker verification mostly when using four multipeak tapers.

The next chapter uses the data augmentation method of this chapter to enable the use of an i-vector based speaker verification system. This system will be used to evaluate the performance of the techniques to be introduced for the fusion of features like OE-MFCC and multitaper fitted LPCC. Chapter 6 presents some experiments that demonstrate the performance of binary key based diarization using OE-MFCC features.

Chapter 4

Recurrent Neural Network based Feature Transformation

As is generally the case with pattern recognition systems, a speaker recognition system can make use of multiple features to enhance its performance. This chapter presents an efficient methodology for feature fusion which also includes decorrelating and dimensionality reducing properties. Principal Component Analysis (PCA) is a traditional technique often used for this purpose in the signal processing literature. The methodology presented here is also based on PCA but using less well known methods which are developed here for application to speech features.

First, some considerations regarding the use of PCA with speech features are addressed. Then an efficient solution for weighted PCA is presented. This solution is based on recurrent neural network (RNN) methods for finding the most dominant eigenvector of a real symmetric matrix. The calculation of the correlation or covariance matrices includes the association of a weight matrix. In this chapter, the weight matrix is used to assign weights for each feature vector. Weighted principal components are then extracted from weighted correlation and weighted covariance matrices.

The weights are obtained by calculating the log-likelihood of the feature vectors for a single Gaussian background model (SG-BM) fitted with the same feature vectors. The convergence rate of the recurrent neural network is reported. As a result of the reduced dimensionality, the savings in the processing time within the i-vector based system are also reported.

4.1 Critical Considerations for PCA on Speech Features

The principal components are commonly considered to be the eigenvectors of the covariance matrix (e.g of MFCC feature vectors) and the associated eigenvalues are the amount of variance interpreted by those eigenvectors. Let $\mathbf{X}$ represent a matrix of feature vectors pooled together from a population of speakers. The covariance between two cepstral coefficients of L frames referred to here as $\mathbf{a}$ and $\mathbf{b}$ is

$$\sigma_{\mathbf{a},\mathbf{b}} = \frac{1}{L}\sum_{l=1}^{L}(\mathbf{a}_l - \bar{a})(\mathbf{b}_l - \bar{b}), \tag{4.1}$$

where $\mathbf{a}$ and $\mathbf{b}$ are rows in $\mathbf{X}$ and the columns of $\mathbf{X}$ are the feature vectors resulting from individual MFCC or similar feature extraction process. $\bar{a}$ and $\bar{b}$ are the means of $\mathbf{a}$ and $\mathbf{b}$. Their correlation can be expressed as

$$\tilde{\mathbf{\Sigma}}_{\mathbf{a},\mathbf{b}} = \frac{1}{L}\sum_{l=1}^{L}\frac{(\mathbf{a}_l - \bar{a})(\mathbf{b}_l - \bar{b})}{\sigma_{\mathbf{a}}\sigma_{\mathbf{b}}}, \tag{4.2}$$

where $\sigma_{\mathbf{a}}$ and $\sigma_{\mathbf{b}}$ are the variances for $\mathbf{a}$ and $\mathbf{b}$, respectively. For $\mathbf{X}$, whose mean is normalised, the covariance matrix is expressed as

$$\mathbf{\Sigma} = \mathbf{X}\mathbf{X}^T. \tag{4.3}$$

It is known that in order to perform PCA, the matrix $\mathbf{X}$ must be mean normalised. From (4.1) and (4.2), if the variance of $\mathbf{X}$ is also normalised, then both equations are equal because $\sigma_{\mathbf{a}} = \sigma_{\mathbf{b}} = 1$. Hence, $\mathbf{\Sigma}$ of (4.3) becomes the correlation matrix, denoted here by $\tilde{\mathbf{\Sigma}}$.

It was demonstrated in Section 2.1.3 that speech features, like MFCC and LPCC, exhibit large differences in the variances of their cepstral coefficients. The logarithms [1] of the covariance and correlation matrices of $\mathbf{X}$ are depicted in Fig. 4.1, where the feature vectors of $\mathbf{X}$ are 13 MFCC coefficients appended with their first and second derivatives. It can be seen that the attributes of the covariance matrix are affected by the variance values of the cepstral coefficients. For example, in the top left corner the highest attributes are associated with the higher variance coefficients. On the other hand, the relationship between the cepstral coefficients expressed by the correlation matrix, do not seem to be affected by those numerical variations. For example, the diagonal of the correlation matrix contain the highest values in the matrix, which is the correlation of each cepstral coefficient with itself. Hence, it seems

[1] The logarithm of the matrices is shown to make the plots more visually comprehending.

to be very important for speech features that the feature variances are normalised before performing any analysis for PCA.

One may also notice a 3×3 structure in the matrices shown in Fig. 4.1. This structure shows the covariance (and correlation) of MFCC cepstral coefficients with their first and second derivatives (velocity and acceleration terms of the feature vector).

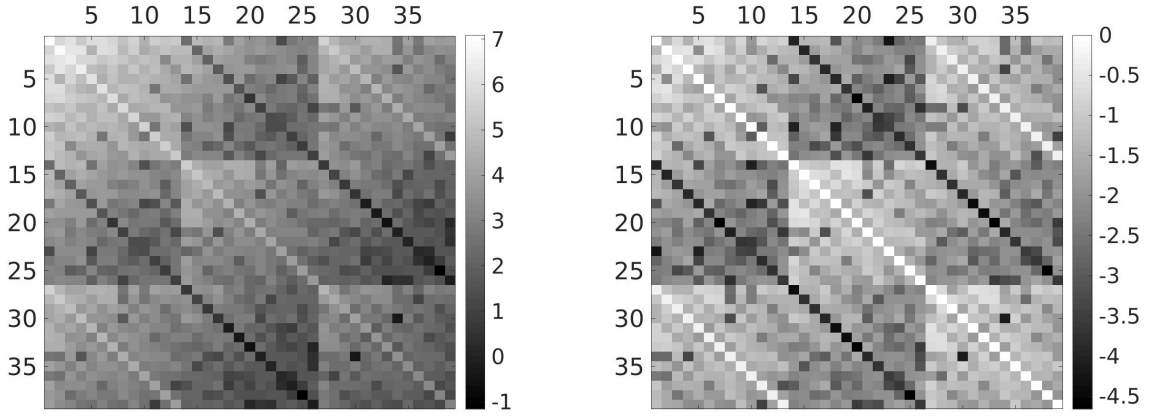


Fig. 4.1 Left image: logarithm of the covariance matrix. Right image: logarithm of the correlation matrix. These matrices are determined using a set of 13 dimensional MFCC feature vectors appended with their first and second derivatives thus the total dimensionality is 39.

Another aspect that must be considered in the analysis of PCA is its robustness to outliers. An experiment is conducted here to estimate the amount of outliers that may exist in a more than 9 million MFCC feature vectors calculated from utterances of $\sim$ 1200 speakers. Fig. 4.2 illustrates the method used. In the experiment, the mean feature vector is first calculated then the Euclidean distance is determined between all feature vectors and the mean feature vector. Now denote the the median of the distances by Q. The first quartile Q_1 is determined as the median of the distances that are smaller than Q. The third quartile Q_3 is determined as the median of the distances greater than Q. The interquartile IQR is then determined, which is $Q_3 - Q_1$. These parameter are used to identify the lower fence, $Q_1 - 1.5 \times IQR$, and the upper fence, $Q_3 + 1.5 \times IQR$, of the distances distribution. Distance values that are greater than the upper fence or smaller than the lower fence indicate outlying feature vectors. For the data under study, the experiment indicated that around 1% of the points are outliers, which is more than 99 thousand feature vectors. Therefore, it can be useful to establish methods for PCA that can be robust to outliers and, possibly, bad feature vectors.

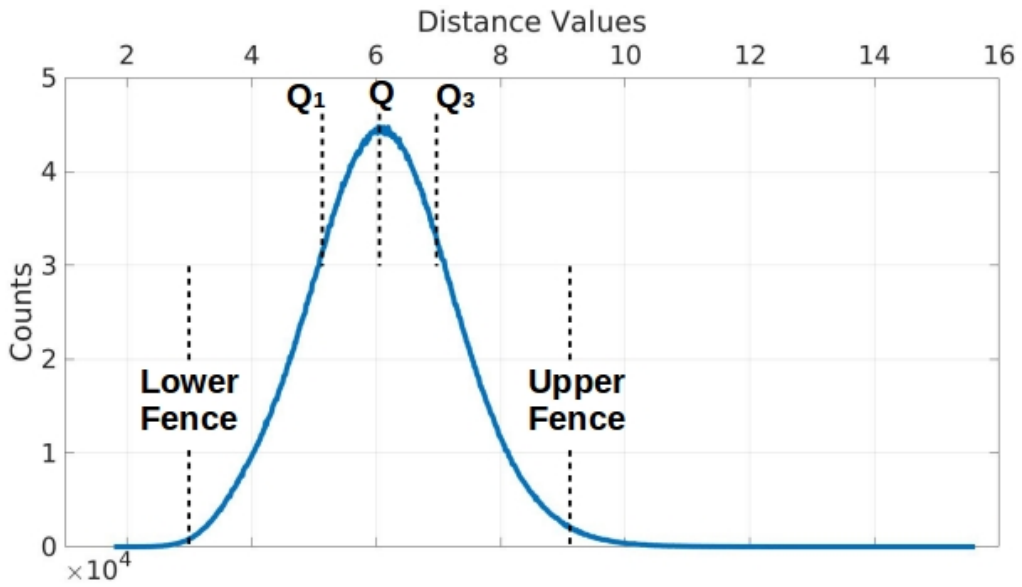


Fig. 4.2 Distribution of the Euclidean distances. This figure illustrates the method used to determine the possible amount of outliers in a set of feature vectors that could be used to extract the principal components.

4.2 Recurrent Neural Network Solution for WPCA

4.2.1 Methodology

This section introduces a class of Recurrent Neural Networks (RNN)s that can be used to extract the principal components of weighted covariance and correlation matrices. The usually used technique for PCA, Singular Value Decomposition (SVD), does not compute or use a correlation or covariance matrix. Thus, it is difficult to include weighting in the analysis. Also, as discussed in Section 2.1.3, PCA based on the Expectation Maximisation (EM) algorithm was found to be more precise (Bailey, 2012). The power iteration method introduced in Delchambre (2014) was an improvement over the EM solution. The proposed RNN-based methodology is found to provide equivalent PCA solution to the power iteration method but has a higher convergence rate, as will be shown shortly.

The type of the RNN can be defined by its architecture. Furthermore an RNN can be designed to model a dynamical system. Since a mathematical process can sometimes be viewed as a dynamical system, Rajasekaran & Pai (2002) formulated the eigendecomposition problem as an equilibrium problem for a dynamical model of a RNN. In that work, the RNN was used to identify the largest eigenvalue and the associated eigenvector of a real symmetric

matrix. Similarly in (Yi et al., 2004), a class of RNN was proposed to determine the largest and smallest eigenvalues and the associated eigenvectors.

The RNN presented in (Rajasekaran & Pai, 2002) had fixed weights which were the elements of a real symmetric matrix. Using those weights, it was shown that when the network input is an arbitrary vector, the output converged to the equilibrium state, i.e. the dominant eigenvector of that real symmetric matrix. This work defines the objective of the learning algorithm of (Rajasekaran & Pai, 2002) and considers it for the eigendecomposition problem here. The RNN based approach considered here is capable of identifying the desired subset of the weighted principal components, extracted in the order of the size of the eigenvalues from the weighted covariance or correlation matrix which is not possible with conventional SVD. This is because the SVD solution does not use or calculate a correlation or covariance matrix but it identifies the principal components directly from a sequence of feature vectors. Thus, it can be difficult to engage any weighting.

A weighted correlation or covariance matrix can be determined by including a weights matrix $\mathbf{W}$ which must be the same size as the feature vectors matrix $\mathbf{X}$. If only the mean of $\mathbf{X}$ is normalised, then its weighted covariance matrix can be achieved as in the following

$$\mathbf{\Sigma}_w = (\mathbf{X} \circ \mathbf{W})(\mathbf{X} \circ \mathbf{W})^T \oslash (\mathbf{W}\mathbf{W}^T), \tag{4.4}$$

where $\circ$ and $\oslash$ indicate the Hadamard product and division, respectively. The weighted correlation matrix $\tilde{\mathbf{\Sigma}}_w$ can also be determined using (4.4) if the variance of $\mathbf{X}$ is normalised.

Weighted covariance and correlation matrices of a set of speech feature vectors are found to be real and symmetric. The methodology described here considers the weighted correlation matrix $\tilde{\mathbf{\Sigma}}_w$ and can be equally applied to the weighted covariance matrix $\mathbf{\Sigma}_w$. The size of $\tilde{\mathbf{\Sigma}}_w$ is $D \times D$, where D is the feature dimensionality. Consider the following eigendecomposition formula

$$\tilde{\mathbf{\Sigma}}_w \mathbf{p}_w = \gamma_w \mathbf{p}_w. \tag{4.5}$$

where $\mathbf{p}_w$ is the weighted principal component associated with eigenvalue γ_w. The learning algorithm for determining $\mathbf{p}_w$ using the RNN is now described.

The structure of the RNN requires two layers: a variable layer and a constraint layer, see Fig 4.3. The number of nodes in each layer is equal to D. Both layers are fully interconnected with the weights being the elements of the weighted correlation matrix $\tilde{\mathbf{\Sigma}}_w$. The initial input to the neurons of the variable layer can be the values of a random column vector ($\mathbf{v}^{(1)}$).

Let $\mathbf{C} = \tilde{\mathbf{\Sigma}}_w$, the output of the neurons of the constraint layer at iteration t will be

$$\boldsymbol{\gamma}^{(t)}(j) = \sum_{i=1}^{D} \mathbf{C}_{ij}\mathbf{v}_i^{(t)} \quad \text{for} \quad j = 1,2,...,D, \tag{4.6}$$

where $i, j = 1,2,...,D$ are, respectively, the rows and columns of $\mathbf{C}$. This describes the so called feed-forward step. Now let $\boldsymbol{\gamma}^{(t)}$ be a column vector of the values of $\boldsymbol{\gamma}^{(t)}(j)$ arranged from $j = 1$ to D.

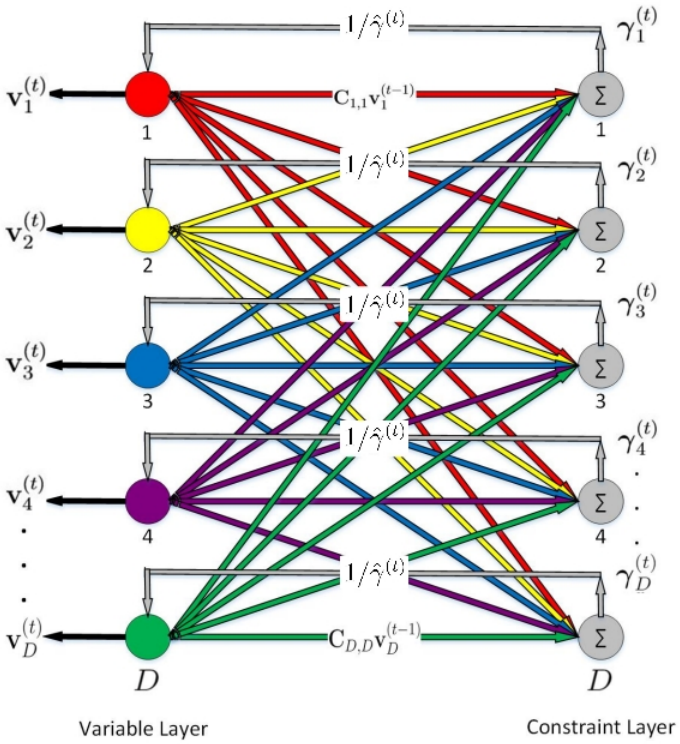


Fig. 4.3 The topology of the RNN network solution for the eigenvalue problem. Two examples of the feed-forward step, expressed by equation (4.6), are given here for clarification. The weights of the links between all $\mathbf{v}_i^{(t)}$ and $\gamma_1^{(t)}$, for $i = 1,2,...,D$, are the elements of the first column of $\mathbf{C}$ which are $\mathbf{C}_{i,1}$, for $i = 1,2,...,D$. The weights of the links between all $\mathbf{v}_i^{(t)}$ and $\gamma_2^{(t)}$, for $i = 1,2,...,D$, are the elements of the second column of $\mathbf{C}$ which are $\mathbf{C}_{i,2}$, for $i = 1,2,...,D$. The rest of the elements of $\boldsymbol{\gamma}^{(t)}$ are obtained in the same way.

In the feedback step, the neural links from the constraint layer to the variable layer are $1/\hat{\gamma}^{(t)}$, where

$$\hat{\gamma}^{(t)} = \max\left(\boldsymbol{\gamma}^{(t)}\right), \tag{4.7}$$

which can also be considered to be the eigenvalue at iteration t. For the next iteration, the input to the variable layer will be

$$\mathbf{v}^{(t+1)} = \frac{\boldsymbol{\gamma}^{(t)}}{\hat{\gamma}^{(t)}}. \tag{4.8}$$

The process described by (4.6), (4.7) and (4.8) is repeated for κ iterations until the network converges to the equilibrium state giving the most dominant eigenvector as

$$\mathbf{v}^{(\kappa)} = \frac{\boldsymbol{\gamma}^{(\kappa-1)}}{\hat{\gamma}^{(\kappa-1)}}. \tag{4.9}$$

Using the variables of the learning algorithm, one can re-write equation (4.5) as

$$\mathbf{C}\mathbf{v}^{(\kappa)} = \hat{\gamma}^{(\kappa)}\mathbf{v}^{(\kappa)}. \tag{4.10}$$

The outcome of the operations on both sides of (4.10) is a column vector. The objective of the learning algorithm here is to minimise a parameter α defined as

$$\alpha = \left|\left|\mathbf{C}\mathbf{v}^{(\kappa)} - \hat{\gamma}^{(\kappa)}\mathbf{v}^{(\kappa)}\right|\right|_1. \tag{4.11}$$

In order to meet the learning objective, i.e. making the value of α approach zero, the learning algorithm must be sufficiently iterated. This will be demonstrated shortly.

It can be seen from (4.5) that the real symmetric matrix, $\tilde{\mathbf{\Sigma}}_w$, scales the eigenvector, $\mathbf{p}_w$, by the eigenvalue, γ_w. If the largest element of $\mathbf{p}_w$ is equal to one then the largest element of $\gamma_w\mathbf{p}_w$ is equal to the eigenvalue γ_w. One can notice that the calculation in (4.6) estimates the right hand side of (4.5), $\gamma_w\mathbf{p}_w$, given the parameters of its left hand side, $\tilde{\mathbf{\Sigma}}_w$ and $\mathbf{p}_w$. By comparing (4.7) and (4.8), one can infer that the maximum value of any $\mathbf{v}^{(t)}$, for $t > 1$, is equal to one. This justifies the calculation of $\hat{\gamma}^{(t)}$ using (4.7) since $\boldsymbol{\gamma}^{(t)}$ is equivalent to $\hat{\gamma}^{(t)}\mathbf{v}^{(t)}$.

The dominant weighted principal component, $\mathbf{p}_w$, of $\tilde{\mathbf{\Sigma}}_w$ is given here by normalising the dominant eigenvector, $\mathbf{v}^{(\kappa)}$, to unity

$$\mathbf{p}_w = \frac{\mathbf{v}^{(\kappa)}}{||\mathbf{v}^{(\kappa)}||}, \tag{4.12}$$

and the associated eigenvalue is now calculated using $\mathbf{p}_w$ and $\tilde{\mathbf{\Sigma}}_w$, as follows

$$\gamma_w = \mathbf{p}_w^T\tilde{\mathbf{\Sigma}}_w\mathbf{p}_w. \tag{4.13}$$

The rest of the weighted principal components are determined as follows. The variance captured by the first principal component, $\mathbf{p}_w$, is removed from $\tilde{\mathbf{\Sigma}}_w$ as in the following

$$\tilde{\mathbf{\Sigma}}'_w = \tilde{\mathbf{\Sigma}}_w - \mathbf{p}_w \gamma_w \mathbf{p}_w^T, \tag{4.14}$$

then the same pre-described learning algorithm can be applied using $\tilde{\mathbf{\Sigma}}'_w$ as the network weights to obtain the second principal component. This procedure is repeated as many times as required to obtain the desired set of d, $d \leq D$, weighted principal components. However, after the order of weighted principal components exceeds the feature dimensionality, all the variance is captured and the additional principal components represent zero variance as illustrated in Fig. 4.4. It can also be noticed from the figure that the principal components are extracted in the order of the amount of variance they represent. This is particularly useful when only a subset of the principal components is needed and those are required to represent the majority of the variance. In such a case, it would not be required to extract all the principal components then selecting the desired subset by examining their associated eigenvalues.

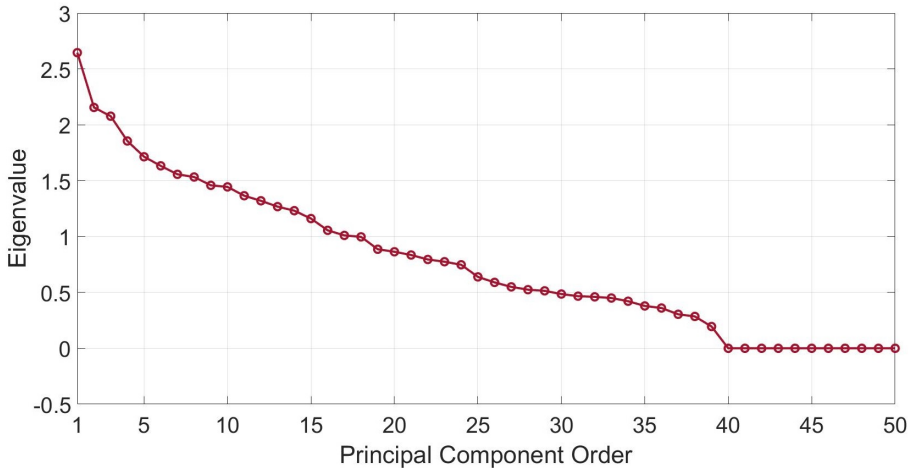
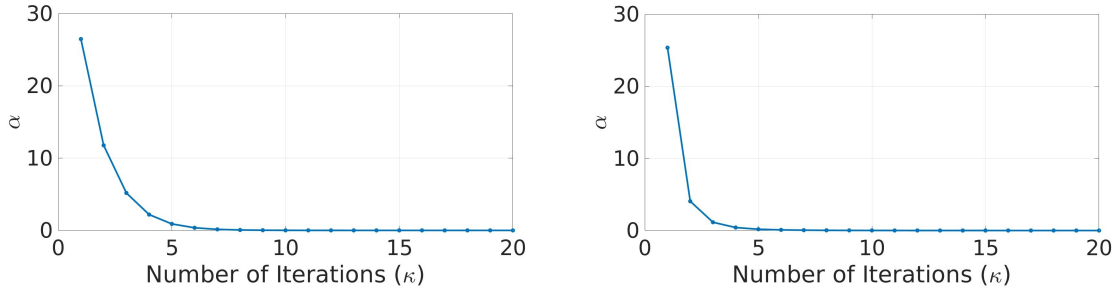


Fig. 4.4 Amount of variance captured by the extracted principal components in term of the eigenvalues. Raw feature dimension is 39. It can be noticed that weighted principal components of order higher than 39 express zero variance.

Fig. 4.5 shows that the proposed RNN solution for the eigendecomposition problem meets the objective of the learning algorithm. One can observe that α approaches zero with a sufficient number of iterations. Fig. 4.5 demonstrates the case when the network input is an arbitrary vector. The use of such an arbitrary vector may not be optimal as it was also addressed in (Delchambre, 2014) with the power iteration method. It was suggested that if

some prior eigenvectors were available, it would then be better to use it to start the iterative process because of the relevance to the problem.



(a) First principal component

(b) Second principal component

Fig. 4.5 Demonstration of how the learning objective, minimising α, of the proposed RNN solution is being met. Examples of the extraction of the first and second dominant principal components.

It is proposed in this work that for every weighted principal component to be extracted, the iterative process is started with the corresponding unweighted principal component determined using SVD. The RNN solution can then be viewed as a process of updating the SVD principal component using the weighted correlation matrix. This strategy can at least increase the convergence rate as discussed in (Delchambre, 2014).

4.2.2 Feature Vectors Weighting Criterion

The weight matrix $\mathbf{W}$ of (4.4) is the same size as the feature vectors matrix $\mathbf{X}$ where the columns of $\mathbf{X}$ are the feature vectors. In this chapter, each column of $\mathbf{W}$ will have the same value for each element such that each feature vector (as a data point) of $\mathbf{X}$ has a different weight. Thus, the weight matrix can be expressed as

$$\mathbf{W} = \begin{bmatrix} w_1 & w_2 & w_3 & \dots \\ w_1 & w_2 & w_3 & \dots \\ w_1 & w_2 & w_3 & \dots \\ w_1 & w_2 & w_3 & \dots \\ \vdots & \vdots & \vdots & \end{bmatrix}. \tag{4.15}$$

Alternatively, each row of $\mathbf{W}$ can have the same value, i.e. each feature has a different weight. The case where each element of $\mathbf{X}$ is assigned a weight requires further investigation which is currently out of the scope of this work.

In this part of the work, each feature vector of the data used to extract the principal components is assigned a weight. This is mainly important in order to decrease the significance of outlying feature vectors and those ones that are noisy or may otherwise represent silence. The proposed weighting criterion here can be described as follows. Using the EM algorithm (Reynolds & Rose, 1995), a GMM is fitted to the feature vectors that are used in the PCA. Then the log-likelihood value of the feature vectors to that GMM can be used directly as weights in this case. With a GMM, Λ, fitted to all the feature vectors of matrix $\mathbf{X}$, each feature vector's weight is then calculated with

$$\mathscr{L}_t = \log\ p(\mathbf{x}_t|\Lambda), \tag{4.16}$$

where $\mathscr{L}_t$ is the log-likelihood of a feature vector $\mathbf{x}_t$.

This criterion is motivated by the concept of acoustic space modelling for speech with the GMM-UBM (Reynolds et al., 2000) and by the methods of model based voice activity detection (VAD) as in (Anguera et al., 2006a). It is therefore anticipated that bad feature vectors will have relatively low log-likelihood values thus lower weights.

All the log-likelihood values are shifted by a scalar amount so that no negative weight is assigned to a feature vector. Let $\mathscr{L}_m$ be the minimum log-likelihood value assigned to a feature vector of $\mathbf{X}$. The non-negative weights of the feature vectors are calculated as

$$\hat{\mathscr{L}}_t = \mathscr{L}_t + |\mathscr{L}_m| \qquad \forall t. \tag{4.17}$$

As a result, the minimum value of $\hat{\mathscr{L}}_t$ is zero (which was the most negative value of $\mathscr{L}_t$). In Delchambre (2014), zero weighted data points were considered to be missing where the power iteration method presented in that work was considered to be more suitable in such a case than conventional PCA methods. It must be noted that the RNN based solution proposed in this work is found to give the same solution as the power iteration method but at a higher convergence rate.

The GMM used here has one component and it is referred to as a Single Gaussian Background Model (SG-BM). One might argue that a GMM-UBM can be used, however, it seems to overfit for this approach. The reason is that higher variability between the weights of the feature vectors can be seen with the SG-BM as illustrated in Fig. 4.6. In the same figure one can see that by using a GMM-UBM, the variability between the weights is less and it decreases by increasing the number of the mixture components.

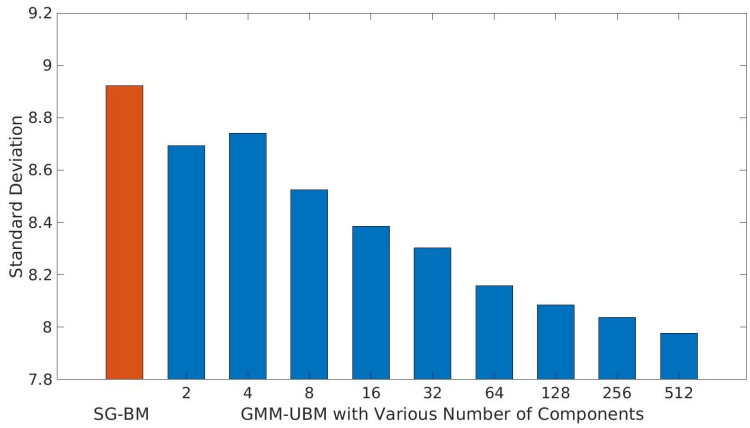


Fig. 4.6 Weight variability in the case of SG-BM versus, the case of GMM-UBM with different number of components.

The weights are associated in the calculation of the weighted correlation (or covariance) matrix using (4.4).

4.2.3 Network Convergence

The power iteration method introduced in Delchambre (2014) was considered a fast solution for weighted PCA as discussed in Section 2.1.3. This method is based on the diagonalisation of the weighted covariance matrix through two spectral decomposition methods. These are power iteration and Rayleigh quotient iteration. In brief, given a weighted covariance Σ_w and a vector of nonzero elements $\mathbf{v}^{(1)}$, the iterative process $\mathbf{v}^{(\kappa)} = \Sigma_w \mathbf{v}^{(\kappa-1)}$ should converge to the dominant eigenvector as $\kappa \to \infty$. However, that method can experience a low convergence rate under some conditions and a further refinement process of the principal components was proposed to allow faster convergence.

The RNN solution for weighted PCA have a higher convergence rate compared to the power iteration method. Fig. 4.7 demonstrates the convergence rates of the RNN method and the power iteration method. The figure shows the number of iterations required to achieve the first weighted principal component of a 39 by 39 weighted correlation matrix. This matrix was obtained with (4.4) for $\mathbf{X}$ using 39 dimensional MFCC features and the weights were obtained using the criterion described previously (Section 4.2.2).

It can be observed that the RNN method converges approximately twice as fast as the power iteration method. The figure also demonstrates that both solutions for weighted PCA are the same. Note that the power iteration process is also initialised here with corresponding unweighted principal component obtained using SVD.

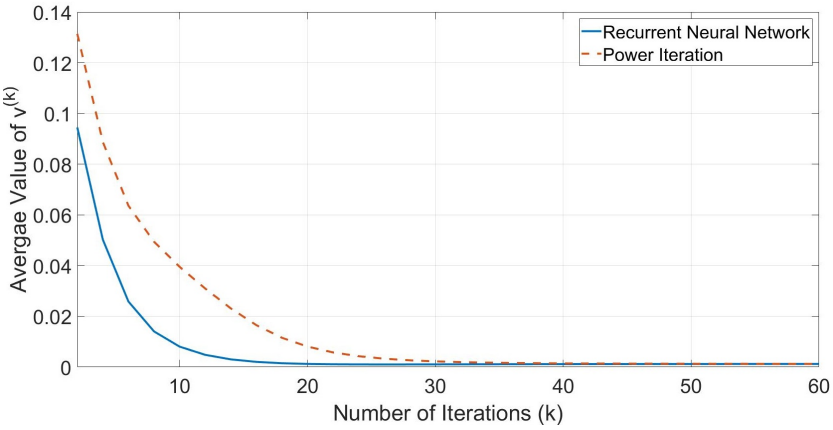


Fig. 4.7 Comparison of the convergence rates of the power iteration method and the recurrent neural network method for extracting the first weighted principal component.

4.3 Evaluation on the i-vector Based Speaker Verification System

This section reports the performance of speaker verification in the i-vector system. The system uses features that are transformed or fused based on the proposed RNN approach for weighted PCA. Note that feature fusion is also, generally, seen here as feature transformation. The performance is reported in terms of the Equal Error Rate (EER). For the sake of comparison, weighted covariance based PCA is also considered and the performance of the system based on the resultant transformed and fused features is provided. Weighted correlation and weighted covariance based PCA are referred to as WCR-PCA and WCV-PCA, respectively.

Additionally, the performance introduced by using classical SVD solution for PCA is also presented. SVD is used to decompose variance normalised features and non variance normalised features which is equivalent to the eigendecomposition of the correlation and covariance matrices, respectively. The unweighted correlation based PCA is referred to as CR-PCA and the unweighted covariance based PCA is referred to as CV-PCA.

The i-vector system parameters and the development data used are the same as the ones described earlier in Section 3.3.1. The evaluation is conducted on the Det5 subset of the 2010 SRE dataset. The data used for performing the principal component analysis is the same as the one used for estimating the GMM-UBM which was also described in Section 3.3.1. Note that the number of speakers of both genders is balanced. The number of iterations of the RNN used for extracting the principal components of WCR-PCA and WCV-PCA is 50.

Each feature and feature combination is normalised using the Cepstral Mean and Variance Normalisation (CMVN) over a sliding window of 3s worth of feature vectors. For the case of covariance based PCA, features and feature combinations of the speech utterances are mean normalised, projected on (multiplied by) the principal components then CMVN over a sliding window is used to normalise the resultant features. For the case of correlation based PCA, features and feature combinations of the speech utterances are subject to mean and variance normalisation, projected on (multiplied by) the principal components then CMVN over a sliding window is also used to normalise the resultant features. For weighted PCA, the weights are only involved in the extraction of the weighted principal components. This is done by associating the weights in the calculation of the weighted correlation and covariance matrices using (4.4). The weights are calculated as described in Section 4.2.2 using the same feature vectors (and feature type) used to estimate the principal components.

Features (Filter Bank)	**Features Dimension**	**EER(%) Splitting**	**EER(%) Splitting plus Adding Channel Effect**
MFCC (24)	39	8.20	3.79
OE-MFCC (28)	76	7.97	3.16
MFCC (24) + LPCC	39 + 39	8.01	3.60
OE-MFCC (28) + LPCC	76 + 39	8.33	3.76

Table 4.1 Effect of using Gaussian noise in data augmentation for different features and feature combinations.

Table 4.1 lists the different sources of features and feature combinations to be used in the experimentation of the RNN approach for WPCA. The table reports some details about the features' parameters. It also highlights the benefits of using the data augmentation method presented earlier (chapter 3) in terms of providing reasonable system performance with the limited development data available.

In Tables 4.2, 4.3, 4.4 and 4.5, d, $1 \leq d \leq D$, indicates the number of the principal components used for the projection of the original features to the new reduced dimension feature space. Thus, d is the resultant feature dimensionality. 'AV' and 'STD' refer to the average and standard deviation of EER, respectively. The amount of variance captured by the reported number of principal components is approximately in the range of 95% to 99% for CV-PCA/WCV-PCA and 85% to 95% for CR-PCA/WCR-PCA. These ranges of variance are found to give the best verification performance in terms of average EER.

The reduction in computation time as a result of reduced system complexity is presented at the end (Section 4.3.3).

4.3.1 Feature Transformation

In this subsection, the effect of the transformation of MFCC and OE-MFCC features using the proposed RNN PCA is investigated. As the number of cepstral coefficients (features dimensionality) is 39 for MFCC, the total number of principal components is also 39. For OE-MFCC, the feature dimensionality is 76 resulting in 76 principal components used for feature transformation.

d	**CV-PCA**	**WCV-PCA**	**CR-PCA**	**WCR-PCA**
39	3.89	4.08	3.82	4.17
35	3.74	2.96	3.51	3.28
34	3.96	3.00	2.96	2.81
33	3.67	2.71	2.97	2.79
32	3.64	3.28	2.91	2.97
31	3.56	3.23	3.34	2.81
30	3.47	3.47	3.21	3.37
29	3.96	3.50	3.63	3.38
28	3.79	3.46	3.31	3.09
27	3.61	3.10	3.64	3.41
26	4.20	3.69	4.12	3.56
AV	**3.76**	**3.24**	**3.36**	**3.15**
STD	0.22	0.29	0.37	0.28

Table 4.2 System performance (in EER%) using transformed MFCC features.

d	**CV-PCA**	**WCV-PCA**	**CR-PCA**	**WCR-PCA**
76	4.52	4.20	4.33	4.41
45	3.87	3.23	3.61	3.10
44	3.46	3.06	3.64	2.80
43	3.04	3.16	3.36	2.52
42	2.98	2.75	3.64	2.75
41	3.46	3.06	3.21	3.21
40	3.36	2.52	3.21	3.10
39	3.42	3.79	2.52	2.28
38	3.18	3.06	2.47	2.36
37	3.26	3.25	2.59	2.25
36	3.21	3.18	2.55	2.47
AV	**3.32**	**3.11**	**3.08**	**2.68**
STD	0.25	0.33	0.49	0.36

Table 4.3 System performance (in EER%) using transformed OE-MFCC features.

The performance of four types of PCA were investigated and the results are listed in Tables 4.2 and 4.3. It can be observed that the performance using CV-PCA is comparable to the reference performance (Table 4.1). In fact, covariance PCA outperformed correlation PCA solely for dimensionality reduction because it gave similar performance in relation to

the one reported in the table even for lower number of components. It is also noticeable that WCV-PCA outperforms CV-PCA, however, WCR-PCA provided the lowest average EER in both cases.

Recall that odd and even filters subsets in OE-MFCC interchangeably capture the speech spectrum. This can in turn cause their cepstral coefficients to be more correlated than conventional MFCC coefficients. Accordingly, these features can, in particular, benefit from the decorrelating effect of a PCA based transformation such as the transformation achieved in this work.

4.3.2 Feature Fusion

The performance of the system is evaluated for two combinations of features fused using PCA. These combinations are MFCC+LPCC and OE-MFCC+LPCC. The performance in terms of EER is reported in Tables 4.4 and 4.5. Similar to single feature type transformation, weighted PCA outperforms unweighted PCA; and PCA of the correlation matrix outperforms PCA of the covariance matrix. The average EER in the fusion of MFCC and LPCC is comparable to that of OE-MFCC which demonstrates the power of OE-MFCC.

d	**CV-PCA**	**WCV-PCA**	**CR-PCA**	**WCR-PCA**
78	4.04	3.92	4.05	4.19
45	4.05	3.10	3.52	2.90
44	3.55	2.79	3.31	3.00
43	3.55	3.07	2.77	2.71
42	3.31	2.79	2.47	2.41
41	3.31	3.05	3.18	2.33
40	3.21	3.33	2.51	2.61
39	3.26	2.97	3.31	2.19
38	2.56	2.40	3.32	2.52
37	3.31	2.79	3.18	2.71
36	3.85	2.95	3.31	2.35
AV	**3.40**	**2.93**	**3.09**	**2.57**
STD	0.40	0.25	0.36	0.26

Table 4.4 EER for fusion of MFCC and LPCC.

WCR-PCA offers the best average performance for all feature combinations studied and this is explained by the following reasons: 1) the use of the correlation matrix instead of the covariance matrix, where the variances of the feature coefficients are normalised thus they have equal contribution to the analysis; 2) using weights for the population feature vectors such that the impact of the outliers is reduced and 3) using an iterative approach for determining the principal components which is found to be more efficient than conventional

approaches (Delchambre, 2014). WCV-PCA also offers an enhancement over CV-PCA, while CV-PCA relatively presented the highest average EER.

d	**CV-PCA**	**WCV-PCA**	**CR-PCA**	**WCR-PCA**
115	6.03	5.72	5.86	5.90
50	3.10	2.68	2.71	2.66
49	3.26	2.32	2.71	2.11
48	2.99	2.59	2.54	1.91
47	3.12	2.36	2.36	2.35
46	2.91	2.22	2.71	1.99
45	2.91	2.19	1.99	2.08
44	3.23	2.50	2.11	1.97
43	2.86	2.22	2.71	1.89
42	3.01	2.01	2.53	1.97
41	2.96	2.97	2.40	2.17
AV	**3.03**	**2.41**	**2.48**	**2.11**
STD	0.13	0.28	0.26	0.23

Table 4.5 EER for fusion of OE-MFCC and LPCC.

A notable aspect of the results presented here is that projection on all the principal components gives a relatively high EER in all cases. This is because having relatively high feature dimensionality corresponds to a higher number of principal components. The higher order principal components will only express a low percentage of the variance. These principal components may not only represent low variance but also noise or other distracting effects embedded in the original features. Projection on these principal components results in sets of attributes that are found to negatively affect the verification performance.

The performance obtained from using all the principal components in all the investigated cases of transformation and fusion supports that explanation. Particularly in Table 4.5, which reports the performance for the case of fusing OE-MFCC and LPCC features, it is evident that using all the principal components (115) gives a high EER compared to the reference performance for the concatenation of OE-MFCC and LPCC features reported in Table 4.1. The reason is that, given the high feature dimensionality of 115 (76 dimensions of OE-MFCC plus 39 dimensions of LPCC), the higher order components, relatively, represent extremely low variance. Another noticeable aspect in the fusion of OE-MFCC and LPCC, is that even CV-PCA shows a considerable reduction in EER. This means that a mere concatenation of different features is not as effective as anticipated especially when the accumulated feature dimensionality becomes relatively high.

Fig. 4.8 depicts the variability in EER in relation to using variable number of principal components in the projection of the features. Compared to the performance of CR-PCA, the EER from WCR-PCA exhibits relatively low variability when using features in different

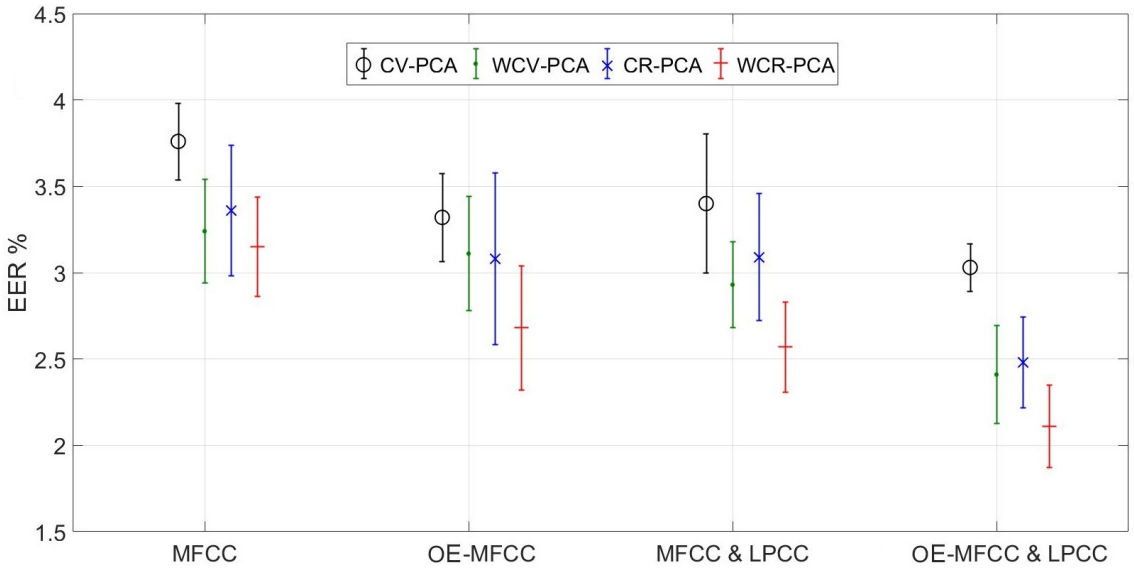


Fig. 4.8 Variability in EER for the overall system performance for all PCA configurations presented for all features and feature combinations.

proportions of dimensionality. This similarly applies to WCV-PCA. The feature vectors weighting process appears to reduce the significance of feature vectors that are more severely affected by noise or those that represent non-speech sounds like breathing.

The best average EER achieved in this work is 2.11%, reduced from 3.76%, using development data that only contained 639 male and 639 female speakers with 5112 utterances in total. Given the limited amount of development data, the relative improvement in the performance is comparable to that of the GMM-UBM/i-vector framework reported in (Khosravani & Homayounpour, 2018) with an EER of 1.13%. Note that in (Khosravani & Homayounpour, 2018), the development data contained 1925 male and 2603 female speakers which enabled a baseline performance of 2.40% EER.

PCA influence on the performance is judged based on the average of the EER values over a range of principal components used for feature transformation. It is difficult to anticipate that the performance at a particular number of principal components will be exactly the same in a different system, for a different evaluation set or with different data used to extract the principal components. However, the average EER exhibits a notable improvement over the performance with non-transformed features.

4.3.3 Computation Time

The complexity of the system described earlier (Section 2.2.2) can be considered to be reduced using the presented methods for feature dimensionality reduction whilst the performance has also been improved. The reduced feature dimensionality reduces the processing time taken by various elements of the i-vector system. The computer used in estimating the computation time here has an Intel Xeon(R) 3.20GHz CPU and 16GB of RAM.

Process	D	d
GMM-UBM Estimation	3.4 ms	2.5 ms
Baum-Welch Statistics Calculation	150 ms	110 ms
Total Variability Learning	1480 ms	940 ms
i-vector Extraction	670 ms	250 ms

Table 4.6 Computation time for the processes affected by features dimension.

The highest original feature dimensionality D was 115 for the concatenation of OE-MFCC and LPCC features. In Table 4.5 we can see that this high feature dimension can be significantly reduced whilst also providing the aforementioned performance improvements. Dimensionality reduction to $d = 43$ is taken in Table 4.6 as a case example to show the reduction in computation complexity. The processing time taken by the estimation of the GMM-UBM is reported per feature vector. Similarly for the total variability subspace training, the processing time is reported per one Baum-Welch statistic supervector. The reported time of calculating the Baum-Welch statistics is for a 150 seconds long speech utterance. The most significant reduction in computation time is in the extraction of the i-vector. This suggests that a variety of speech features can be combined to improve speaker recognition performance using the proposed methodology with relatively low system complexity.

The processing time taken by the principal component analysis is also investigated. Fig. 4.9 illustrates the time taken by the three PCA methods used versus dimensionality of features and feature combinations. The processing time is presented per feature vector. Obviously, the iterative methods (for 50 iterations) consume less time than the classical method used (i.e. SVD). Also, the time taken by SVD increases substantially for each additional dimension. The power iteration and the RNN methods have almost the same processing time. However, it has been shown in Fig. 4.7 that in order for the principal components to converge, RNN requires fewer iterations than the power iteration method. Hence, the RNN method is superior to power iteration in situations where the principal components require a higher number of iterations to converge (for example: 500 iterations) as in (Delchambre, 2014).

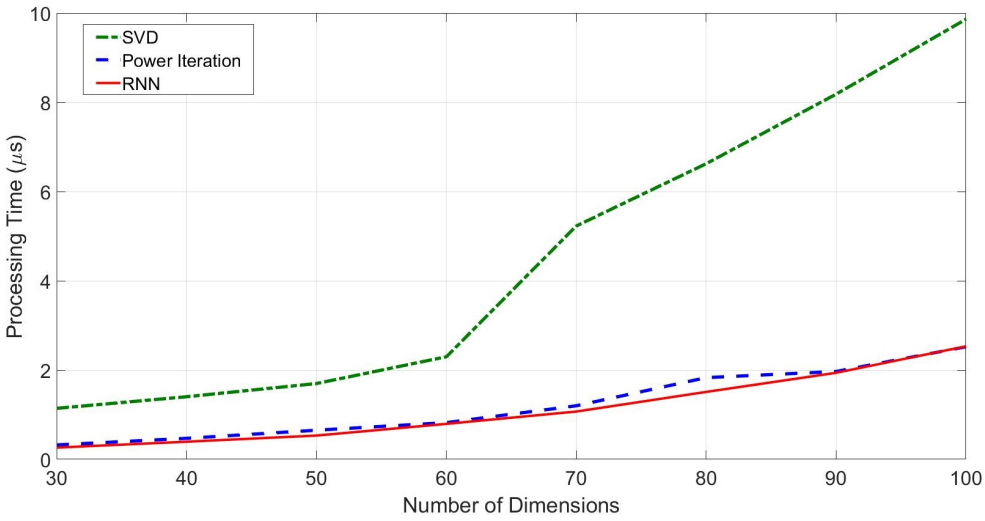


Fig. 4.9 Computation time required to perform PCA using singular value decomposition (SVD), power iteration and the recurrent neural network (RNN).

4.4 Summary

This chapter elaborated on critical aspects regarding the conduct of principal component analysis. First, it was demonstrated how the estimation of a covariance matrix can be affected by the variances of speech features (i.e. cepstral coefficients). It was experimentally shown that a variance normalisation is useful and that PCA based on the correlation matrix is superior to that based on the covariance matrix. The amount of possible outliers in a feature vector set was also demonstrated in this chapter.

The work in this chapter introduced a new RNN-based PCA approach for the eigendecomposition of weighted correlation and covariance matrices. The feature vector weighting criterion presented aims to down-weight outlying feature vectors and those ones that could be distorted by noise and similar affects. In this regard, weighted PCA was shown to outperform unweighted PCA, for dimensionality reduction and feature fusion, in the framework of i-vector based speaker verification.

In comparison to classical SVD and power iteration approaches for PCA, RNN-based weighted PCA framework appeared to be efficient in terms of speed and convergence. The next chapter focuses on speaker diarization and includes some experiments that will demonstrate a broader application of weighted PCA where different features can be assigned different weights.

Chapter 5

Spatial Features and Channel Selection in Binary Key Based Diarization

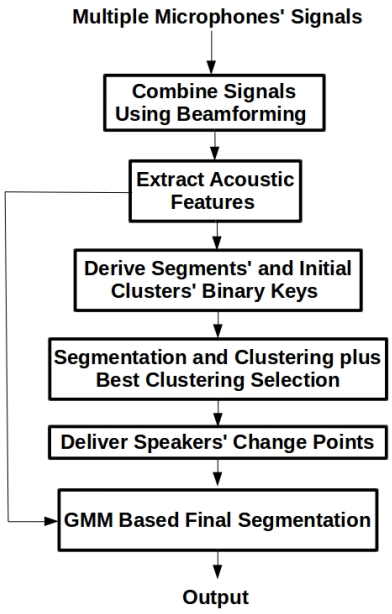
The focus of this chapter moves on to a different aspect of speaker recognition systems which is speaker diarization. For a multi-speaker conversation, an unsupervised diarization system attempts to blindly identify segments of speech belonging to the same speaker. Acoustic features, like MFCC, that are usually extracted from a summation of the microphones' signals are the fundamental front-end processing in diarization systems. This is used for the baseline system (Fig. 5.1a). On the other hand, Time Delay Of Arrival (TDOA) features estimated between the speech signals that are simultaneously received by multiple microphones indicate speakers' locations. Hence, these features can help in the diarization process.

Binary key based diarization[1] is investigated here where the efforts are put into improving the performance of this fast system. It will be shown that a concatenation of MFCC features extracted from all available channels [2] provides better performance than combining signals using the beamforming technique. Since that is computationally expensive, two channel selection methods are introduced to provide cost-effective alternative sources of acoustic features that maintain the improved performance. One of the methods aims to select spatially diverse channels, Fig. 5.1c, and the other aims to select the best quality channels, Fig. 5.1d. This chapter also investigates how TDOA features can be used in binary key based diarization where non-linear transformation is proposed for that purpose, Fig. 5.1b.

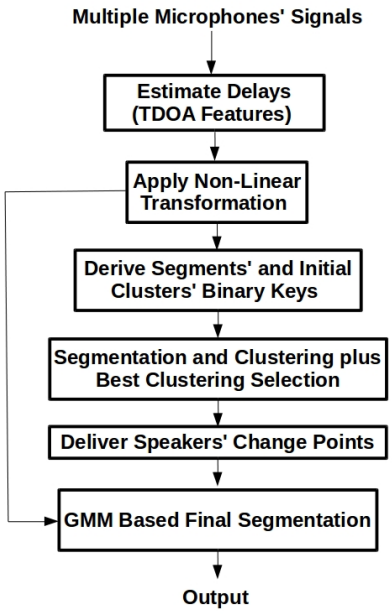
The evaluation section first presents the system performance based on features extracted form the beamformed signals in comparison to features extracted from selected channels. Then it reports the system performance using transformed TDOA features as well as their fusion with all sources of acoustic features mentioned above as illustrated in Fig. 5.2.

[1] Full description of this diarization approach is given in Section 2.3.2.

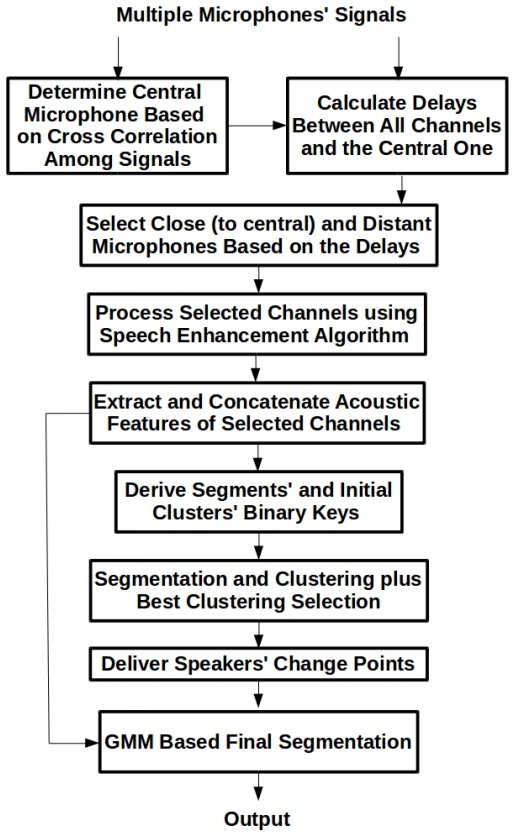
[2] The terms microphone and channel are used interchangeably.



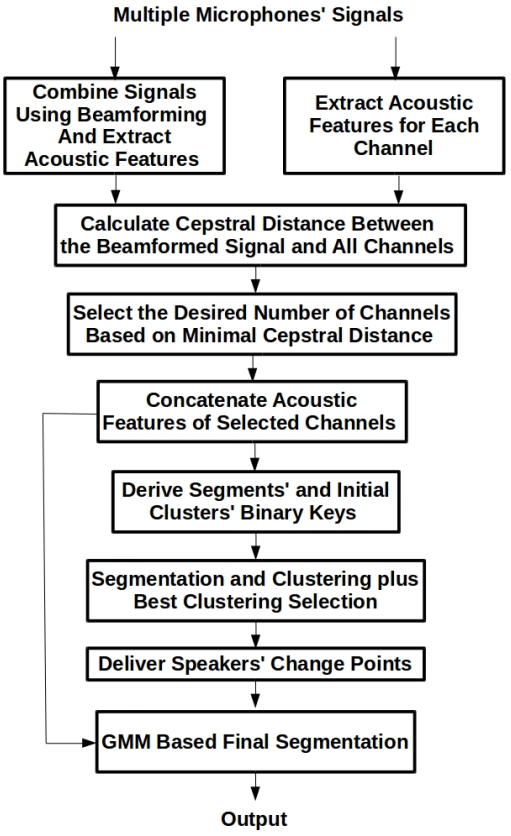
(a) Acoustic features of the beamformed signal. This is the baseline system. Its performance is reported in Section 5.4.2.



(b) Transformed spatial (TDOA) features. A nonlinear transformation, $\mathscr{X} : [0,1]^N \to \mathbb{R}^N$, is identified in Section 5.2.2.



(c) Acoustic features of distant microphones. Detailed description in Section 5.1.1. A parameter, $\tilde{\tau}_{i,ref}$, expressed in (5.2) is used in the selection.



(d) Acoustic features of best microphones. Detailed description in Section 5.1.2. A parameter, $\tilde{\Delta}_{i,ref}$, estimated in (5.6) is used in the selection.

Fig. 5.1 Binary key based diarization using the front-ends proposed in this chapter (apart from the case of Fig. 5.1a which is the baseline system).

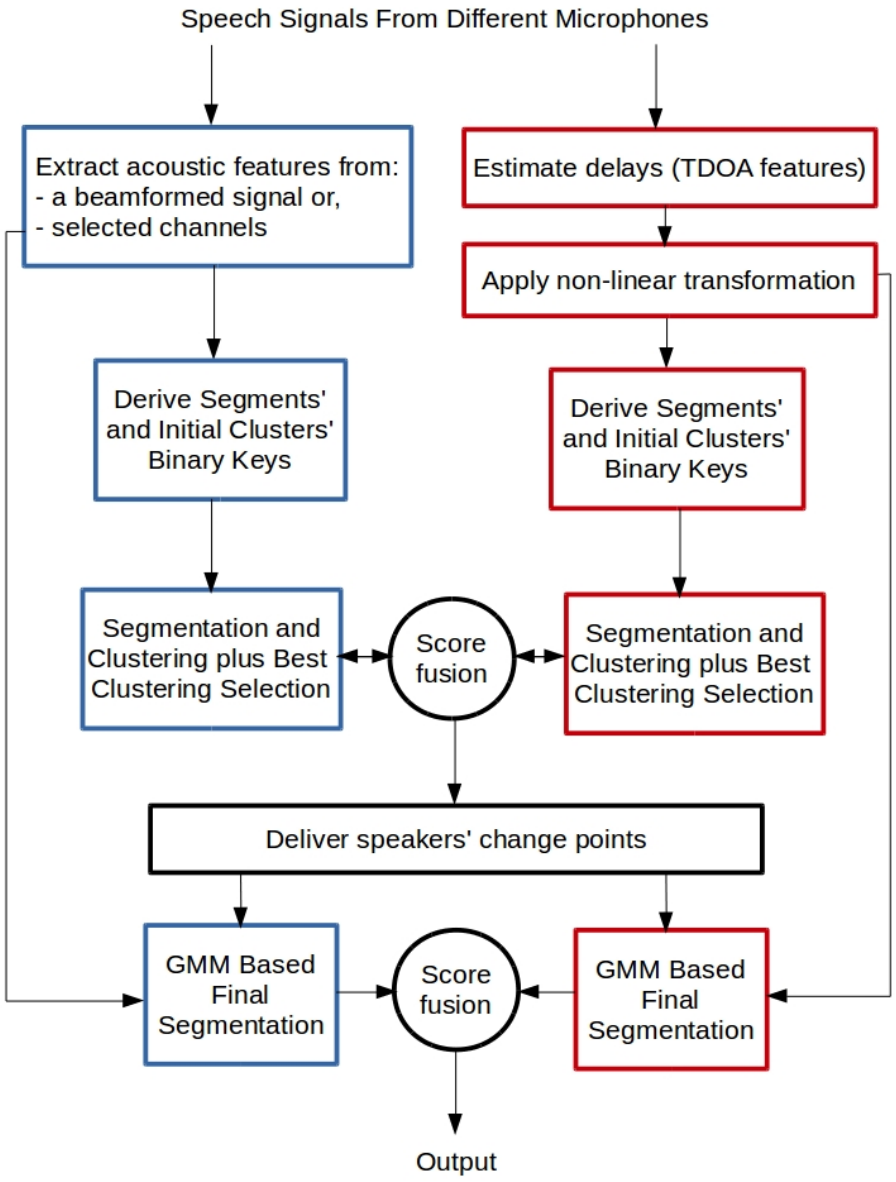


Fig. 5.2 Diagram of the final binary keys-based diarization system of this chapter which integrates acoustic and spatial features.

5.1 Acoustic Feature Concatenation of Selected Channels

This section presents two methods that aim to select suitable channels to concatenate their features in the diarization system. The goal of the first method is to find two sets of channels that are distant from each other. The second method aims at finding the highest quality channels.

5.1.1 Selection of Distant Microphones

The fundamental theory behind this selection method is that when many microphones are available, then those that can deliver a diversity in information are more desired. In other words, if a subset of microphones is to be selected, a desired selection method could be one that identifies the microphones with the redundant information. Selecting microphones that are distant from each other can be useful due to the following aspects:

- In speaker diarization, the feature stream is uniformly segmented whether for the purpose of initial clustering (in uniform initialisation) or segmentation. As a result, the starts and ends of segments do not fall into the correct change points between speakers. This effect can be slightly reduced if a concatenation of features extracted from close and distant channels is used instead of one set of features extracted from one channel (or the beamformed signal).

- A microphone located far from a number of other microphones can be closer to some speakers but not others and that information would be duplicated through the energy pattern of the recorded speech signal. Also, for a relatively distant microphone, the signal sensitivity threshold can be triggered by some speakers more than others.

- Variability in speakers' locations can cause their recorded speech by a particular microphone to be subject to effects of a different nature because of the room impulse response. Additionally, the difference in the effect of the room impulse response on the speech recorded by each microphone is emphasised if the recording microphones are not placed close to each other (Anguera et al., 2007). These undesired disturbances that affect distant microphones differently can actually help in the diarization since a particular speaker's speech is affected by slightly different conditions from other speakers.

The diversity may also, but not necessarily, increase statistical independence within the extracted features especially when diagonal covariance matrices are used in the anchors'

Gaussian models or the GMMs in the final re-segmentation process. Note that diagonal covariance matrices better describe the covariance if the variables (the features) are more statistically independent (Deco & Obradovic, 2012).

Based on the aforementioned theory, a concatenation of features extracted from two distant microphones would present an improvement in the performance as will be shown in the results. In addition, it was found that using two groups of distant microphones can present even better performance. The suitable number of selected microphones for each group will be chosen empirically.

Before beamforming became common practice for combining microphones' signals, the centrally located microphone was usually selected based on cross correlation and was considered a good signal source for the extraction of speech features Anguera et al. (2007). A centrally located microphone is considered a 'close' microphone in the methodology of this section. The distance distribution of other microphones is decided based on signal time delay of arrival at any of the available microphones in relation to the central one.

For this channel selection method, the delays between the signals are calculated using the cross correlation method over segments of 250 ms length and 10 ms shift. For two speech segments from one of the microphones, s_i, and the reference (central) microphone, s_{ref}, the delay τ (the lag) in samples, is the one that maximises the following cross correlation function (see Appendix A.4 for its definition)

$$\mathscr{C}_{(\tau)} = \sum_{n=1}^{N} s_i(n) s_{\text{ref}}(\tau + n), \tag{5.1}$$

where N is the total number of samples within the segment.

Let τ_m be the one that maximises (5.1). The average of the absolute delays over all of the segments of one of the microphones i and the central ref microphone can then be expressed as

$$\tilde{\tau}_{i,\text{ref}} = \frac{1}{S} \sum_{j=1}^{S} \tau_m(s_i(j), s_{\text{ref}}(j)), \tag{5.2}$$

where S is the total number of segments.

For a total number of $\tilde{M}$ microphones, the farthest microphone from the central ref one is selected as the one with the highest $\tilde{\tau}_{i,\text{ref}}$

$$\arg\max_{\forall i} \ \tilde{\tau}_{i,\text{ref}} \quad \text{for} \quad i = 1, 2, ..., \tilde{M} - 1, \tag{5.3}$$

and the closest microphone to the central ref microphone is the one with the lowest $\tilde{\tau}_{j,\text{ref}}$

$$\underset{\forall j}{\arg\min} \ \tilde{\tau}_{i,\text{ref}} \quad \text{for} \quad j = 1, 2, ..., \tilde{M} - 1. \tag{5.4}$$

The first order coefficient of MFCC features reflects the distribution of speech spectral energy between low and high frequencies of speech. It can be used to demonstrate the diversity between distant microphones. Fig. 5.3 shows the distribution of this coefficient extracted from signals of the central, the nearest and the farthest microphones of the IS1001a AMI meeting. One can see that the distribution of this coefficient is similar between the central and the nearest microphone and dissimilar to that of the farthest microphone. This can support the idea that near and far microphones are triggered differently by speakers' speech from the point of speech loudness and speakers locations. On the other hand, it might be difficult to argue that features from distant microphones can enrich statistical independence despite the evident variation in the distributions. The reason is that all microphones are simultaneously making observations about the same event.

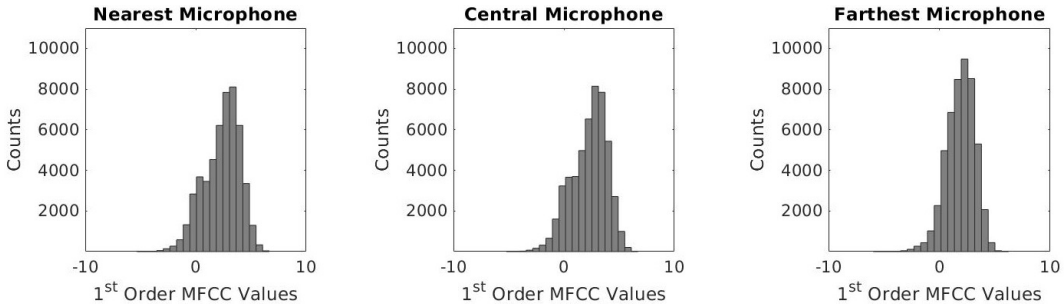


Fig. 5.3 The distribution of MFCC first order coefficient extracted from the speech signal recorded at a central, near and distant microphone.

Fig. 5.4 illustrates the correlations between the first order MFCC coefficient extracted from the signal received at the central microphone and those of the rest of the microphones individually. This is addressed for eight meeting excerpts [1] of the AMI data . Each meeting has four participants and the conversation is recorded by two circular microphone arrays. The first array has 8 microphones and is situated between the four participates. The second array has 4 microphones and it is 1.09 meters away from the first array.

Fig. 5.4 shows the mean correlation over these eight meetings as well as the standard deviation. It can be seen that the correlation decreases as the microphone distance from the central microphone increases. This implies that the diversity in the characteristics of the

[1] Summary of these eight meetings is given in Section 5.4.1 Table 5.1.

recorded speech signal seems to increase with distance. It must be noted that all coefficients of MFCC provide similar correlation pattern. The sudden change in the correlation value between the 7^{th} and the 8^{th} microphones indicates the transition in microphone selection from one microphone array to another.

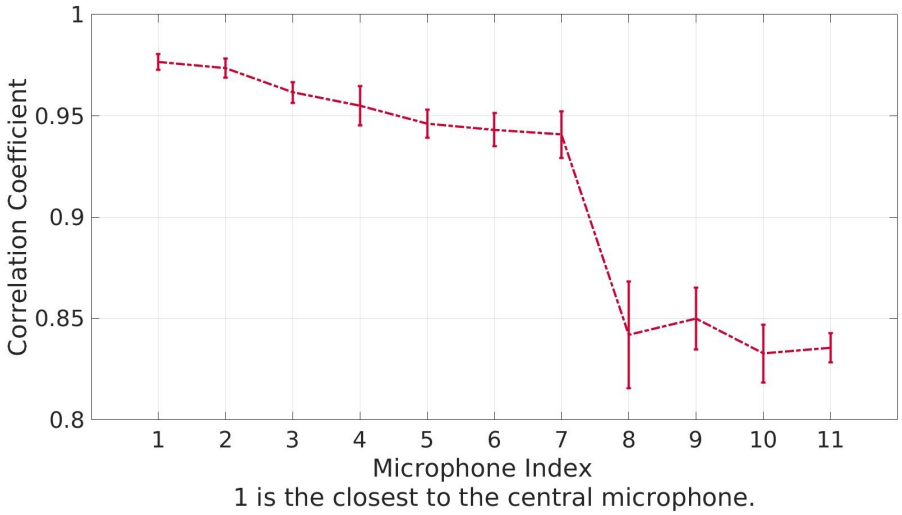


Fig. 5.4 The correlation coefficient of channels' 1^{st} order MFCC cepstral coefficient as a function of distance from the central microphone. The correlation coefficient is determined between the 1^{st} order MFCC coefficient obtained from the central microphone and the one obtained from the first closest microphone, the second closest microphone and so on until the eleventh microphone. The line indicates the mean of the correlation coefficient across eight meetings of the AMI data whilst the error bars indicate the standard deviation of the correlation coefficients across those eight meetings.

This selection method does not consider the quality of the selected channels. Hence, the application of speech enhancement techniques to the microphones' signals, as a pre-processing step, is necessary before speech features are extracted. In this work, a form of Wiener filtration called Two-Step Noise Reduction (TSNR) was used (Plapous et al., 2006).

5.1.2 Selection of Best Quality Channels

The theory behind this selection method is very different from the one behind the selection of distant microphones. This method aims to select one or more of the least distorted channels among the available ones. Reverberation is a considerable source of distortion to the recorded speech especially in meeting rooms and where the recording microphone is distant from the speaker(s). Reverberated speech develops when the speech is reflected off surrounding

objects, like room walls, and overlaps with the original speech at the acquisition (recording) point.

The use of cepstral distance to identify good quality channels requires a reference channel which is assumed to provide a relatively clean speech signal. When information about the meeting room setting is provided, there might be prior knowledge about a particular microphone which can provide a good quality signal that can be used as a reference signal. It can be argued that there might be no need to conduct channel selection in such a case. In practice, however, such information is generally not available and the choice of a reference channel is a difficult task.

The beamformed signal obtained using the method of (Anguera et al., 2007) is proposed to be used as a reference signal here for the following reasons:

1. The segments of the signals to be combined are aligned based on the estimated TDOA delays which strengthens the speech signal and weakens random noise effects;

2. In the summation stage, as described in Section 2.3.3, the signals are weighted according to their qualities (using (2.37)).

Although some aspects were noted on beamforming earlier, the beamformed signal seems to be a good reference choice for the selection of the least distorted channel using cepstral distance. The beamformed signal is an enhanced signal that was found to have better diarization performance in comparison to the signal received at the most centrally located microphone (Anguera et al., 2007). It will be shown in the results that the proposed cepstral distance based channel selection with the beamformed signal as a reference provides channel selection that has better diarization performance than the beamformed signal itself for the development data. MFCC features are used here as the cepstral representation of the speech signal in the cepstral distance calculation. The cepstral distance between two feature vectors is calculated as (Flores et al., 2018)

$$\Delta_{i,\text{ref}} = \frac{10}{\log 10}\sqrt{2\sum_{r=1}^{R}|f_i(r) - f_{\text{ref}}(r)|^2}, \tag{5.5}$$

where $f_i(r)$ and $f_{\text{ref}}(r)$ are MFCC cepstral coefficients of two feature vectors of channel i and the beamformed signal ref, respectively. R is the total number of MFCC cepstral coefficients (feature vector dimensionality). The term $10/\log 10$ in (5.5) is related to the definition of the cepstral distance as the logarithmic spectrum envelop distance (Kitawaki et al., 1982).

Let $\mathbf{X}_i$ and $\mathbf{X}_{\text{ref}}$ denote the entire set of feature vectors extracted from channel i and the beamformed signal ref. The rows of $\mathbf{X}_i$ and $\mathbf{X}_{\text{ref}}$ are the cepstral coefficients of MFCC and the columns are the feature vectors. The average cepstral distance between all feature vectors of channel i and the beamformed signal ref is determined as in the following

$$\tilde{\Delta}_{i,\text{ref}} = \frac{1}{T}\sum_{t=1}^{T}\frac{10}{\log 10}\sqrt{2\sum_{r=1}^{R}|\mathbf{X}_i(r,t)-\mathbf{X}_{\text{ref}}(r,t)|^2}, \tag{5.6}$$

where T is the number of feature vectors.

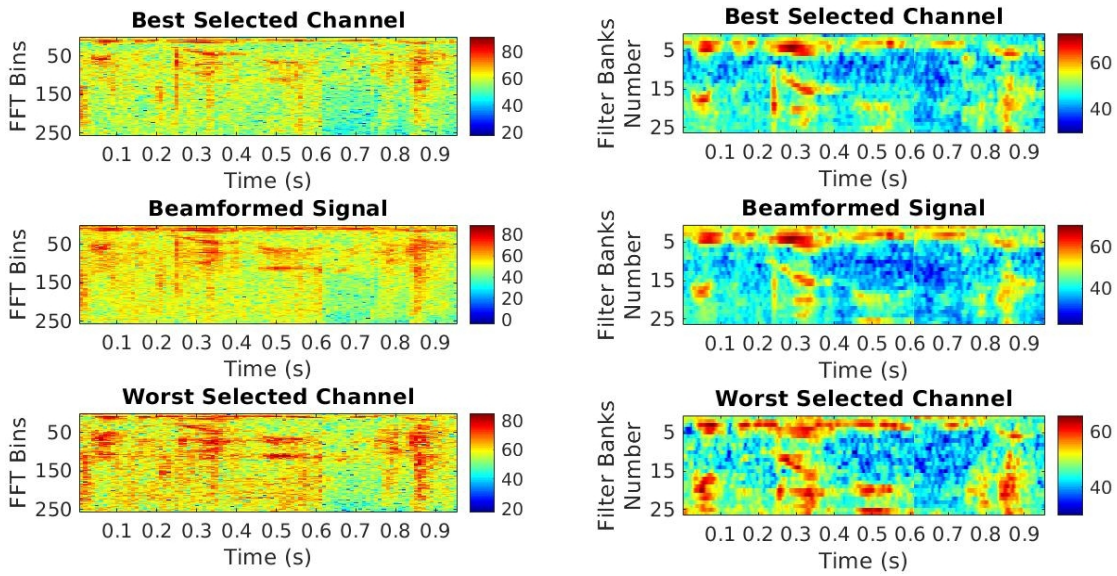
The average cepstral distance between all of the channels and the beamformed signal is calculated. Then the best channel is selected as the one that produces minimal cepstral distance from the beamformed signal ref

$$\text{best channel} = \arg\min_{\forall i} \ \tilde{\Delta}_{i,\text{ref}} \tag{5.7}$$

where $i = 1,2,...,\tilde{M}$ and $\tilde{M}$ is the number of microphones. Fig. 5.5a depicts the FFT spectrum of peer one second segments of speech from the beamformed signal as well as two channels that are selected as the best and worst channels using the proposed method. The meeting example under study is the IS1001a AMI meeting. Before making inferences about the spectrums shown, it should be noted that the regions with the highest values represent strong energy instances in speech phonemes. As such, empty (silent) regions in the spectrum should present the lowest values. These (empty) instances may not have the lowest values in noisy channels, especially if the noise energy is affecting the general spectrum (for example, white noise).

It can be seen in Fig. 5.5a that the channel selected as the worst one exhibits what can be described as distorted speech spectrum compared to the spectrum of the beamformed signal. The dispersion of higher energy regions (spectrum smearing) strongly suggests that this channel experiences a reverberation effect. On the other hand, the best selected channel provides a more distinct speech spectrum than the beamformed signal and the worst selected channel. It can be noticed that the beamformed signal presents higher phoneme energies than the best selected channel likely due to the alignment of segments before summation. However, it seems to have lighter coloured background than the best selected channel which indicates the presence of noise. The probable reason is that the noise is not independent between channels so that the combined effect is not fully cancelled by the weighted summation.

In the process of MFCC feature extraction, the spectrum of Fig. 5.5a is decomposed using a filter bank before calculating the cepstral coefficients. The filter bank spectrum is



(a) FFT spectrum

(b) Spectrum quantified with a filter bank.

Fig. 5.5 Spectrums of one second of speech extracted from the beamformed signal and two channels selected as the best and worst ones. Fig. 5.5a presents quite an informative imaging of the quality of the spectrums. Fig. 5.5b provides additional insights on the filter bank decomposition of the spectrums that will actually be used in the extraction of MFCC features.

shown in Fig. 5.5b. It can be noticed that the spectrums of the best selected channel and the beamformed signal are more alike in this case. The channel selected as the worst one continued to show dispersion of high speech energy.

It is evident that this selection method performs as anticipated. The beamformed signal which is used as a reference signal clearly provides a distinguishable factor between good and bad quality channels based on cepstral distance.

5.2 TDOA Features Fitting in Binary Key Based Diarization

This section addresses the distribution of TDOA features and presents an inference in binary keys system statistics. Then, it identifies a suitable objective transformation of those features. The outcome of this part of the work should enable the integration of spatial (TDOA) features in the binary keys system in order to improve its performance.

5.2.1 Distribution of TDOA Features

The main idea in binary key based speaker diarization resides in the methodology of obtaining a Binary Key Background Model (KBM). The KBM comprises of a number of anchor models which their selection methodology, described in Section 2.3.2, makes the binary keys discriminative since they are derived from projecting speakers' feature vectors onto the same KBM. The speaker characterised by speech segments which are transformed to acoustic features (usually MFCC), is the main subject to detect in a conversation. Success of the anchor models (the KBM) concept, for speaker diarization, in the acoustic feature space poses a question about the possibility of generalising the concept to other feature spaces if appropriate observation about the speakers can be obtained. Especially that, in the case of acoustic features, the anchor models are trained from the conversation's feature segments and not a plethora of features of external speakers.

As mentioned earlier, in the system description (Section 2.3.2), the number of clusters should decrease by one for each step of the clustering process. However, when TDOA features are used in this system, it was observed that the clusters do not constantly decrease by one at each iteration but instead, they rapidly decrease to one cluster after few iterations. The reason is that the binary keys based on a TDOA features space are not sufficiently discriminative.

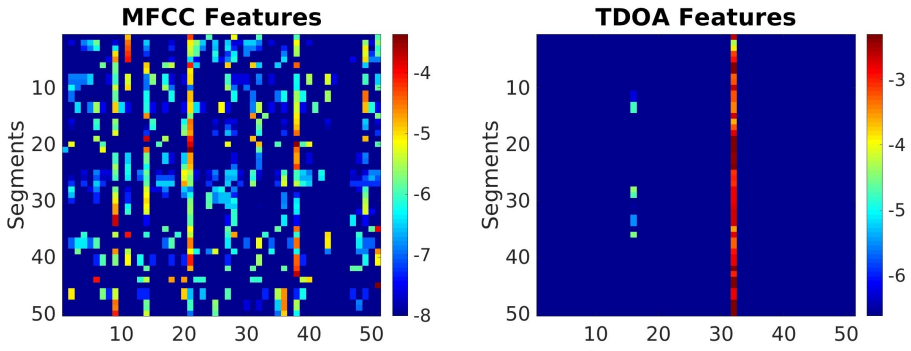


Fig. 5.6 AMI corpus sample, Carletta et al. (2006). These are images of matrices, each row is a section of a segment's cumulative vector. The horizontal axes indicate the indices of the attributes of the cumulative vectors. For the TDOA feature space, one can observe relatively high similarity between the attributes of the cumulative vectors.

For a conversation sample of the AMI corpus, Carletta et al. (2006), Fig. 5.6 shows the attributes of the cumulative vectors. The figure presents a comparison between the case of MFCC features and TDOA features. It can be easily observed that, for TDOA features,

there is very small variability across the segments' cumulative vectors, unlike the case with MFCC features. Hence, the resulting binary keys of the TDOA feature space are found to be less discriminative than those of the MFCC feature space. Similarly, Fig. 5.7 illustrates the situation with a sample of the RT-05S data, Fiscus et al. (2005).

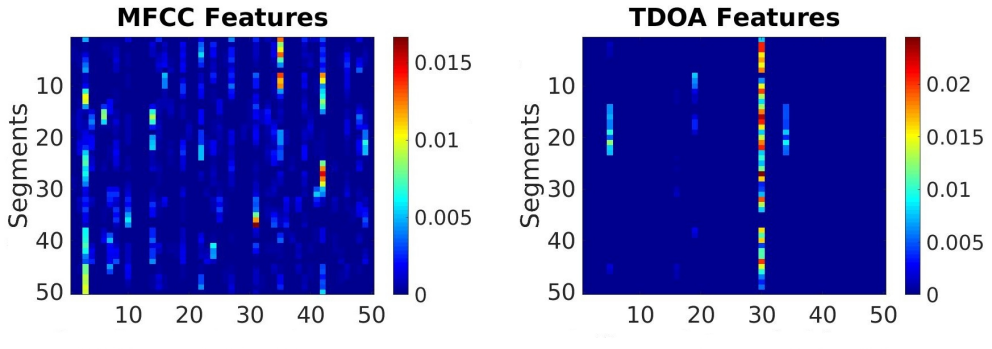


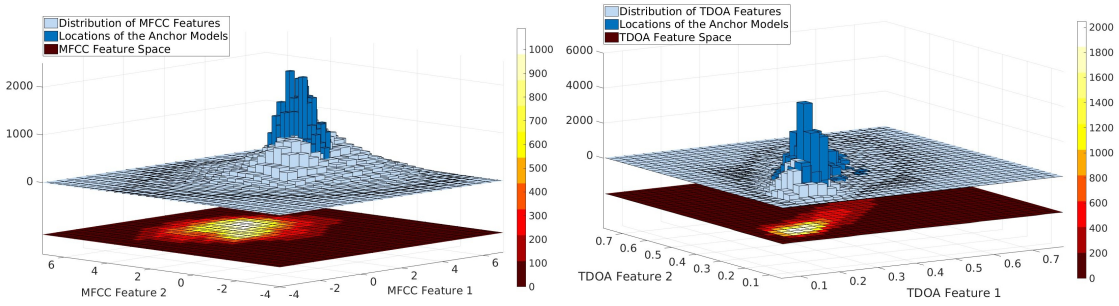
Fig. 5.7 RT-05S dataset sample, Fiscus et al. (2005). These are images of matrices, each row is a section of a segment's cumulative vector. The horizontal axes indicate the indices of the attributes of segments' cumulative vectors.

It can be inferred that, in a TDOA feature space, the anchor models collectively represented by the KBM are not appropriate for binary key based diarization. There are two possible reasons for this; firstly, the methodology of obtaining the KBM is not appropriate for a TDOA feature space and secondly, the statistical conditions of TDOA features make them inappropriate to estimate anchors for the KBM. Since the methodology of selecting the anchor models has shown success in MFCC acoustic feature space, it is more plausible to use it. Instead, inference can be conducted about the selected anchors in the MFCC space and the conclusions can be used to seek possible actions for TDOA features.

Fig. 5.8a shows a plot of 2D MFCC feature space and the means of 896 anchor models that have been selected within this space. The selection process was described in Section 2.3.2. It can be observed that MFCC features have an approximately spherical histogram (skewness = -0.0048 and kurtosis = 2.9218) [1]. In other words, they approximate a normal distribution which has a skewness of zero and kurtosis of 3 (Hoyle, 1995). It can be inferred that the anchor models reside in the area where the data points are more dense which is the centre of the feature space in the case of MFCC. By comparing this with the case of TDOA features illustrated in Fig. 5.8b, it can be observed that for TDOA features, the anchors also reside in the area where the data points are more dense. Hence, one can understand the behaviour of the selection procedure of anchor models. However, it can be noticed that

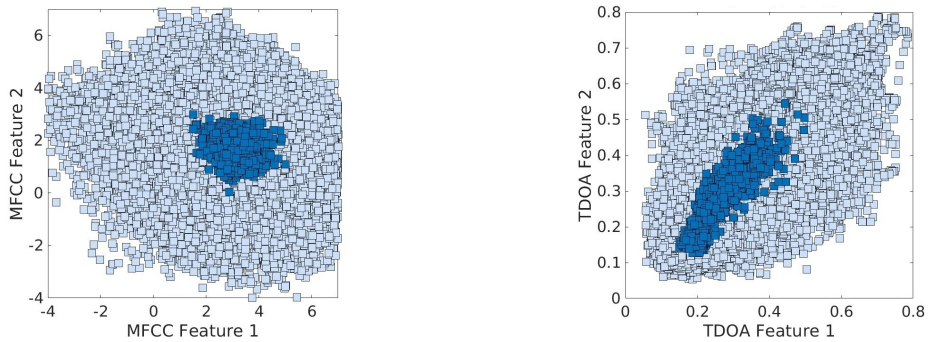
[1] The reader may refer to Appendix A.5 for the meanings of the skewness and kurtosis.

TDOA features exhibit a skewed distribution (skewness = 0.96 and kurtosis = 3.29), hence, the anchors are not located in the centre of the TDOA feature space. Bearing in mind the locations of MFCC space anchors, one can argue that MFCC anchors are suitable because they are located in the centre of the feature space. Thus, the features log-likelihood scores on these anchors are more viable.



(a) MFCC distribution and anchors

(b) TDOA distribution and anchors



(c) Anchors locations in MFCC space

(d) Anchors locations in TDOA space

Fig. 5.8 Top row: features histogram (light blue) and anchor model means (dark blue) on top and the density of the feaure distribution at the bottom. Bottom row: another view to clarify the anchor models locations in their respective feature spaces.

Figures 5.8c and 5.8d can help to complete the picture regarding the locations of the anchor models in the MFCC and TDOA feature spaces. Finally, it can be concluded that the distribution of features affects the estimation of a suitable KBM for deriving discriminative binary keys.

The distribution of TDOA features is usually modelled by a Gaussian Mixture Model (of one component) in the popular Bayesian Information Criterion (BIC) based speaker diarization system. Their skewed distribution might be the reason behind the limitation of their representation with a single Gaussian model as in (Martínez-González et al., 2017; Pardo et al., 2007). A normalisation of TDOA features could be necessary for most diarization systems that use these features.

Let q_n denote a stream of raw TDOA features extracted between a pair of microphones. Let α_s denote the skewness value. The distribution of q_n can be assumed to fit a skew-normal (SN) distribution which was introduced in (Azzalini, 2013) and described in Appendix A.6. The skew-normal distribution of q_n is expressed by SN such that $q_n \sim SN(\tilde{\mu}, \tilde{\sigma}^2, \alpha)$ where $\tilde{\mu}$ and $\tilde{\sigma}^2$ are the mean and variance of q_n, respectively. The skewness of the distribution of q_n is determined by the shape parameter α, such that when $\alpha = 0$ the skewness vanishes ($\alpha_s = 0$) and $SN(\tilde{\mu}, \tilde{\sigma}^2, \alpha)$ becomes $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$.

It is proposed, here, to transform the TDOA features in order to normalise their distribution. The transformed TDOA features will be referred to as TTDOA. The goal is to find a transformation that can alter the parameter α making it approach zero.

5.2.2 Nonlinear Transformation of TDOA Features

Nonlinear transformations can be used to modify the distribution of TDOA. A nonlinear transformation, $\mathscr{X} : [0,1]^N \rightarrow \mathbb{R}^N$ is defined here, which takes an N dimensional vector of the TDOA feature space, $[0,1]^N$, and transforms it to the space $\mathbb{R}^N$ such that the statistical distribution of the vector is normalised. Nonlinear transformations alter the relative distances between the values of the features, thus, they change the shape of their distribution (Weinberg & Abramowitz, 2008). Let the transformation of TDOA features (q_n) be represented by $\tilde{q}_n$.

There are some transformations that can be used to correct the positive skewness of a distribution (Sheskin, 2003). A number of operations are investigated here to identify the suitable transformation $\mathscr{X} : [0,1]^N \rightarrow \mathbb{R}^N$. Those include: the square root $\tilde{q}_n = \sqrt{q_n}$, the logarithm $\tilde{q}_n = \log_{10} q_n$, the reciprocal $\tilde{q}_n = 1/q_n$ and the arcsine (inverse sine) $\tilde{q}_n = 2 \times \text{arcsine} \times \sqrt{q_n}$. The Box-Cox transformation (Box & Cox, 1964), is a power transformation which is also useful and it can be regarded as an adaptive transformation. This is because it covers some of the transformations mentioned earlier depending on a parameter λ_{bc} (the meaning and estimation of this parameter will be explained shortly). The Box-Cox transformation is expressed as follows

$$\tilde{q}_n = \begin{cases} \dfrac{{q_n}^{\lambda_{bc}} - 1}{\lambda_{bc}} & \text{for} \quad \lambda_{bc} \neq 0 \\ \log q_n & \text{for} \quad \lambda_{bc} = 0. \end{cases} \tag{5.8}$$

Regardless of the shift and the scaling by λ_{bc} in (5.8), the Box-Cox transformation can assume an infinite number of forms including the logarithmic [1] transformation when $\lambda_{bc} = 0$, the square-root transformation when $\lambda_{bc} = 0.5$ and the reciprocal transformation when $\lambda_{bc} = -1$.

The main goal here is to find the transformation that best normalises the distribution of q_n. A pure normal distribution can be difficult to achieve as TDOA features are real life measurements. In practice, one can still assume normal statistics if the distribution is unimodal and symmetric (Wu et al., 2010).

5.2.3 Box-Cox Parameter Estimation Based on Local log-likelihoods Maximisation

The main advantage of Box-Cox transformation is that it is not a fixed operation (it is parametric) because the basis of the actual operation of this transformation depends on the parameter λ_{bc}. The value of λ_{bc} that best normalises the distribution of a TDOA feature stream $\mathbf{q} = \{q_1, q_2, ..., q_N\}$ using (5.8) is the one that maximises the log-likelihood as in the following

$$\mathscr{L}_{max}(\lambda_{bc}) = -\frac{1}{2} N \log \hat{\sigma}^2_{\tilde{\mathbf{q}}}(\lambda_{bc}) + \log \mathscr{J}(\lambda_{bc}; \mathbf{q}), \tag{5.9}$$

where $\log \mathscr{J}(\lambda_{bc}; \mathbf{q})$ is the Jacobian transformation, it is equal to $(\lambda_{bc} - 1) \sum \log q_n$ for the special case in (5.8) of the simpler case $\mathbf{q}^{(\lambda_{\mathbf{bc}})}$, and

$$\hat{\sigma}^2_{\tilde{\mathbf{q}}}(\lambda_{bc}) = \frac{r_s(\lambda_{bc})}{N}, \tag{5.10}$$

where $r_s(\lambda_{bc})$ is the residual sum of squares of the variance of $\mathbf{q}^{(\lambda_{\mathbf{bc}})}$.

A modification to the estimation of the transformation parameter λ_{bc} is also proposed here. This modification is useful when there is a need to normalise sub-distributions associated with particular subjects (speakers) in a set of observations (the TDOA features). The task of speaker diarization is to detect the change points between speakers so these change points are initially unknown. Alternatively, $\mathbf{q}$ can be divided into uniform segments and, as these are randomly selected samples, normalising the distribution for each segment will theoretically result in normalising the distribution of all $\mathbf{q}$ as well as the distributions of the speakers' TDOA features. However, the segments cannot be transformed according to different values

[1] The base of the logarithm is not important (Quinn & Keough, 2002), as $\upsilon \log_b(a) = \log_c(a)$, where the constant $\upsilon = \log_b(c)$.

of λ_{bc} because there will be an increased chance that the segments belonging to a particular speaker will have greatly different values (since this is a power transformation).

Therefore, all $\mathbf{q}$ will be transformed based on the same new value of λ_{bc}. Simply stated, λ_{bc} is to be selected according to (5.9) modified to maximise the average of the local log-likelihoods estimate for each segment. Let $\mathbf{q}_s$ represent a segment of $\mathbf{q}$ with size $\tilde{S}$ and let S be the total number of segments. The new selected value of λ_{bc} will be the one that maximises the following average log-likelihood function

$$\tilde{\mathscr{L}}_{max}(\lambda_{bc}) = \frac{1}{S}\sum_{S}\left(-\frac{1}{2}\tilde{S}\log\hat{\sigma}^2_{\tilde{\mathbf{q}}_\mathbf{s}}(\lambda_{bc}) + \log\mathscr{J}\left(\lambda_{bc};\mathbf{q}_s\right)\right). \tag{5.11}$$

This proposed modification may not optimally normalise the distribution of all segments since the objective log-likelihood function of (5.11) is the maximum likelihood estimate of $\boldsymbol{\theta}$ for (5.13). However, the majority of segments will be fairly normalised because λ_{bc} will be selected based on an average log-likelihood. Additionally, an overlap can be allowed between the segments in order to smoothen the local log-likelihood estimates to avoid spikes that may affect the resulting average log-likelihood.

The basis for the estimation of λ_{bc} in the original form of Box-Cox (equation (5.9)) is explained here. Assume that, at the optimum value of λ_{bc}, the transformation of $\mathbf{q}$, denote by $\tilde{\mathbf{q}}$, conforms to the normal distribution theory assumptions with variance $\sigma^2_{\tilde{\mathbf{q}}}$ and expectations

$$E\{\mathbf{q}^{(\lambda_{\mathbf{bc}})}\} = \mathbf{A}\boldsymbol{\theta}, \tag{5.12}$$

where $\mathbf{A}$ is assumed to be a known matrix of dimensions $N \times N$ and $\boldsymbol{\theta}$ is a vector of unknown parameters, associated with $\mathbf{q}^{(\lambda_{\mathbf{bc}})}$. The reader may refer to the skew-normal probability density function of (A.22) for $\mathbf{q}$, $q_n \sim SN(\tilde{\mu},\tilde{\sigma}^2,\alpha)$. To estimate the parameters of the Box-Cox transformation, the probability density function of $SN(\tilde{\mu},\tilde{\sigma}^2,\alpha)$ makes use of the the Jacobian transformation of the normal probability density function of $\tilde{q}_n$:

$$\mathscr{F}_{\mathbf{q}} = \frac{1}{(2\pi)^{0.5N}\sigma^N_{\tilde{\mathbf{q}}}}\exp\left(-\frac{(\mathbf{q}^{(\lambda_{\mathbf{bc}})}-\mathbf{A}\boldsymbol{\theta})'(\mathbf{q}^{(\lambda_{\mathbf{bc}})}-\mathbf{A}\boldsymbol{\theta})}{2\sigma^2_{\tilde{\mathbf{q}}}}\right)\mathscr{J}\left(\lambda_{bc};\mathbf{q}\right), \tag{5.13}$$

where $\mathscr{J}\left(\lambda_{bc};\mathbf{q}\right)$ is the Jacobian transformation expressed as

$$\mathscr{J}\left(\lambda_{bc};\mathbf{q}\right) = \prod_{n=1}^{N}\frac{\partial q_n^{(\lambda_{bc})}}{\partial q_n}. \tag{5.14}$$

For a given λ_{bc}, (5.13) is the likelihood for a least-squares problem. The maximum-likelihood estimates of $\boldsymbol{\theta}$'s, are the least-squares estimates for $\mathbf{q}^{(\lambda_{\mathbf{bc}})}$ and the estimate of $\sigma^2_{\tilde{\mathbf{q}}}$ at a fixed λ_{bc} is $\hat{\sigma}^2_{\tilde{\mathbf{q}}}(\lambda_{bc})$ given in (5.10). Hence, finding optimum λ_{bc} was reduced to maximising the log-likelihood function of (5.9).

5.3 Integration of Acoustic and Spatial Features

This section describes the primary integration framework of this chapter based on score fusion. It also describes the possibility of using WPCA of Chapter 4 to perform RNN-PCA based feature level fusion.

5.3.1 Score Fusion of Independent Binary Key Based Systems

Fusion of the scores of diarization systems that use acoustic features and spatial features almost always result in superior performance in comparison to the case when features are used individually. System fusion is normally a weighted combination of scores. An independent diarization system deals with each type of features and then the decision scores are fused. The BIC based speaker diarization system completely depends on the log-likelihood scores. For all the BIC based system components, the log-likelihood scores are commonly combined with a weight of 0.9 for MFCC features and a weight of 0.1 for the TDOA features Martínez-González et al. (2017).

The binary key based system will include three aspects of fusion where each is expected to be a weighted sum of scores. These scores are: the Jaccard coefficient values in the clustering-and-segments-reassignment phase, the values of within cluster sum of squares (WCSS) in the best clustering selection phase and the log-likelihood values in the final re-segmentation phase.

If the spatial (TTDOA) features weight is denoted by w_s and acoustic (MFCC) features weight by w_a. For $0 < w_a < 1$:

$$w_s = 1 - w_a. \tag{5.15}$$

The score fusion within the three decision stages of the system is described below.

Clustering and Segments Assignment

This includes a weighted sum of the Jaccard coefficients. The method of segments assignment to clusters is first explained. For a particular conversation excerpt, let $\boldsymbol{v}^a_\theta$ and $\boldsymbol{v}^s_\theta$ be two clusters' binary keys of the acoustic and spatial spaces, respectively. Each of $\boldsymbol{v}^a_\theta$ and

$\boldsymbol{v}^s_\theta$ corresponds to the exact same duration of the conversation. Similarly, let $\boldsymbol{v}^a_{seg}$ and $\boldsymbol{v}^s_{seg}$ be two segments' binary keys of the acoustic and spatial spaces, respectively. Each of $\boldsymbol{v}^a_{seg}$ and $\boldsymbol{v}^s_{seg}$ corresponds to the same, but relatively smaller, duration of the conversation. Based on both acoustic and spatial spaces, the segments scores to the clusters is fused according to the following weighted sum of the Jaccard coefficients

$$\mathscr{J}_{seg} = w_a \frac{\sum_{i=1}^{B} \boldsymbol{v}^a_\theta(i) \wedge \boldsymbol{v}^a_{seg}(i)}{\sum_{i=1}^{B} \boldsymbol{v}^a_\theta(i) \vee \boldsymbol{v}^a_{seg}(i)} + w_s \frac{\sum_{i=1}^{B} \boldsymbol{v}^s_\theta(i) \wedge \boldsymbol{v}^s_{seg}(i)}{\sum_{i=1}^{B} \boldsymbol{v}^s_\theta(i) \vee \boldsymbol{v}^s_{seg}(i)}, \tag{5.16}$$

where B is the size of the binary keys, $\vee$ is the boolean OR and $\wedge$ is the boolean AND. Based on maximum values of $\mathscr{J}_{seg}$, every segment is assigned to a cluster, then each space's clusters are re-modelled using the corresponding space's segments.

Next, two clusters are merged if they are found to be the most similar among the existing ones. Let $\boldsymbol{v}^a_{\theta 1}$ and $\boldsymbol{v}^a_{\theta 2}$ be the binary keys of two clusters of the acoustic space. Let $\boldsymbol{v}^s_{\theta 1}$ and $\boldsymbol{v}^s_{\theta 2}$ be the binary keys of the same clusters in the spatial space. Note that the same clusters means that they correspond to the same duration of the underlying conversation. Using both spaces (acoustic and spatial), the similarity between two clusters is determined as in the following

$$\mathscr{J}_\theta = w_a \frac{\sum_{i=1}^{B} \boldsymbol{v}^a_{\theta 1}(i) \wedge \boldsymbol{v}^a_{\theta 2}(i)}{\sum_{i=1}^{B} \boldsymbol{v}^a_{\theta 1}(i) \vee \boldsymbol{v}^a_{\theta 2}(i)} + w_s \frac{\sum_{i=1}^{B} \boldsymbol{v}^s_{\theta 1}(i) \wedge \boldsymbol{v}^s_{\theta 2}(i)}{\sum_{i=1}^{B} \boldsymbol{v}^s_{\theta 1}(i) \vee \boldsymbol{v}^s_{\theta 2}(i)}. \tag{5.17}$$

Afterwards, two clusters are merged if they obtained the maximum value of $\mathscr{J}_\theta$ among the existing clusters.

Best Clustering Selection

The previous process starts with a relatively large number of clusters then, by cluster merging, the number of clusters decreases until it reaches one cluster. However, before any merging takes place after each iteration, the clustering structure is saved. In other words, the segments labels per clusters are saved at each iteration.

As explained in Section 2.3.2, the Within Cluster Sum of Squares (WCSS) is used to identify the best clustering structure, supposedly, the number of hypothetical speakers. The best clustering selection is made using a graphical framework based on the values of WCSS. As previously illustrated in Fig. 2.10, a straight line connects the WCSS values belonging to the case of the highest number of clusters (i.e. the number of initial clusters N_{init}) and the case of one cluster. Then, the values of WCSS for the rest of the clustering structures with different number of clusters Θ, where $1 < \Theta < N_{init}$, are calculated and the plot of those

values forms a curve under that straight line. The point in the curve with the highest distance from the straight line, i.e. the curve's elbow, is identified and it signifies the best clustering structure (hypothetically the correct number of clusters).

In this work, the focus is on integrating spatial (TDOA) features with acoustic features in the framework of binary key based diarization. For best clustering selection, the fusion will also be a weighted sum of WCSS values of acoustic and spatial spaces clusters. Following the previous clustering and segments reassignment step, a particular clustering structure $\mathbb{C}$ will be the same for the acoustic and spatial spaces. This means that at $\mathbb{C}$, there exists the same number of clusters Θ in both spaces where each cluster comprises of the same corresponding segments. The fusion of WCSS values for $\mathbb{C}$ is determined as

$$\mathscr{W}_{\Theta}(\mathbb{C}) = \sum_{i=1}^{\Theta} \left(w_a \sum_{\boldsymbol{g}^a \in \theta_i^a} ||\boldsymbol{g}^a - \tilde{\boldsymbol{g}}_i^a|| + w_s \sum_{\boldsymbol{g}^s \in \theta_i^s} ||\boldsymbol{g}^s - \tilde{\boldsymbol{g}}_i^s|| \right) \tag{5.18}$$

where θ_i^a is an acoustic space cluster, $\boldsymbol{g}^a$ is a segment's binary key within cluster θ_i^a and $\tilde{\boldsymbol{g}}_i^a$ is the cluster's centroid. θ_i^s is the corresponding spatial space cluster similarly the meanings of $\boldsymbol{g}^s$ and $\tilde{\boldsymbol{g}}_i^s$.

Final Re-segmentation

The previous step identified the best clustering (i.e. the number of clusters). Note that there is the same number of clusters in the acoustic and spatial spaces where each cluster is defined by the corresponding segments. A cluster θ in either space, comprises segments that extend for the same duration of a conversation. A GMM is trained for each cluster using the feature vectors covered by its extent. Let $\mathscr{G}^a$ be the GMM for a cluster of acoustic feature vectors covering the extent T. Let $\mathscr{G}^s$ be the GMM for a cluster of spatial feature vectors covering the same extent T. Note that T does not necessarily indicate a set of subsequent feature vectors, however, the feature vectors covered by T are assumed to belong to a particular speaker. Let $\mathbf{x}_t^a$ and $\mathbf{x}_t^s$ be an acoustic and a spatial feature vectors, respectively, each representing the same time extent t of a conversation. For all $t \in T$ and $t \notin T$, the fused score (log-likelihood value) in this final re-segmentation process is determined as

$$\mathscr{L}_t = w_a \log \mathscr{P}(\mathbf{x}_t^a|\mathscr{G}^a) + w_s \log \mathscr{P}(\mathbf{x}_t^s|\mathscr{G}^s) \tag{5.19}$$

where $\mathscr{P}(\mathbf{x}_t^a|\mathscr{G}^a)$ is the posterior probability of $\mathscr{G}^a$ generating $\mathbf{x}_t^a$. The same applies to $\mathscr{P}(\mathbf{x}_t^s|\mathscr{G}^s)$. The values of $\mathscr{L}_t$ are smoothened over a one second window. This indicates the segmentation of one second worth of feature vectors (i.e. a conversation time) per speakers that delivers the final outcome of the diarization system which is 'who spoke when?'. The values of w_a and w_s will be experimentally optimised for each of the three stages of the diarization process as will be shown in the results.

5.3.2 WPCA Based Fusion

To demonstrate broader benefits of weighted PCA, this subsection explains the possibility of fusing acoustic and spatial features using the RNN based technique of Chapter 4. However, no extensive investigation will be conducted in this chapter. The framework of weighted PCA here mainly differs in the weighting criterion as well as the method for achieving the projection of the original feature vectors onto the weighted principal components. After concatenating acoustic and spatial feature vectors (say in a matrix $\mathbf{X}$), each feature's position along the corresponding column of the weight matrix $\mathbf{W}$ is assigned a different weight. Therefore, unlike the case of Chapter 4, the columns of $\mathbf{W}$ are similar but the rows are different depending on the weight assigned to each feature.

Since the principal components will be retained from the same feature vectors to be projected (not feature vectors of many speakers), the scoring (projection) on these principal components can be performed in a least-squares sense as in the following

$$\hat{\mathbf{x}} = (\mathbf{P}_w^T \mathbf{w}^2 \mathbf{P}_w)^{-1} \mathbf{P}_w^T \mathbf{w}^2 \mathbf{x}, \tag{5.20}$$

where $\mathbf{x}$ is the original feature vector, $\mathbf{w}$ is a diagonal matrix of the weights [1] and $\mathbf{P}_w$ is the set of weighted principal components. See Appendix (A.7) for the basis of (5.20).

In the framework of WPCA based fusion, the values of the weights need a single calibration. This will be separately studied in the results.

5.4 Experimental Evaluation and Discussion

This section reports the baseline performance of the binary keys diarization system as well as the performance presented by the proposed methodologies. First the development and evaluation corpora are described. The baseline performance is reported for these corpora

[1] $\mathbf{w}$ is a column of $\mathbf{W}$ transformed into a diagonal matrix. Recall that all columns of $\mathbf{W}$ are the same.

based on MFCC features extracted from the beamformed signal. Then, the performance of the proposed channel selection methods are compared to the beamformed signal. Afterwards, diarization performance using transformed TDOA (TTDOA) features is reported. Finally, the performance for diarization system fusion of TTDOA features with acoustic (MFCC) features is presented. The acoustic features used in the fusion include features from the beamformed signals and features from selected channels separately.

5.4.1 Corpora

The AMI meeting corpus (Carletta et al., 2006) contains a high volume of meeting excerpts in audio and visual forms. Two separate sets of this meeting corpora are selected as development and evaluation data. The development set consists of eight meeting excerpts collected at the IDIAP research institute in Switzerland, it will be referred to here as the AMI development set. The evaluation set consists of another eight meeting excerpts collected at the TNO Human Factors Research Institute in Netherlands, referred to here as the AMI evaluation set.

Excerpt Name	No. of Channels	No. of Speakers	Excerpt Length (s)
IS1001a	12	4	909.14
IS1002d	12	4	1263.10
IS1003a	12	4	913.57
IS1004a	12	4	796.75
IS1005a	12	4	1024.85
IS1006a	12	4	850.60
IS1007a	12	4	965.97
IS1009a	12	4	838.91

Table 5.1 Description of the AMI development set.

Excerpt Name	No. of Channels	No. of Speakers	Excerpt Length (s)
TS3004a	18	4	1345.32
TS3005a	18	4	1318.84
TS3006a	18	4	1252.86
TS3007a	18	4	1609.34
TS3008a	18	4	1352.23
TS3009a	18	4	1505.89
TS3010a	18	4	1041.02
TS3011a	18	4	1509.76

Table 5.2 Description of the AMI evaluation set.

The RT-05S NIST set (Fiscus et al., 2005) is also used in the evaluation and it contains ten meetings. This latter set will be referred to as the NIST evaluation set. One excerpt

of the NIST evaluation set comes from the IDIAP corpus, so it is not included in the AMI development set. Summaries of the different corpora used in this section are reported in Tables 5.1, 5.2 and 5.3.

Excerpt Name	No. of Channels	No. of Speakers	Excerpt Length (s)
AMI200412101052	12	4	943.87
AMI200502041206	16	4	2231.65
CMU200502281615	3	4	1083.43
CMU200503011415	3	4	1208.40
ICSI200105311030	6	7	3642.26
ICSI200111131100	6	9	3429.60
NIST200504121303	7	9	3104.97
NIST200504270939	7	4	2381.98
VT200503041300	2	5	1340.59
VT200503181430	2	5	2663.19

Table 5.3 Description of the RT-05S NIST evaluation set.

5.4.2 Baseline System Performance

This subsection reports the baseline system performance for the corpora under investigation. For convenience of comparison to other achievements in the literature, the system uses traditional MFCC features (Hamming window) extracted from the beamformed signal of each meeting excerpt. All available channel signals are used in the beamforming process. The first best delays selected by the Viterbi algorithm are used in the alignment of the segments. Delays and channel weights are estimated every 250 ms for 500 ms segment size. The MFCC features are extracted from the beamformed signal for speech frames of 25 ms in size at 10 ms rate (every 10 ms). The number of filters in the filter bank are 24 and the number of cepstral coefficients are 19 excluding the 0^{th} order coefficient as it only represents summation of the filters energies.

The size of the KBMs is 896, as was previously optimised in (Anguera & Bonastre, 2011), which is also the size of each binary key. The KBM Gaussians are estimated over subsequent feature segments of 2 seconds in size with 75 % overlap (a rate of 0.5 seconds). In the clustering phase, the sizes of the segments are 1 second extended by ± 1 seconds for a total segment size of 3 seconds. To determine the cumulative vectors, the top 10 log-likelihood scores on the KBM Gaussians are selected for each feature vector. For binary keys derivation from the cumulative vectors, a ratio of 20 % of the highest score positions are set to 1's, the rest are set to 0's. The Jaccard coefficient is used as a similarity metric between the binary keys as in Anguera & Bonastre (2011). The number of uniform initial clusters is 16. In the

final re-segmentation, cluster's (hypothetical speaker's) feature vectors are modelled by a GMM of 128 mixtures. The log-likelihood values of segments' feature vectors to each GMM are smoothened using a one second sized window. The baseline performance is reported in Table 5.4. Minimising the DER and SER is the primary performance objective. As explained in Section 2.3.4, DER indicates speaker time that is not attributed correctly to a speaker and SER is speaker time that is attributed to the wrong speaker. Non-speech feature vectors are identified using the reference files associated with each excerpt in order to precisely evaluate the diarization performance as in (Delgado et al., 2015a).

Dataset	DER (%)	SER (%)	False Alarm (%)	Missed Speech (%)
AMI Development Set	36.41	35.9	25.0	9.4
AMI Evaluation Set	41.25	40.3	37.5	3.1
RT-05S NIST Evaluation Set	30.90	21.3	10.7	28.6

Table 5.4 Baseline System Performance. By the end of this chapter, considerable improvements will be shown compared to this baseline performance that only uses MFCC features extracted from beamformed signals.

5.4.3 System Performance for Acoustic Features Extracted from Selected Channels

This subsection presents a study of system performance in light of the proposed channel selection methods. The extraction parameters of MFCC, as acoustic features, are the same as those of the baseline system. Initially, for the AMI development set, Table 5.5 shows the effect of using a concatenation of features extracted from each individual channel in comparison to those extracted from the beamformed signal.

Signal	Feature Dimension	DER (%)	SER (%)	False Alarm (%)	Missed Speech (%)
Beamformed	19	36.41	35.9	25.0	9.4
All Channels	228	28.06	27.6	21.9	3.1

Table 5.5 Performance comparison between the case of MFCC features extracted from the beamformed signal and a concatenation of MFCC features extracted from each channel for the AMI development set.

A considerable reduction in DER and SER can be noticed in Table 5.5 as an effect of features concatenation. However, as mentioned earlier, concatenation of features comes at the cost of increasing the dimensionality. For the AMI development set, where each excerpt is recorded using 12 microphones, MFCC feature dimensionality grew from 19 (of the single

beamformed signal) to 228 which increased the processing time. Accordingly, the channel selection methods proposed attempts to achieve similar improvement in the performance but with a lower dimensionality.

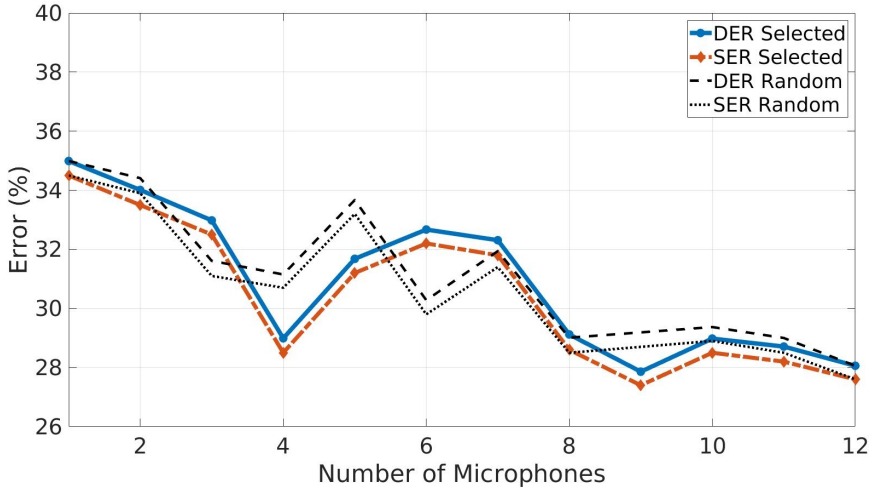
5.4.3.1 Distant Channels

The AMI development set is used to approximate the suitable number of selected channels that provide an improvement in the diarization accuracy. This channel selection method cannot be evaluated on the RT-05S NIST evaluation set because there are four meetings with three channels and less. As such, only the AMI evaluation set is used for testing.

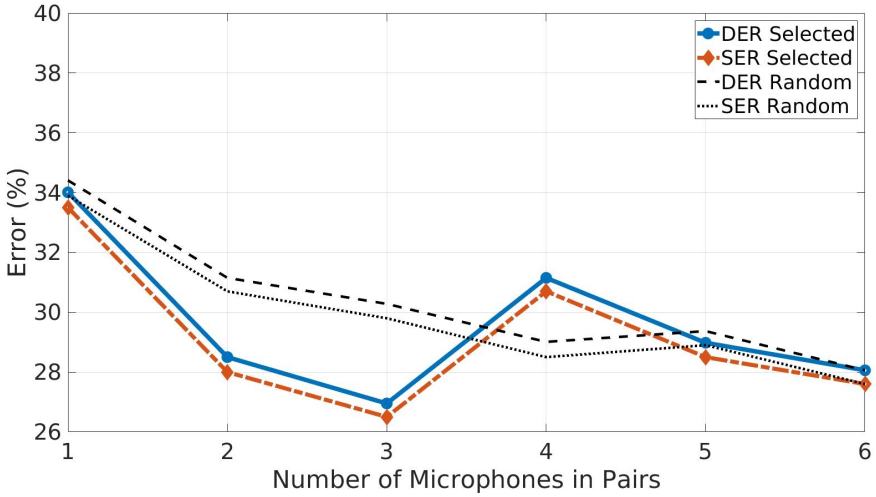
The diarization performance is investigated for two cases of this selection method. The first case is the selection of the most centrally located channel and then additional channels are selected starting from the most distant one from the central microphone. The second where two groups of distant channels are selected. One group includes the central microphone and the ones close to it and the second group is one that is distant from the first group. In this case, the selection also starts from concatenating the features from the centrally located microphone and the most distant one. Then, features of two more channels are added, one of them is the nearest to the central one and the other is the second most distant one, and so on.

Fig. 5.9a demonstrates the system performance in terms of DER and SER in relation to the first case of this selection method. A trade off must be made between dimensionality growth and system performance. It can be noticed that concatenating features of three distant channels in addition to the central one (total of 4) provided equivalent performance to the one when all channels' features are used (Table 5.5). After that, an increase in the errors can be noticed supposedly due to decrease in diversity. Then, as expected, the performance moves toward the one achieved when all channels' features are used. One might notice that the lowest error occurred at nine selected channels, however, it is not a favourable operation point given the high dimensionality of features at that case.

Fig. 5.9b demonstrates the second case of selecting two groups of distant microphones. It can be observed that the choice of three pairs (distant groups of three microphones each) provides an appealing trade off between the performance and the number of channels (six in total). The DER and SER at this point is also lower than the case of concatenating all channels' features. The case of Fig. 5.9b, better demonstrates the achievement of the desirable diversity using this selection criterion. Selecting two distant groups of microphones is assumed to provide more diversity between the channels, compared to the case of Fig. 5.9a, hence better performance is achieved.



(a) Case 1



(b) Case 2

Fig. 5.9 Effect of distant microphones selection (AMI development set) on DER and SER. The figures contrast the performance of the selection criteria (DER Selected and SER Selected) with the case of using randomly selected microphones in addition to the centrally located one (indicated by DER Random and SER Random). Case 1: microphone index 1 means that only features of the central microphone are used. Then features of the rest of the channels are added starting from the most distant channel. Case 2: features of two groups of distant channels are used. Index 1 means the pair of the central and the most distant channels.

It is evident, from Fig. 5.9b, that using a concatenation of features of only one pair of distant microphones provides a decrease in DER and SER by about 2% (DER of 34.01% and SER of 33.5%) compared to the case of using only features extracted from the beamformed signal (Table 5.4). This can also confirm the basis of this selection method. For the AMI evaluation data there was a decrease of about 4% (DER of 37.06% and SER of 36.1%) using features extracted from a central and one distant microphones.

The theoretical behaviour of the plots of Fig. 5.9 was anticipated to be as in the following. A concatenation of a few number of distant channels improves the performance as the diversity is assumed to be high. By adding more channels, the errors are expected to increase as a result of decrease in the diversity. Then the performance is supposed to improve again by adding more channels as it approaches the case of concatenating all channels features. The plots of Fig. 5.9 fairly accommodated the expectations. Sharp changes occurred due to the fact that the DER and SER depend on the outcomes of three components within the system: clustering, best clustering selection and the final re-segmentation. In the process of changing the amount of features, a small variation in one component's outcome can cause non-smooth changes in the subsequent ones.

Tables 5.6 and 5.7 show system performance in light of both cases of the proposed channel selection method at the points that gave the lowest errors based on the AMI development set. By comparing the results in tables 5.6 and 5.7 to those of Table 5.4, one can notice that there is a maximum relative improvement in DER of around 8% on the evaluation set. While the development set experienced a maximum relative improvement in DER of about 25%. It is normal that the relative improvements on the AMI development and evaluation sets are different because of the difference in meetings conditions. However, this proposed selection method provides a cost effective alternative to beamforming. Using features of only one pair of distant microphones, this method presented relative improvements of 6.59% and 10.15% for the AMI development and evaluation sets, respectively.

Dataset	**DER (%)**	**SER (%)**	**False Alarm (%)**	**Missed Speech (%)**
AMI Development Set	28.99	28.5	31.2	9.4
Baseline	36.41	35.9	25.0	9.4
AMI Evaluation Set	38.03	37.1	15.6	3.1
Baseline	41.25	40.3	37.5	3.1

Table 5.6 System performance for the AMI development and evaluation sets as an effect of features concatenation of one central channel and three distant channels.

Dataset	DER (%)	SER (%)	False Alarm (%)	Missed Speech (%)
AMI Development Set	26.95	26.5	25.0	12.5
Baseline	36.41	35.9	25.0	9.4
AMI Evaluation Set	38.87	37.9	6.2	6.2
Baseline	41.25	40.3	37.5	3.1

Table 5.7 System performance for the AMI development and evaluation sets as an effect of features concatenation of two distant groups of channels. Each group has three microphones.

5.4.3.2 Best Quality Channels

The beamformed signal is used here as a reference to assess the quality of the channels using the cepstral distance and based on 19 MFCC coefficients extracted from the speech signals.

The performance of speaker diarization based on acoustic features from channels selected with this method is investigated using the AMI development data. The development data is used to estimate the sufficient number of best quality channels which the concatenation of their features improves the performance. In this method, the channels' signals are not processed by any form of speech enhancement techniques. Therefore, the quality of the channels is assessed using the proposed method without any speech enhancement.

Fig. 5.10 shows the changes in DER and SER in relation to varying the number of best selected channels. Concatenated features of these best selected channels are used in the diarization. In the figure, the diarization performance using features of the first selected best channel is superior to that of the beamformed signal. This channel presented lower DER by about 4%. Then concatenating additional features of more good quality channels decreased the DER and SER further. The increase in DER and SER at 4 channels is caused by a sudden inexplicable increase in the DER and SER of meeting IS1001a. For this meeting, the DER was 39.94% for three channels, then it jumped to 61.34% for four channels and then it returned to 39.00% with five channels. The SER exhibits a similar pattern.

The lowest DER and SER is achieved with five channels. This best performance point is followed by an increase in DER and SER, possibly as a result of adding features from poor quality channels. Then the error decreases again as a result of approaching the performance when all channels' features are used. For comparison, Fig. 5.10 also presents results on the concatenation of features selected from worst quality channels and random channels. Worst quality channels are selected as the ones that have the highest cepstral distance from the beamformed signal. One can notice from the figure that the curve for worst quality channels provides degraded performance in the beginning. Then, it improves when the number of channels increases. On the other hand, the curve of the random selection has no particular trend. In all cases, the performance improves when the number of channels

increases. However, the relatively high increase in dimensionality causes the processing time to increase noticeably.

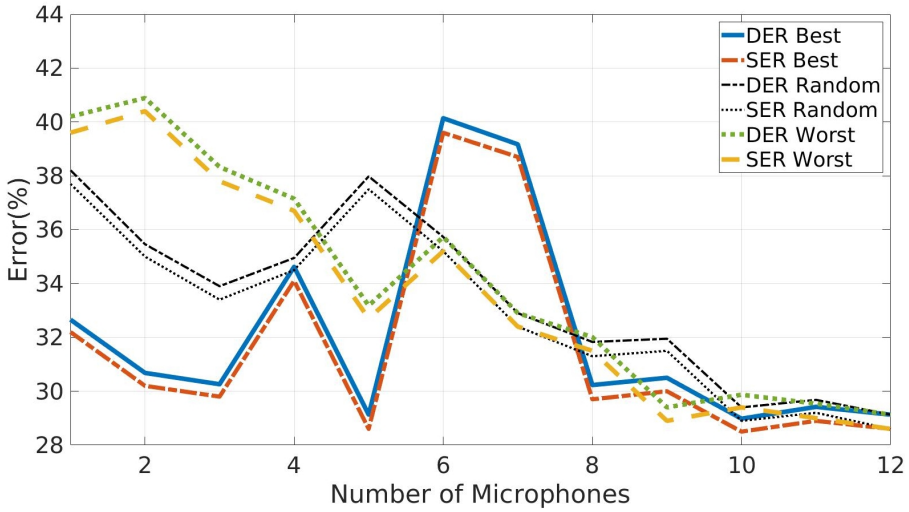


Fig. 5.10 Effect of best quality microphones selection (AMI development set) on DER and SER. An improvement over the case of the beamformed signal can be noticed from 1 to 5. The error increased between 6 and 7 as a result of including lower quality channels, which is expected. Then the error decreases as a natural result of increasing the number of channels and features. DER-Best and SER-Best indicate the cases when the concatenation start from the best channels. DER-Worst and SER-Worst indicate the opposite case. DER-Random and SER-Random indicate random selection.

By using a concatenation of features in the diarization, it can be difficult to tell if this selection method is performing as anticipated. It can be more informative to report the system performance, for illustrative purposes, when features of the selected channels are individually used in the diarization. Fig. 5.11 demonstrates the system performance in relation to using features from individual channels, starting from the best quality channel as selected by this method. The curve of Fig. 5.11 implies that the channel quality estimation process is performing fairly as anticipated. In general, the DER and SER are increasing as the quality of the channel is decreasing. However, the concatenation of features from good quality channels results in lower DER and SER as reported in Fig. 5.10.

Given the similar level of errors for the range of channels 7 to 11 (Fig. 5.11), it can be inferred that the qualities of those channels are similar. Therefore, those low quality channels were not optimally ranked by the selection method minimising the impact on the final results. Thus it appears that the method provides good selection performance, in general, in addition to identifying the worst quality channel (channel 12) which provides the highest error.

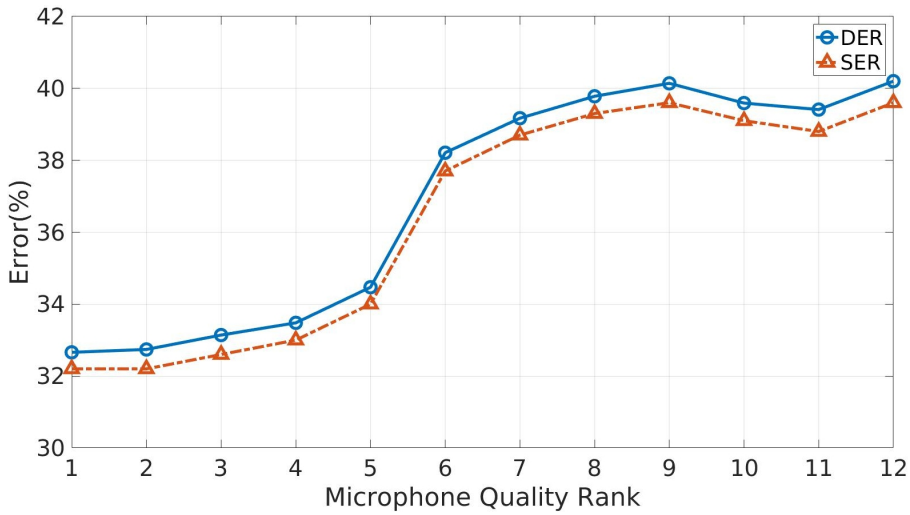


Fig. 5.11 Effect of microphones selection (AMI development set) on DER and SER starting from the best quality microphone. This plot demonstrates the efficiency of this selection method in distinguishing signals' qualities, specifically, for the range between 1 and 9.

In the evaluation, features concatenated from up to five of the best channels are used with the AMI and the RT-05S NIST evaluation sets. Table 5.8 shows the results for these sets. The proposed method of best quality channels selection presents an improvement in DER of about 20% for the AMI development set and 6% for the AMI evaluation set in comparison to the case of using MFCC features extracted from the beamformed signals. For the RT-05S NIST evaluation set, the proposed method introduces a relative improvement of 14.43% in DER and 20.65% in SER.

Dataset	**DER (%)**	**SER (%)**	**False Alarm (%)**	**Missed Speech (%)**
AMI Development Set	29.13	28.6	15.6	9.4
Baseline	36.41	35.9	25.0	9.4
AMI Evaluation Set	38.84	37.9	9.4	6.2
Baseline	41.25	40.3	37.5	3.1
RT-05S NIST Evaluation Set	26.44	16.9	8.9	32.1
Baseline	30.90	21.3	10.7	28.6

Table 5.8 System performance for the AMI development set and for the evaluation sets as an effect of features concatenation of a maximum of five best quality channels.

5.4.4 Integrating TDOA Features in Binary Key Based Diarization

This section investigates the performance of binary key based speaker diarization when TDOA features are integrated. A transformation of these features is needed before their integration

in the system. Fig. 5.12 depicts the resulting distributions from transforming TDOA features using a number of transformations. One can notice that the reciprocal transformation has made it worse and it has increased the skewness and the kurtosis. The arcsine and square-root transformations introduce a moderate normalising effect which did not reach the desired level. On the other hand, the logarithmic and the Box-Cox transformations appear to help normalise the distribution of these features to a great extent. The Box-Cox transformation introduced a skewness of approximately zero and, at the same time, a kurtosis that is very close to that of the normal distribution. In the experimentation, the values of Box-Cox transformation parameter, λ_{bc}, are found to be in the range of -1.75 to 0.61.

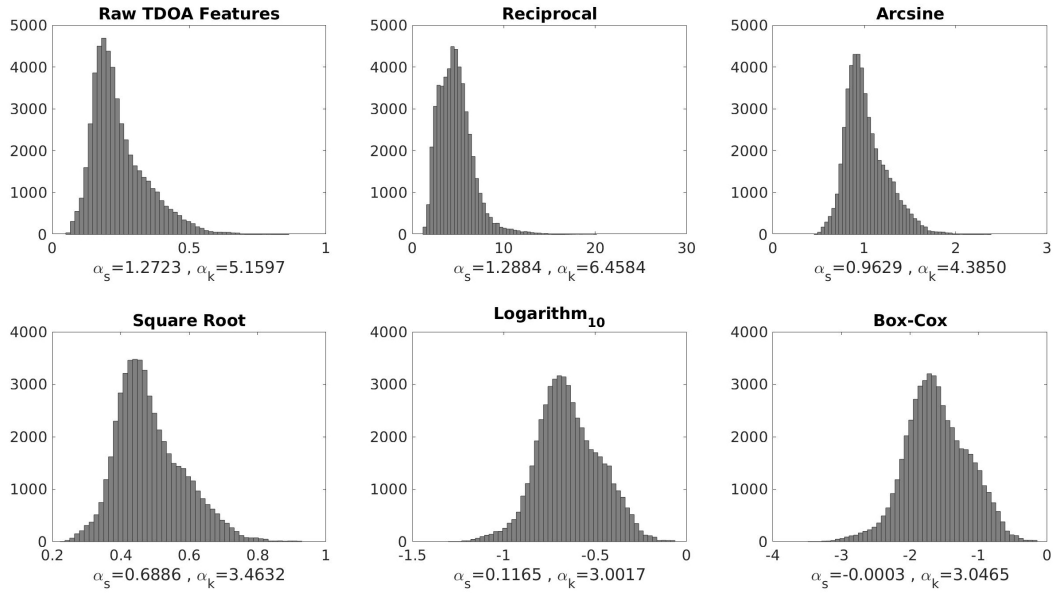


Fig. 5.12 Distributions of transformed TDOA features using the transformations under investigation. The skewness and kurtosis are reported below wherein the captions for each sub-figure identify how the individual transformations affect these parameters. TDOA features are calculated from the IS1001a meeting of the AMI corpus Carletta et al. (2006).

The example of IS1001a AMI meeting used in Figs. 5.12 and 5.13, contains four speakers. As stated earlier, the goal of transforming TDOA features can surpass normalising their overall distribution and instead to normalise TDOA features for each speaker. Fig. 5.13 shows that the distribution of raw TDOA features for each speaker are positively skewed.

Box-Cox transformation of the entire TDOA feature stream also provides a good normalising affect on the distribution of TDOA features for each of the speakers individually, as shown in Fig. 5.13 and as indicated by the included skewness and kurtosis values. However, the local skewness and kurtosis of the speakers' distributions are not the same as that of

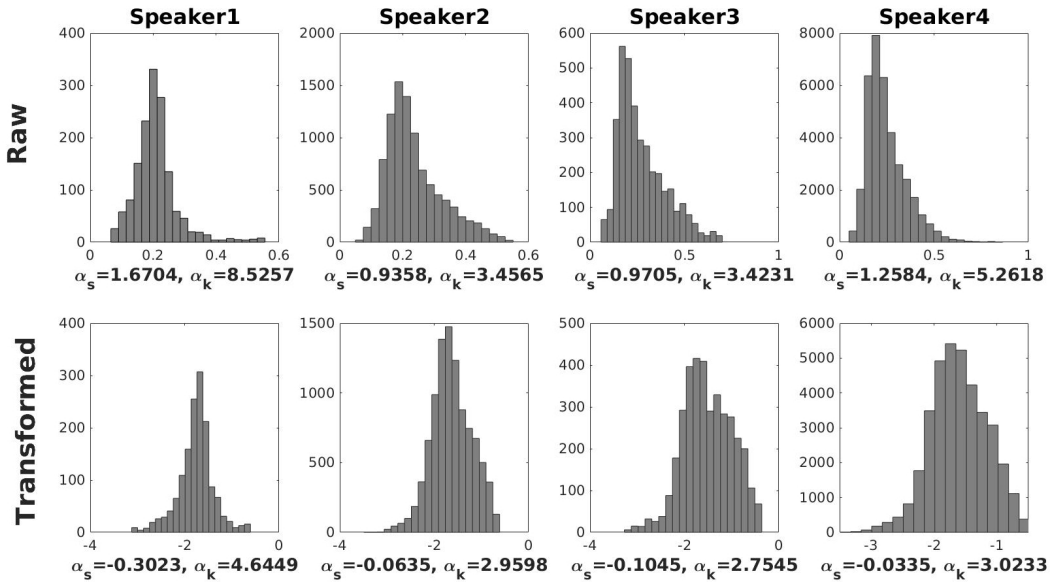


Fig. 5.13 Top row: distribution of raw TDOA features of each speaker of the AMI IS1001a meeting. Bottom row: for the same meeting and microphone pair, the distribution of each speaker's features after performing Box-Cox transformation. The changes in the ditribtuion are precisely described by the skewness and kurtosis parameters below each sub-figure.

the entire stream (as shown in Fig. 5.12). There could be two reasons for this: firstly, this is a natural result since speakers' TDOA are not randomly chosen samples, secondly, the transformation parameter λ_{bc} was selected to be the one that maximises the log-likelihood of equation (5.9) for all the TDOA sequence. As a result, the speaker with the greater contribution to the conversation has a higher effect on the selection of λ_{bc} in relation to the other speakers. For example, one can notice that the original distribution of raw features of Speaker 4 as well as the distribution of speaker 4's transformed features resembles, to a large extent, those of the entire sequence. The reason is that this speaker has contributed to about 70% of the conversation.

In the case of the modified Box-Cox transformation, Fig. 5.14 shows the effect of varying the segment length and that the proposed method can reduce the local skewness of speakers' distributions. It can be seen that using the original Box-Cox transformation, the distribution of Speaker 4 has relatively low absolute skewness because it has the greatest contribution to the conversation and thus the highest impact on the selection of λ_{bc}.

Nonetheless, the standard Box-Cox transformation presented superior performance to the proposed modification as will be shown shortly. The reason is that the Box-Cox transformation has a better normalisation effect in terms of the distribution of the entire TDOA sequence

and consequently the skewness and kurtosis. While the modification targeted the normalisation of the distribution of each speaker's features, the distribution of transformed TDOA (TTDOA) became less normal compared to the one presented by the standard Box-Cox.

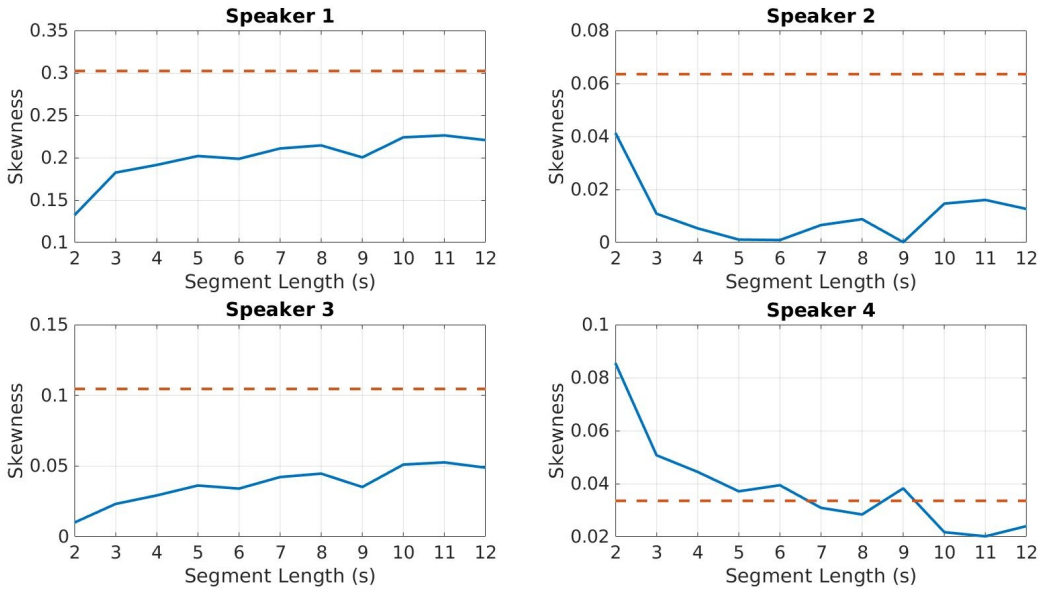


Fig. 5.14 The local absolute skewness of speakers' distributions in relation to segment length in the modified Box-Cox transformation. The dotted lines represent the local absolute skewness of the speakers distributions as a result of the standard Box-Cox transformation.

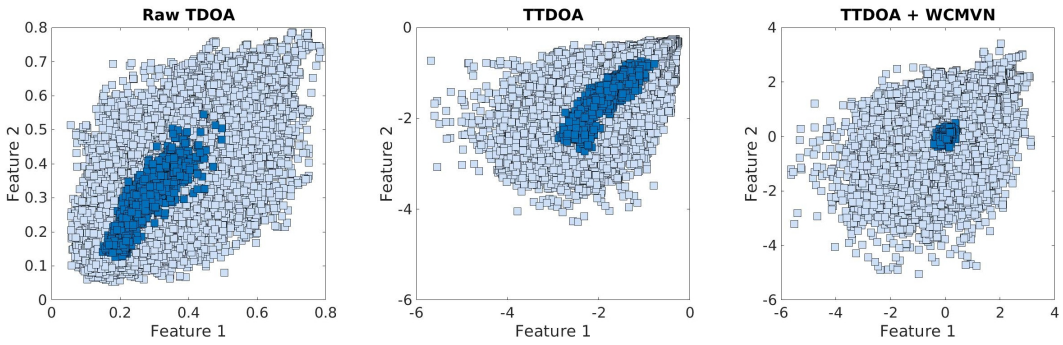


Fig. 5.15 Locations of the anchor models in the feature space. A comparison between raw TDOA, TTDOA and TTDOA further processed by mean and variance normalisation over a sliding window (WCMVN).

Fig. 5.15 shows the effect of using the Box-Cox transformation on the locations of the selected anchors in the TDOA feature space. For TTDOA, the feature distribution became more spherical and the locations of the anchor models are more centred. For illustrative

purposes, the figure also shows a case when TTDOA are further processed with mean and variance normalisation over a 3 seconds sliding window (WCMVN). This additional normalisation resulted in the features distribution to be even more spherical and the anchor models to be more concentrated in the centre of the feature space. However, WCMVN is not suitable for diarization and the actions to normalise the distribution of TDOA features appear to be currently limited to the Box-Cox nonlinear transformation.

5.4.4.1 TTDOA Features Based Diarization

The proposed modified Box-Cox transformation aims to take into account the distribution of each speaker. However, as shown in Table 5.9, TDOA features present better diarization performance when they were transformed with the standard Box-Cox transformation. Each excerpt of the AMI development set is recorded using 12 microphones. This provides the TDOA features with 11 dimensions where the delays are computed between each microphone and a central microphone. Table 5.9 shows a performance comparison, based on the development set, when TDOA features are transformed using the standard or modified Box-Cox transformations.

Transformation	**DER (%)**	**SER (%)**	**False Alarm (%)**	**Missed Speech (%)**
Standard Box-Cox	35.27	34.8	15.6	6.2
Modified Box-Cox	39.49	39.0	43.8	21.9

Table 5.9 System performance for the AMI development set with spatial features transformed using standard Box-Cox and modified Box-Cox transformations.

The reason that the standard Box-Cox method appears to be more suitable transformation for the diarization problem is explained as follows. Standard Box-Cox transformation normalises the distribution of the overall TDOA stream of a meeting excerpt. Assume that the TTDOA stream contains all possible TTDOA samples in the confined space of the meeting room; segments of this presumably normally distributed TTDOA stream are random samples that should have similar distributions as that of the overall TTDOA. Hence, the within segment distribution can be more statistically compatible with the procedure of obtaining the KBM as the anchor models are Gaussian models.

For the AMI development set, TDOA features transformed with standard Box-Cox transformation have superior performance in comparison to the case when acoustic (MFCC) features are used (see Table 5.4). This means that if speaker locations are appropriately measured and modelled, they would then provide sufficiently discriminative properties of

the speakers for speaker diarization. However, this can also depend on other conditions, for example, the room setup and speakers' locations.

While the evaluation sets also exhibit acceptable diarization performance using TTDOA features, see Table 5.10, the DER and SER were higher than those when MFCC features are used (Table 5.4). Nonetheless, a notable aspect of using TTDOA features for diarization is that they are slightly better than MFCC features in terms of the false alarm and missed speech errors. The fractional relative improvements regarding the combination of these types of errors were 36.62% for the AMI development set, 46.05% for the AMI evaluation set and 9.16% for the RT-05S NIST evaluation set. The fusion of TTDOA features with MFCC features also present considerable reduction in DER and SER as will be shown next.

Evaluation Set	**DER (%)**	**SER (%)**	**False Alarm (%)**	**Missed Speech (%)**
AMI	49.84	48.9	12.5	9.4
RT-05S NIST	43.35	33.8	7.1	28.6

Table 5.10 System performance on the evaluation sets with spatial features transformed using standard Box-Cox technique.

5.4.4.2 TTDOA and MFCC Features Based Diarization

The integration of MFCC and TTDOA features in binary key based diarization requires two independent systems, each system deals with one of those features. Through the three stages of the system, the scores of both systems are fused in a weighted sum fashion. Recall that those stages are comprised of: clustering and segments assignment, best clustering selection and the final re-segmentation. The AMI development set is used to learn the suitable weights that should be assigned for MFCC and TTDOA features in the score fusion. DER is used as the calibration parameter where the aim is to minimise it by finding the optimum weights. Recall that w_a and w_s denote the acoustic (MFCC) and spatial (TTDOA) features weights, respectively.

It was mentioned in Section 5.4.4.1 that TTDOA features have a noticeable influence on the system performance in terms of false alarm and missed speech errors. Accordingly, one can notice from Fig. 5.16 how using TTDOA features in addition to MFCC features in the clustering and segments assignment phase considerably affect the DER and SER for the AMI development set. The lowest DER and SER is achieved when $w_a = 0.5$, which represents the case when TTDOA and MFCC features are given the same weights, or in other words, the fusion is simply the sum of both systems' scores.

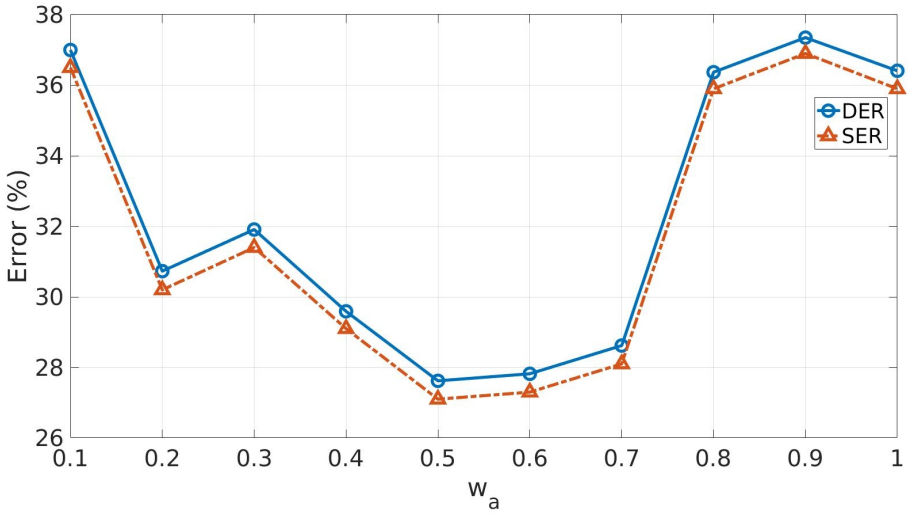


Fig. 5.16 Effect of fusion weights of MFCC and TTDOA features on system performance in the clustering-and-segment-reassignment phase for the AMI development set.

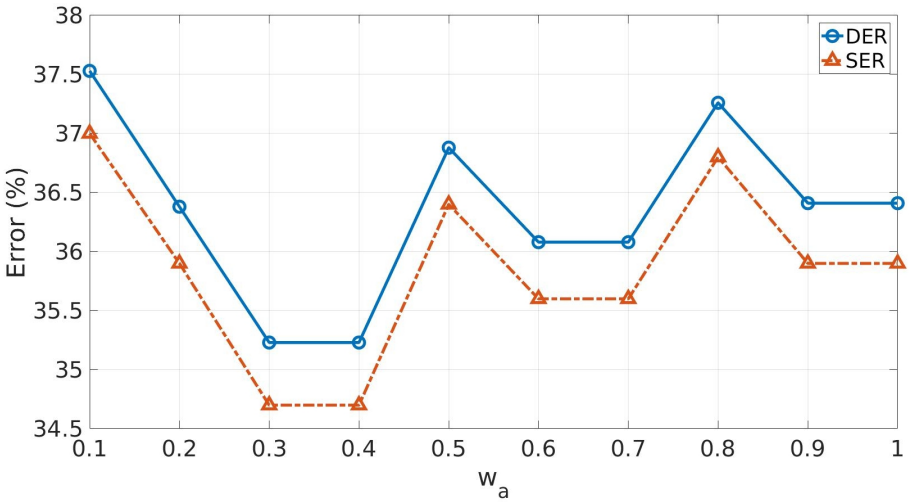


Fig. 5.17 Effect of fusion weights of MFCC and TTDOA features on system performance in the best clustering selection phase for the AMI development set.

Fig. 5.17 illustrates the effect of TTDOA and MFCC features fusion on the process of best clustering selection. The best DER and SER at this phase is achieved when $w_a = 0.3$ and $w_a = 0.4$. However, it can be noticed that there is not an adequate range of stability where the performance is improved over the case of only using MFCC features (Table 5.4). Hence, there is no robust point of weights ratio (between MFCC and TTDOA features) that one can safely choose to use in the fusion and a simple sum of WCSS scores seems to be a good choice. One can also have the decision of the best clustering structure made using only one type of the features.

Results for the final re-segmentation phase are shown in Fig. 5.18. Here, for the AMI development set, the lowest DER and SER are obtained when w_a is less than 0.3. The performance within that range of weights is not as stable as the one for the range above $w_a = 0.3$. In this range, one can notice that the best working points are provided when MFCC features have the weights: $w_a = 0.5$, $w_a = 0.6$ and $w_a = 0.7$. It appears that any of these weights is favourable since the performance is less variant in the vicinity of those particular weights. This conclusion is further confirmed by the plot of Fig. 5.19 which shows results on the AMI evaluation set for various weights in this final re-segmentation phase. One can notice that the weights range $w_a = 0.5$ to $w_a = 0.6$ provided the lowest error with $w_a = 0.5$ falling in the middle of invariant performance as particularly indicated by the SER values.

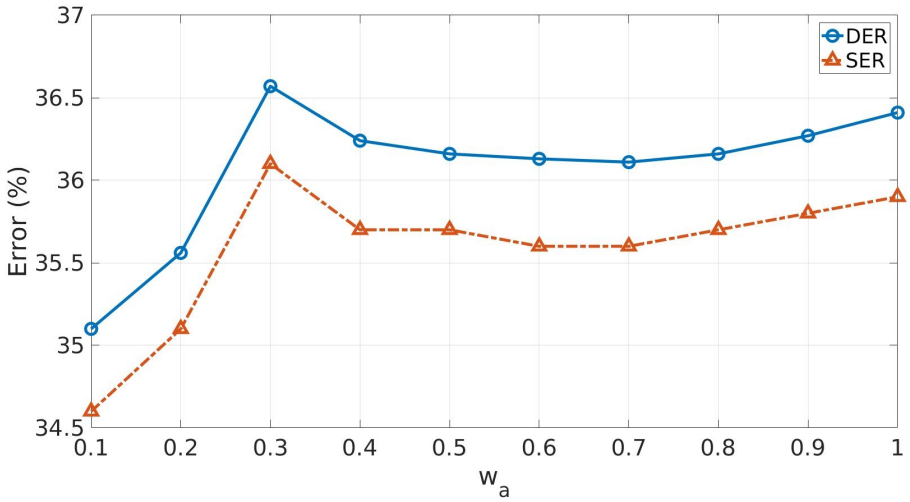


Fig. 5.18 Effect of fusion weights of MFCC and TTDOA features on system performance in the final re-segmentation phase for the AMI development set.

After this investigation, it can be concluded that the system's performance appears to improve when the score fusion is only a simple (unweighted) summation. Furthermore, the fusion in the clustering-and-segment-reassignment stage has the greatest influence on the

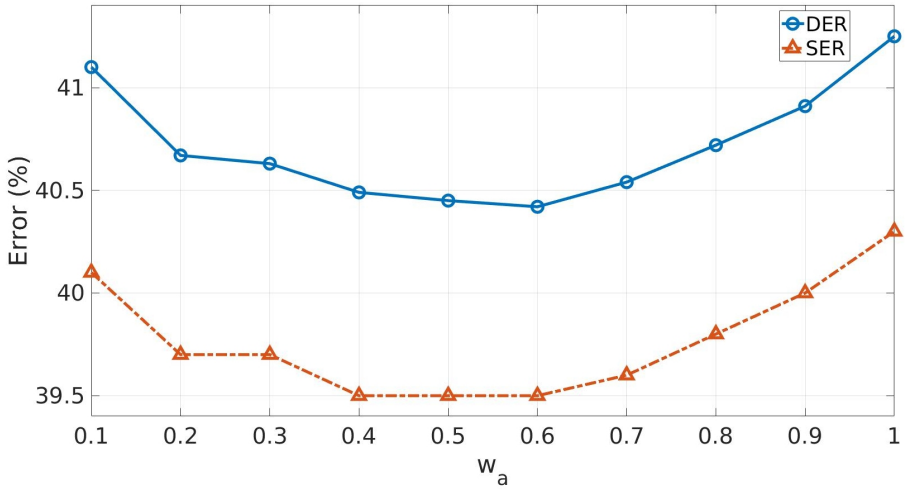


Fig. 5.19 Effect of fusion weights of MFCC and TTDOA features on system performance in the final re-segmentation phase for the AMI evaluation set.

performance and the best performance was achieved at the point when both features are equally weighted, i.e. $w_a = w_s = 0.5$. When the fusion is a simple summation of scores, it provides the advantage of minimised fusion dependencies.

Table 5.11 reports the system performance for the fusion of MFCC features (extracted from the beamformed signal) and TTDOA for the three datasets used in this study. The fusion of these features provided variable amounts of relative improvements in the errors. The least absolute improvement in DER is 4.6% for the RT-05S NIST evaluation set. The absolute improvement in DER for the AMI development and evaluation sets is 10.11% and 8.78%, respectively. The least relative improvement is 14.88% in the DER for the RT-05S NIST evaluation set. For the AMI development set, the relative improvement in DER is 27.76% and it is 21.28% for the AMI evaluation set.

Data Set	**DER (%)**	**SER (%)**	**False Alarm (%)**	**Missed Speech (%)**
AMI Development	26.30	25.8	31.2	6.2
Baseline	36.41	35.9	25.0	9.4
AMI Evaluation	32.47	31.5	15.6	6.2
Baseline	41.25	40.3	37.5	3.1
RT-05S NIST Evaluation	26.30	16.7	3.6	25.0
Baseline	30.90	21.3	10.7	28.6

Table 5.11 Fusion of TTDOA features and MFCC features extracted from the beamformed signals.

Table 5.12, shows the fusion of TTDOA features and concatenated MFCC features of selected distant pairs of channels. The performance provided when only using concatenated features of selected distant pairs was previously shown in Table 5.7. Table 5.13, on the other hand, shows the performance of the fusion of TTDOA features and concatenated features of a maximum of five best selected channels. The performance provided by using concatenated features of selected best quality channels was previously shown in Table 5.8.

Data Set	DER (%)	SER (%)	False Alarm (%)	Missed Speech (%)
AMI Development	24.54	24.0	18.8	6.2
Baseline	36.41	35.9	25.0	9.4
AMI Evaluation	36.99	36.0	21.9	9.4
Baseline	41.25	40.3	37.5	3.1

Table 5.12 Fusion of TTDOA features and concatenated MFCC features of three distant pairs of channels.

Data Set	DER (%)	SER (%)	False Alarm (%)	Missed Speech (%)
AMI Development	27.27	26.8	21.9	9.4
Baseline	36.41	35.9	25.0	9.4
AMI Evaluation	36.64	35.6	14.3	7.1
Baseline	41.25	40.3	37.5	3.1
RT-05S NIST Evaluation	24.87	15.3	7.1	26.8
Baseline	30.90	21.3	10.7	28.6

Table 5.13 Fusion of TTDOA features and concatenated MFCC features of best quality channels (maximum of five).

By comparing the results between Tables 5.12 and 5.7 and between Tables 5.13 and 5.8, it can be observed that fusion with TTDOA features results in an improvement in the performance in all cases. For the case of selecting distant groups of microphones (Table 5.12), the AMI development set experienced further improvement in the performance over the case when MFCC from the beamformed signal is used (Table 5.11). Also, for the case when TTDOA features are fused with MFCC features extracted from selected best quality microphones (Table 5.13), the RT-05S NIST evaluation set exhibited additional improvement in the performance over the case when TTDOA features are fused with MFCC extracted from the beamformed signal (Table 5.11).

The DER and SER of the RT-05S NIST evaluation set for the results reported in Table 5.13 represent state-of-the-art performance for this dataset using this fast diarization system. As previously reported in (Anguera & Bonastre, 2011), the DER for this dataset was 24.96% using the conventional BIC based diarization system with MFCC features extracted from

the beamformed signal. The baseline performance using the binary keys system results in a DER value of 30.90% for this dataset with MFCC features extracted from the beamformed signal. The methodology presented here removed the performance gap between the BIC based and the binary key based systems. Precisely, the DER of 24.87% for the NIST-RT05S set indicates that the proposed methodologies makes the binary key based diarization a very competitive approach. Although using a concatenation of features and the fusion with TTDOA features increases the computation time, the system is still very fast with average run time for the RT-05S dataset of 0.0516 $\times$RT.

System	**Features**	**DER (%)**	**SER (%)**	$\times$**RT**
Binary Keys (Beamformed)	MFCC	30.90	21.3	0.026
Binary Keys TTDOA Integration (Beamformed)	MFCC + TTDOA	26.30	16.7	0.034
Binary Keys (Best Channels)	MFCC	26.44	16.9	0.039
Binary Keys TTDOA Intergation (Best Channels)	MFCC + TTDOA	**24.87**	**15.3**	0.051
Information Bottleneck (Vijayasenan et al., 2008)	MFCC + TDOA	–	17.7	0.340
BIC based System (Anguera & Bonastre, 2011)	MFCC	24.96	–	1.19
Online i-vector with Information Bottleneck (Madikeri et al., 2015)	MFCC	–	16.1	–
PLDA i-vector with Information Bottleneck (Madikeri et al., 2015)	MFCC	–	16.5	–
Robust GMM based Modelling (Peso, 2016)	TDOA	–	17	–

Table 5.14 Summary of diarization systems performance in terms of SER (%) for the RT-05S NIST set.

From a DER and SER perspective, other than the BIC based diarization system, the performance presented here is superior to other diarization methods, integrated diarization systems and acoustic/spatial features fusion. Recently, a method for robust TDOA features modelling was presented in (Peso, 2016). The SER was found to be 17.0% using TDOA features with the number of speakers known. In (Madikeri et al., 2015), an online i-vector extractor system is integrated with information bottleneck based diarization to produce SER of 16.1%. The same study reported an SER of 16.5% based on a PLDA i-vector system instead of the online i-vector. Also, previously in (Vijayasenan et al., 2008), TDOA features were fused with MFCC features using an information bottleneck system resulting in SER of 17.7%. Table 5.14 provides a summary of those results and the results achieved here.

5.4.5 WPCA Based Fusion of Acoustic and Spatial Features

This section presents a short experiment that demonstrates how weighted PCA of Chapter 4 can be used to fuse MFCC and TTDOA features. Each feature type, MFCC and TTDOA, is assigned a different weight and the weights are associated with a conversation's feature matrix (after mean and variance normalisation) using equation (4.4) which determines the weighted 'correlation' matrix. Before performing the WPCA analysis, feature vectors that represent silences are removed. This experiment uses the IS1000 set of the AMI corpus (Carletta et al., 2006). The first four meetings (IS1001a, IS1002d, IS1003a and IS1004a) will be reserved for final evaluation. Meetings IS1005a, IS1006a, IS1007a and IS1009a will be referred to as the calibration subset and will be used to identify the suitable weighting and number of components. Since the acquisition conditions are similar for all the recordings used in this experiment, it can be expected that the parameters optimised on the calibration subset will produce good results on the final evaluation subset.

The meetings of the IS1000 set were recorded using 12 microphones each. The signals of those microphones are combined using beamforming and 19 MFCC coefficients are then extracted. Delays are calculated between each microphone's signal and the central microphone, then they were transformed using Box-Cox transformation to provide 11 dimensional TTDOA features. The combination of MFCC and TTDOA features results in 30 dimensional feature vectors. Feature vectors are mean and variance normalised, the weighted principal components are retained using the RNN framework then the mean and variance normalised feature vectors are projected onto the principal components using (5.20).

At the beginning, an arbitrary number of principal components are selected, 15 in this case, then the weights are varied to find the suitable ones for each feature (MFCC and TTDOA). Using the calibration subset, Fig. 5.20 illustrates the DER and SER in light of weights variation. w_a indicates the weight of acoustic (MFCC) features and the spatial (TTDOA) features weight w_s is equal to $1 - w_a$.

From Fig. 5.20, the lowest error is achieved when TTDOA features are assigned higher weight than MFCC features. Given the case of score fusion when the best performance was achieved at equal weighting of MFCC and TTDOA, the case with WPCA is probably due to the fact the MFCC dimensionality is higher than TTDOA. Hence, in the analysis of WPCA, higher weighting was required for TTDOA features in order to emphasise their effect on the analysis.

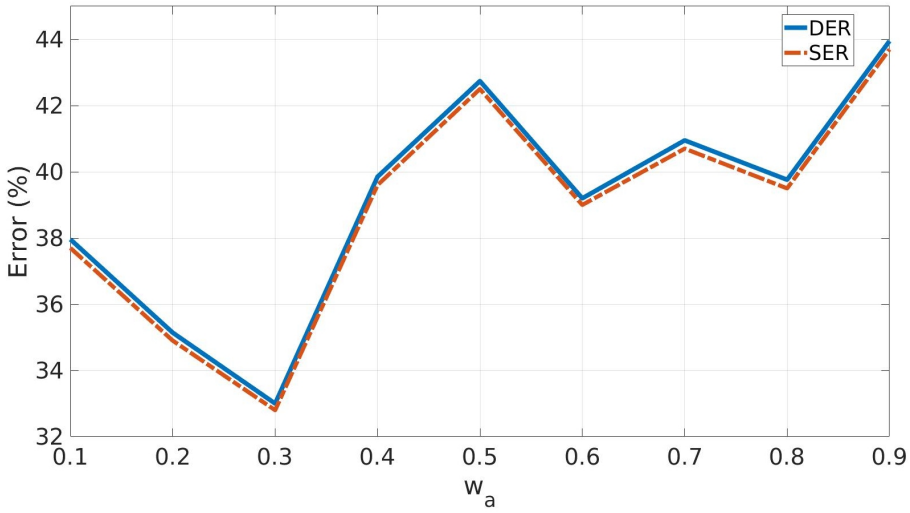


Fig. 5.20 Feature weighting in WPCA. Number of components is 15. It can be seen that the lowest error was given when MFCC features are assigned the weight 0.3 and TTDOA are assigned the weight 0.7.

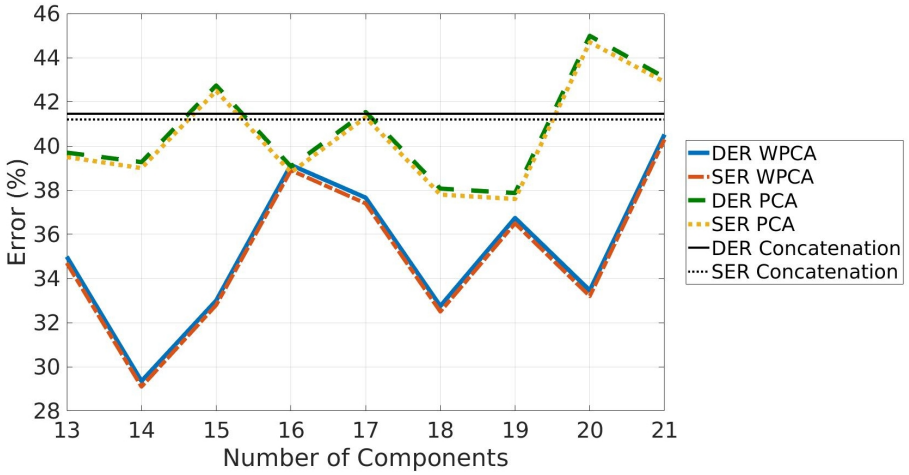


Fig. 5.21 System performance for the fusion of MFCC and TTDOA features using concatenation, WPCA and PCA with variable number of principal components. The calibration set was used here (IS1005a, IS1006a, IS1007a and IS1009a). Features' weights are: $w_a = 0.3$ and $w_s = 0.7$.

At the weights $w_a = 0.3$ and $w_s = 0.7$, Fig. 5.21 illustrates the performance across a range of principal components for the scoring of the original features. It also compares RNN based WPCA to traditional PCA based on the Singular Value Decomposition (SVD) technique. It can observed that WPCA outperforms traditional PCA in the majority of the cases. The three best number of components indicated by the exhibited error are chosen to experiment the system performance with the final evaluation (meetings: IS1001a, IS1002d, IS1003a and IS1004a). The results are shown in Table 5.15 with comparison to traditional PCA and to the case of feature concatenation.

Method - Feature Dimension	**DER (%)**	**SER (%)**	**False Alarm (%)**	**Missed Speech (%)**
Concatenation - 30	31.73	31.00	43.80	12.50
WPCA - 20	28.87	28.10	12.5	6.2
PCA - 20	36.37	35.60	6.2	6.2
WPCA - 18	32.03	31.30	0	18.8
PCA - 18	37.69	36.90	18.8	12.5
WPCA - 14	30.99	30.20	0	12.5
PCA - 14	39.09	38.30	6.2	6.2
Score Fusion - 30	25.87	25.1	6.2	25.0

Table 5.15 System performance using the evaluation subset of the IS1000 data with feature fusion by concatenation, WPCA and PCA. The number of components chosen are the best ones indicated by the system error as shown in Fig. 5.21. The bottom row shows the case of score fusion for this subset at $w_a = w_s = 0.5$.

It can be seen in Table 5.15 that WPCA outperforms traditional PCA in all of the cases. WPCA also outperforms concatenation of features in terms of false alarm and missed speech. Additionally, traditional PCA noticeably improved false alarm and missed speech errors but it has worsened the DER and SER. The case of score fusion provided better DER and SER but relatively high missed speech error in comparison to PCA based fusion. Given both cases, of score fusion and PCA based fusion, there is trade-off to be made between accuracy and computational complexity which can be decided in favour of the underlying application. For example, the performance of the clustering phase, indicated by false alarm and missed speech errors, appears to be better in case of PCA based fusion which can be preferable for speaker detection and counting.

This experiment demonstrated another aspect of weighted PCA where different features can be weighted as appropriate. While PCA is commonly seen as a sole dimensionality reduction technique, WPCA can additionally be used as an effective feature fusion technique that can be very useful for speaker recognition in general.

5.5 Summary

The work in this chapter addressed two aspects related to speaker diarization and it proposed methodologies that noticeably improved the performance of binary key based diarization at a computation speed that reached a maximum of 0.056 $\times$RT. Acoustic features, usually MFCC, are the main input to diarization systems. Therefore, one of the aspects focused on making use of richer sources (when multiple channels exist) for the extraction of MFCC features as opposed to their extraction from a single (beamformed) signal.

While a concatenation of MFCC features of all available channels' was shown to improve the performance, the proposed channel selection methods provide a suitable solution for the problem of increased dimensionality that would otherwise slow down the system. A selection of distant microphones can possibly help to capture similar diverse information captured by all the channels in an enclosed space; especially, if the major diversity comes from the acoustical conditions of the space. Selection of the best quality channels, on the other hand, can be particularly helpful when low quality, faulty or badly located microphones exist.

Interestingly, it has been observed during the experimentation that the best quality channel selection method tends to, generally, choose channels that are selected as close channels by the distant channel selection method. This implies that microphones that are closer to the speaker(s) provide better quality speech signals. This could be due to the fact that the magnitude of the direct signal is higher than those of multipath signals. Furthermore, the power of peripheral environmental noise can be lower than the power of direct speech signals.

The second aspect focused on the integration of spatial features as an additional input that was found to be helpful in other systems. This chapter highlighted the issue of the distribution of TDOA features which, to the best of the author's knowledge, was not addressed before. A non-normal distribution of these features is found to be an obstacle for their integration in binary key based diarization. As learned from the system behaviour when MFCC features are used, it is found to be mandatory for the features to be normally distributed such that the anchor models are situated in the centre of the feature space. That is, in turn, believed here to be essential for the derivation of discriminative binary keys.

The next chapter presents methods for non-uniform initialisation of the binary key based diarization. It also presents an acquisition of MFCC features that is based on the selection of least distorted channels' subbands.

Chapter 6

Subband Based Diarization and System Initialisation

The focus of the first section of this chapter is similar to the the idea of channel selection presented in the previous chapter. From the literature, it came to the attention that reverberation effect can vary across the speech spectrum. Therefore, the first section here presents the idea of selecting the least distorted subbands of the available channels instead of the entire channel. The reverberation effect is to be assessed using an appropriate measure presented in this chapter that is expected to be adaptive to the acoustic conditions of a meeting session. The performance of binary key based diarization is to be evaluated using acoustic features extracted from selected subbands of the available channels. This will be compared to the performance when features are extracted from channels combination based on beamforming.

The second part of this chapter focuses on the initialisation of binary key based diarization. A number of methods will be presented for this purpose. One of the effective aspects of these methods is that they make use of the cumulative vectors and binary keys that are already derived as a part of the actual diarization process. System performance based on these methods will be compared to the uniform initialisation method. Also, the most robust method among the ones proposed will be identified. The additional processing time presented by using non-uniform initialisation will be reported.

6.1 Speaker Diarization Based on Spectrum Subbands

This section introduces a methodology for channels' subbands selection in relation to the amount of the reverberation exhibited. It also presents an evaluation of the performance of binary key based diarization using OE-MFCC features introduced in Chapter 3.

6.1.1 Selection of Least Reverberated Channels' Subbands

The amount of reverberation can be influenced by the location of the recording microphone in an enclosure, such as a meeting room. This is because each microphone receives a different amount of reflected signals and is exposed to a different proportion of the direct signal in comparison to reflected signals. This was partly addressed with the quality based channel selection methodology in Chapter 5. However, the amount of reverberation and the reverberation time in particular can vary depending on frequencies (Ismail, 2013). Hence, for each channel (microphone) there is potentially a range of its frequencies that might be less affected by reverberation.

This motivates the idea of extracting acoustic features from subbands of channels that are less affected by reverberation. It is important to clarify that a channel refers to the speech signal recorded by a microphone and a subband refers to a range of frequencies in the spectrum of that signal. In brief, the speech spectrum will be first divided into a number of subbands and the reverberation effect on those subbands will then be characterised over all of the available channels. The least reverberated subbands among the underlying channels are then chosen. In the end, the entire speech spectrum is retained from different channels and MFCC coefficients are extracted from the log-energies of the mel-filters that correspond to each subband.

6.1.1.1 Average Joined Gradient Estimates of Reverberation

Reverberation is known to cause smearing in the speech spectrum. This effect can be visually observed when stacking together the spectral estimates of short speech frames like it is done in the extraction of MFCC. Consider the layout in Fig. 6.1, reflected speech is delayed and it overlaps with the directly propagated speech at the acquisition point. In the speech spectrum represented by a sequence of frames, reflected speech signal causes extensions in the speech energy from one frame to another as illustrated by the reverberated spectrum in Fig. 6.2a. On the other hand, the spectrum of clean (non reverberated) speech signal has sharper transitions from one frame to the other as shown in Fig. 6.2b.

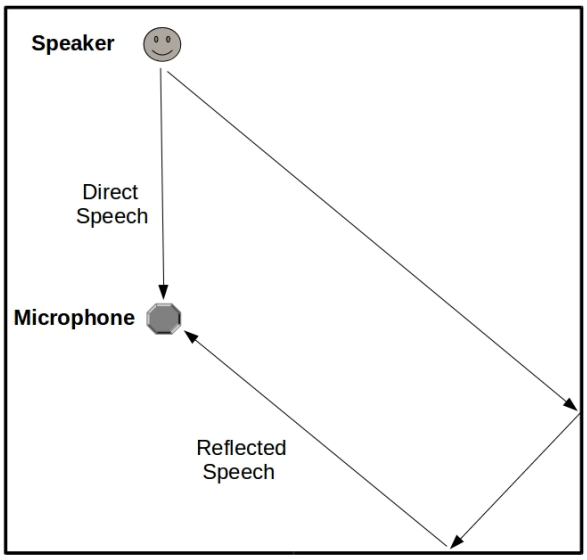
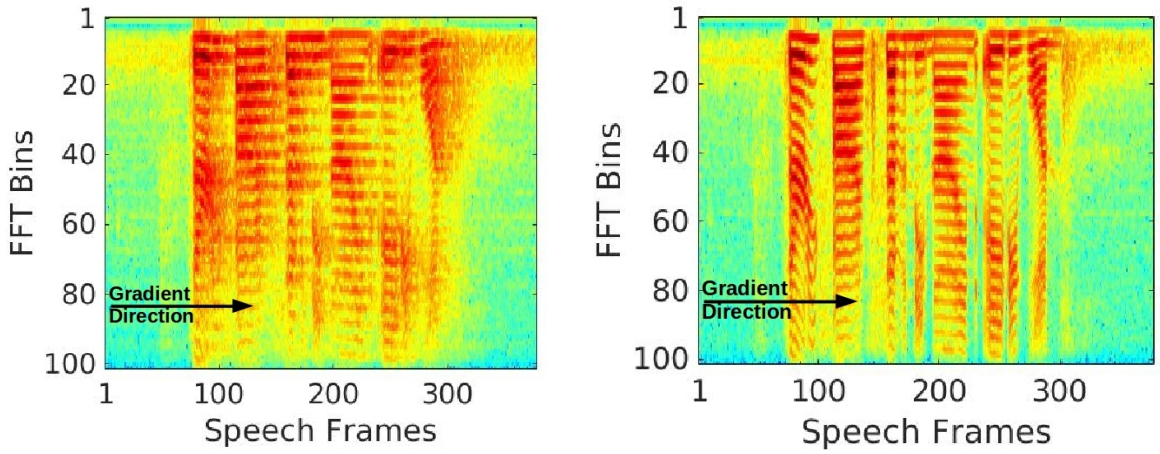


Fig. 6.1 Hypothetical room setup illustrating how reverberated speech can develop.

It is proposed here, to estimate the gradients of the spectrum across the speech frames and to use it to characterise the degree of reverberation. It is assumed that the smearing effect of reverberations minimises the gradient. Hence, less reverberated speech will have higher gradient values.



(a) Reverberated Speech (b) Clean Speech

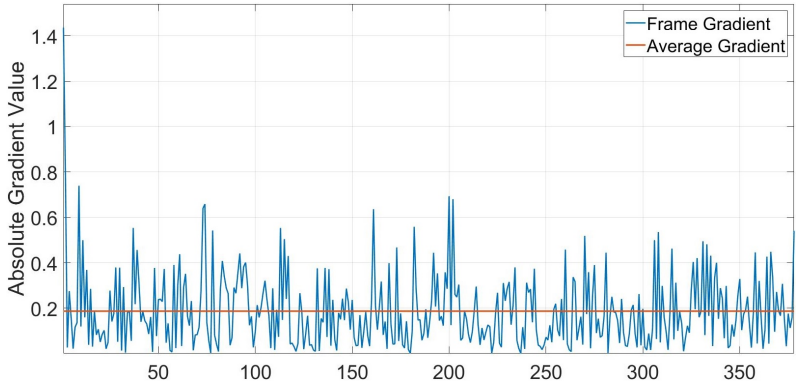
Fig. 6.2 Speech sample of the YOHO data (Campbell & Higgins, 1994) for a female uttering the numbers "35 79 81". Artificial reverberation of 0.7s was added to produce the reverberated sample.

In order to estimate the gradient, the speech signal is first divided into frames of 25 ms length. Overlap is not allowed between frames since it would result in some continuity of the speech spectrum from one frame to another which would affect the gradient. The spectrum is determined for each frame as $\log_{10}$ of the magnitude of the discrete Fourier transform (i.e. FFT). In an attempt to equalise the gradient estimates over different channels, the mean and variance of the spectrum is normalised across the speech frames.

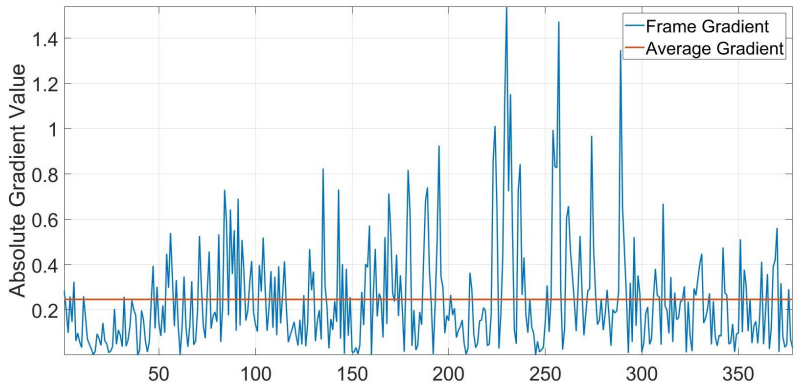
Let k denote a fraction of the spectrum (one FFT bin). Let $\eta(k,t,j)$ be k's value at speech frame t of channel j. The mean of the absolute gradient of $\eta(k,t,j)$ for T frames is determined as

$$\xi_{k,T,j} = \frac{1}{T}\sum_{t=1}^{T} |\Delta_t \eta(k,t,j)|, \tag{6.1}$$

where Δ_t is the gradient function over time.



(a) Reverberated Speech



(b) Clean Speech

Fig. 6.3 These plots illustrate the absolute value of the gradient across t (the x-axis))as calculated in (6.1). This is the gradient across the 80^{th} bin of the spectrums shown in Fig. 6.3. The average of the gradient is also shown.

The mean of the absolute gradient, calculated above, is plotted in Fig. 6.3 for the sample of Fig. 6.2. It can be noticed form Fig. 6.3 that clean speech has higher average gradient than reverberated speech. This is because clean speech spectrums in general should have sharp transition between the frames (across t). Thus, the differences are higher between the spectrum magnitudes of the frames.

The average gradient of a specific subband of the spectrum is calculated as

$$\bar{\xi}_{k_1,k_2,j} = \frac{1}{k_2 - k_1} \sum_{k=k_1}^{k_2} \xi_{k,T,j}, \tag{6.2}$$

where k_1 and k_2 are, respectively, the low and high frequency limits of the subband.

Channels that exhibit high reverberation times can have similar $\bar{\xi}_{k_1,k_2,j}$ values which could make the degree of reverberation to be indistinguishable when measured by $\bar{\xi}_{k_1,k_2,j}$. For example, it is possible that longer spread of the speech energy causes the value of $\bar{\xi}_{k_1,k_2,j}$ to increase which is the opposite to what was originally assumed here. The possibility of such conditions can be tackled by introducing a threshold to discard overly extended smearing of the spectrum.

The threshold is determined using all of the channels for which the reverberation to be characterised over the subband k_1 to k_2. It is calculated as

$$\iota_{k_1,k_2} = \frac{1}{M} \sum_{j=1}^{M} \bar{\xi}_{k_1,k_2,j}, \tag{6.3}$$

where M is the number of channels.

This threshold is basically the average of all channel's $\bar{\xi}_{k_1,k_2,j}$ values obtained in (6.2). The value of ι will be used to transform the gradient estimates into binary values. For $k_1 \leq k \leq k_2$, the new gradient estimates, or the joined gradient, will be obtained by the following transformation

$$\hat{\xi}_{k,t,j} = \begin{cases} 1 & \text{for} \quad \xi_{k,t,j} \geq \iota_{k_1,k_2} \\ 0 & \text{for} \quad \xi_{k,t,j} \leq \iota_{k_1,k_2} \end{cases} . \tag{6.4}$$

where $\xi_{k,t,j}$ is the j^{th} channel gradient value for specific fraction of the spectrum k (equivalent to an FFT bin) at frame t. Then, the new Average Joined Gradient (AJG) estimate of channel

j over T frames and for a subband that extends from k_1 to k_2 is determined as

$$\hat{\xi}_{k_1,k_2,j} = \frac{1}{k_2 - k_1} \sum_{k=k_1}^{k_2} \left(\frac{1}{T} \sum_{t=1}^{T} \hat{\xi}_{k,t,j} \right). \tag{6.5}$$

The higher the reverberation effect the higher the smearing it causes in the spectrum which minimises the value of $\hat{\xi}_{k_1,k_2,j}$ as assumed here. The channel that exhibits the lowest reverberation at subband k_1 to k_2 is selected using the AJG estimate of (6.5) as

$$j_{selected} = \arg\max_{\forall j} \hat{\xi}_{k_1,k_2,j}. \tag{6.6}$$

As stated earlier, this selection method is designed to account for hypothetical variations in the degree of reverberation across the speech spectrum of the available channels. Three subband selection cases will be investigated. In one case, the spectrum is divided into two equal subbands each is to be selected from a different channel. The second case considers three equal subbands and the third case considers four equal subbands.

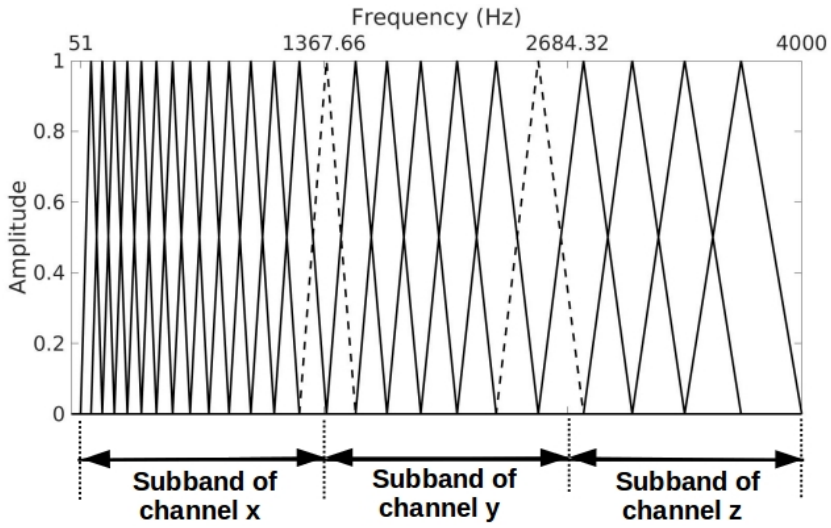


Fig. 6.4 The framework of feature extraction from selected subbands based on MFCC methodology. This figure illustrates the case of the spectrum being divided into three equal subbands each is selected from a different channel. Cepstral coefficients are to be calculated separately from the filters that cover each subband. Dotted filters indicate that they will be included in the feature extraction of both of the adjacent subbands.

Acoustic feature extraction from selected subbands will be based on the MFCC framework. Fig. 6.4 illustrates the case where the spectrum is divided into three equal subbands.

The filters that cover the subbands' edges will be included in the feature extraction for both of the adjacent subbands.

6.1.1.2 Detection of Simulated Reverberation Effects

This subsection demonstrates the accuracy of the Average Joined Gradient (AJG) estimates in detecting the degree of reverberation. Simulated reverberation effects are added to a clean speech sample using the tool designed for the REVERB challenge by Kinoshita et al. (2013). The tool performs convolution between the speech sample and a designated Room's Impulse Response (RIR). Three different reverberation times are added and tested:

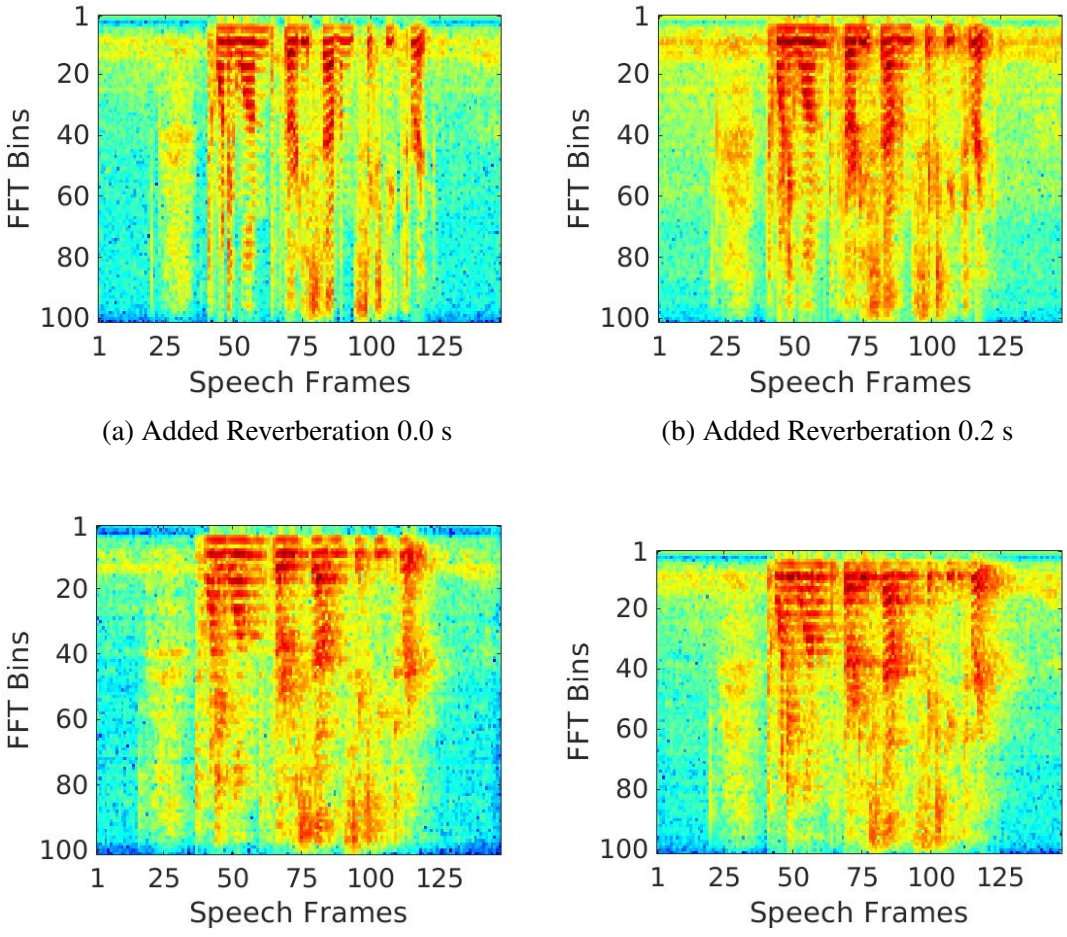
- 0.2s using the RIR of simulation room 1 recording angle A;
- 0.5s using the RIR of simulation room 2 recording angle A;
- 0.7s using the RIR of simulation room 3 recording angle A.

Added Reverberation	$\bar{\xi}_{k_1,k_2,j}$	$\hat{\xi}_{k_1,k_2,j}$
0.0s (Original Speech)	0.356	0.603
0.2s	0.166	0.308
0.5s	0.159	0.276
0.7s	0.152	0.260
Standard Deviation	0.098	0.162

Table 6.1 Values of the average gradient ($\bar{\xi}_{k_1,k_2,j}$) and average joined gradient ($\hat{\xi}_{k_1,k_2,j}$) in relation to different added reverberation times as well as the original speech sample "21 37 63" of the YOHO data (Campbell & Higgins, 1994).

A speech sample from the YOHO data (Campbell & Higgins, 1994) for a male uttering the numbers "21 37 63" is used here to test the gradient estimations of its convolution with the three rooms' RIR as shown in Fig. 6.5. For the entire speech spectrum, Table 6.1 shows the values of the average gradient ($\bar{\xi}_{k_1,k_2,j}$) determined using (6.2) and the average joined gradient ($\hat{\xi}_{k_1,k_2,j}$) estimated using (6.5). The value of these gradients, $\bar{\xi}_{k_1,k_2,j}$ and $\hat{\xi}_{k_1,k_2,j}$, is expected to decrease as reverberation time increases. This is because an increase in reverberation causes spectrum smearing to increase. Thus, the transitions of the spectrum between speech frames would become smoother.

One can notice from Table 6.1 that both $\xi_{k_1,k_2,j}$ and $\hat{\xi}_{k_1,k_2,j}$ decrease as the reverberation time increases which accommodates the assumptions made here. The maximum values are given for non-reverberated speech. More importantly, the values of $\hat{\xi}_{k_1,k_2,j}$ have higher standard deviations which makes this measure more precise in distinguishing close reverberation times.



(a) Added Reverberation 0.0 s

(b) Added Reverberation 0.2 s

(c) Added Reverberation 0.5 s

(d) Added Reverberation 0.7 s

Fig. 6.5 The speech spectrum of the speech frames for a speech sample of a male saying the numbers "21 37 63" from the YOHO data (Campbell & Higgins, 1994). The figure shows the spectrum of the original sample (0.0 s) as well as the spectrums with added reverberation. Recall that the frames are 25 ms in size and they are not overlapped.

6.1.1.3 Evaluation on Speaker Diarization

The performance of binary key based diarization is to be evaluated here using acoustic features extracted from selected channel's subbands (see Fig. 6.4, as an example). The three

cases of subbands division to be investigated are summarised in Table 6.2. Despite that the proposed method aims to account for possible differences in reverberation effects over the subbands, the extent of reverberation over the spectrum is unknown. In an attempt to tackle such uncertainty when examining a subband, the AJG ($\hat{\xi}$) will be estimated over the subband plus 50% extensions with the adjacent subbands as described in the second column of Table 6.2.

Number of Subbands	Spectrum Limits for Estimating $\hat{\xi}_{k_1,k_2,j}$	Subbands for Feature Extraction	Mel-Filters Subset	Feature Dimension
2	51 - 3012.5 Hz, 1037.5 - 4000 Hz.	51 - 2025 Hz, 2025 - 4000 Hz.	1 - 17, 17 - 24.	23
3	51 - 2025 Hz, 709.33 - 3342.65 Hz, 2025 - 4000 Hz.	51 - 1367.66 Hz, 1367.66 - 2684.32 Hz, 2684.32 - 4000 Hz.	1 - 14, 14 - 20, 20 - 24.	23
4	51 - 1531.25 Hz, 543.75 - 2518.75 Hz, 1531.25 - 3506.25 Hz, 2518.75 - 4000 Hz.	51 - 1037.5 Hz, 1037.5 - 2025 Hz, 2025 - 3012.5 Hz, 3012.5 - 4000 Hz.	1 - 11, 11 - 17, 17 - 21, 21 - 24.	23

Table 6.2 Summary of the MFCC based feature extraction framework from selected channels' subbands. The third column shows the subbands to be selected from different channels. The exact subband of a channel used in the feature extraction can be slightly extended as a result of the actual number of filters used, refer to Fig. 6.4.

The choice of the best subband division case will be made based on experiments performed using 16 meeting excerpts of the AMI corpus (Carletta et al., 2006). The first eight meetings are the IS1000 set summarised in Table 5.1 and the second set of eight meetings is the TS3000 set summarised in Table 5.2. The final evaluation will be made on the RT-05S set described in Table 5.3.

The binary key based system here also uses the same setup previously described in Section 5.4.2. In Chapter 5, the baseline performance was produced using MFCC features with 19 dimensions extracted using 24 triangular mel-filters (Table 5.4). For consistent comparison with the feature extraction setup of this section, MFCC features are extracted using the same filter bank of 24 mel-filters but with 23 dimensions (the 0^{th} order coefficient

is excluded). The difference in feature dimensionality provided new results as shown in Table 6.3.

Dataset	DER (%)	SER (%)	False Alarm (%)	Missed Speech (%)
IS1000	34.66	34.2	12.5	12.5
TS3000	44.68	43.7	37.5	3.1
IS1000 & TS3000	39.99	39.2	25.0	7.8
RT-05S	32.13	22.6	26.8	25.0

Table 6.3 Binary key based system performance for the datasets under investigation using 23 dimensional MFCC features extracted from beamformed signals. These results are the reference to which the subband selection results are compared.

Table 6.4 shows the system performance using the combination of IS1000 & TS3000 sets for the 2, 3 and 4 division cases of the subband selection and feature extraction process describe here (as summarised in Table 6.2). Both of the 2 and 3 subband cases improve the accuracy over the reference performance shown in Table 6.3. In theory, having smaller sections of the spectrum selected from different channels is not expected to degrade the performance. However, the case of 4 subbands has slightly degraded the performance which is believed to be caused by the feature extraction framework. In the case of 4 subbands, the third subband (2025 - 3012.5 Hz) is decomposed using 5 mel-filters and the fourth subband (3012.5 - 4000 Hz) is decomposed using 4 mel-filters. Since those are overlapped filters, they can have poor transformation of the spectrum because they are expected to have relatively high residual in the correlation matrix of their log-energies as discussed in Section 3.2.1.2 (also see Fig. 3.5).

No. of Subbands	DER (%)	SER (%)	False Alarm (%)	Missed Speech (%)
2	38.87	38.1	26.6	9.4
3	35.02	34.3	14.1	9.4
4	40.61	39.9	23.4	7.8

Table 6.4 Binary key based system performance for the combination of IS1000 and TS3000 sets for the cases of 2, 3 and 4 subbands.

From Table 6.4 one can notice that the best results are obtained for the case of three subbands. This finding is further investigated and evaluated on the RT-05S set and the results are shown in Table 6.5. The same table also presents separate results for each of the IS1000 and the TS3000 sets. The results of Table 6.5 appear to show a noticeable improvement over the case of using MFCC features extracted from the beamformed signal

(Table 6.3). The results appear to provide evidence that the methodology presented in this section might considered to be a better practice than the process of combining all channels' signals into a single beamformed signal. Beamforming is a time domain process that does not take into account the spectral properties for individual microphones and in particular any deficiencies in particular ranges of a microphone's spectrum. These results appear to show that channels' subbands selection discards channels with spectrums that may have been distorted by reverberation effects or other degradation.

Dataset	**DER (%)**	**SER (%)**	**False Alarm (%)**	**Missed Speech (%)**
IS1000	30.00	29.5	6.2	15.6
TS3000	39.45	38.5	21.9	3.1
RT-05S	28.21	18.6	16.1	21.4

Table 6.5 The performance of the binary key based diarization system for each of the IS1000, TS3000 and RT-05S datasets in the case of three equal subbands spectrum division.

The diarization accuracy shown in Table 6.5 is slightly lower than the case of using a concatenation of features of selected channels which was investigated in Chapter 5. However, the method presented in this section provides a reduction in the feature dimensionality. Therefore, there is a trade-off to be made when deploying the speaker diarization system.

6.1.2 Evaluation of Diarization Performance using OE-MFCC

In Chapter 3, OE-MFCC features were presented and they were shown to improve the performance of speaker verification over regular MFCC features based speaker verification. This is of interest here too as regular MFCC features based speaker diarization have been used in the literature (see e.g Delgado et al. (2015a) and Anguera & Bonastre (2011)) to capture the performance of binary key based diarization.

Therefore, it would be interesting to investigate the performance of speaker diarization using OE-MFCC features. The same filter bank of 24 triangular mel-filters is used here. 11 cepstral coefficients are extracted from each of the odd and even filters subsets for a total of 22 cepstral coefficients. This feature extraction configuration is comparable to extracting 23 dimensional MFCC features from the beamformed signal (Table 6.3) using a bank of 24 filters. However, as discussed in Section 3.3.3, the first 2 coefficients obtained from the even filters subset should be omitted because they exhibit high correlation with the same coefficients obtained from the odd subset. Accordingly, the total number of cepstral coefficients becomes 20.

The binary key based system setup is the same one described in Section 5.4.2. Using a Hamming window in the spectral estimations, Table 6.6 shows the performance of speaker diarization using OE-MFCC. In comparison to the results of Table 6.3, one can see that OE-MFCC has improved the results for TS3000 and RT-05S sets. On the other hand, MFCC outperformed OE-MFCC for the case of the IS1000 set, yet OE-MFCC has reduced Missed Speech error to zero.

Dataset	**DER (%)**	**SER (%)**	**False Alarm (%)**	**Missed Speech (%)**
IS1000	39.70	39.2	18.8	0.0
TS3000	43.28	42.3	37.5	3.1
RT-05S	29.13	19.5	14.3	23.2

Table 6.6 Speaker diarization performance using OE-MFCC features with Hamming window based spectral estimations.

It was shown in Section 3.3.3 that the extraction of OE-MFCC features improves when it is based on multitaper spectral estimations. That observation is confirmed by the results shown in Table 6.7 for speaker diarization where the spectral estimates for OE-MFCC are made using four multipeak tapers as in (Kinnunen et al., 2010). In comparison to Table 6.6, one can observe noticeable improvements in the system's performance provided in Table 6.7.

Dataset	**DER (%)**	**SER (%)**	**False Alarm (%)**	**Missed Speech (%)**
IS1000	27.44	26.9	21.9	3.1
TS3000	41.26	40.3	25.0	3.1
RT-05S	25.76	16.2	7.1	19.6

Table 6.7 Speaker diarization performance using OE-MFCC features with four multipeak multitaper based spectral estimations.

Other than speaker verification, the experiments here demonstrated the superiority of OE-MFCC over traditional MFCC for speaker recognition in general. It is anticipated that further improvements can be obtained when a concatenation of OE-MFCC features of selected channels are used in speaker diarization.

6.2 Initialisation of Binary Key Based Diarization

Diarization systems usually adopt uniform initialisation, as discussed in Section 2.3.4, which can provide acceptable performance. However, uniform initialisation may not be appropriate

for all diarization systems. In binary key based diarization, cluster merging takes place after segment re-allocation without the use of any algorithms to identify a best segmentation path as in Martínez-González et al. (2017) with the Viterbi algorithm. Such algorithms are useful for providing increased accuracy but require more computation time. A disadvantage of binary key based diarization is at the beginning of the process when uniform initialisation is deployed. This is because it is very likely that the system merges clusters containing more than one speaker very soon. In such a case, for example, the segments of a speaker with low contribution to the conversation are expected to be merged with another speaker's model.

In binary keys diarization, cluster and segment modelling, whether by cumulative vectors or binary keys, can facilitate the development of efficient initialisation methods like the ones to be introduced in this section. The most important aspect of the source of this improvement in efficiency is the compatibility of such methods with the diarization system from a computational load perspective. A number of methods are proposed here based on cumulative vectors and others are based on variations of binary keys.

Algorithm 1 The Cluster Purification Framework

Input: $\mathbf{X}$, C ▷ $\mathbf{X}$ are the features vectors of a conversation
▷ C is the number of initial clusters

Split $\mathbf{X}$ into 3s segments with 1s overlap between adjacent segments
Derive a model for each segment and store it in SModels
▷ SModels are binary keys or cumulative vectors

Split $\mathbf{X}$ into C uniform portions
Derive a model for each portion and store it in CModels
▷ CModels are binary keys or cumulative vectors
Initialise a vector Labels with size S to zeros ▷ S is the number of segments
PreLabels = Labels
Set $D = 1$
while $D \neq 0$ **do**
    for $i = 1$ to C **do**
        for $j = 1$ to S **do**
            Similarities (i, j) = *metric* (CModels(i), SModels(j))
            ▷ *metric* is the cosine similarity or the Jaccard coefficient
        end for
    end for
    for $i = 1$ to C **do**
        $\forall j \in S$ find the set of segments, S_i, with maximum scores to cluster i in Similarities (i, j)
        $\forall s \in S_i$, Labels(s) = i
        Derive new cluster model (CModel) from the feature vectors of the segments s
        CModels(i) = CModel
    end for
    D = sum(abs(PreLabels - Labels))
    PreLabels = Labels
end while
Output: Labels, CModels

A new idea (**Algorithm 1**) is presented for initialisation which aims to purify the initial clusters. The purification process starts with uniform clusters and iteratively assigns segments to clusters and estimates new cluster models from those segments until segments re-allocation converges. Thus the initial clusters are expected to consist of homogeneous segments before any merging takes place. Cluster purification is used to obtain the initial clusters, see Fig. 6.6a & 6.7a, and as a method to start the K-means algorithm (**Algorithm 2**), see Fig. 6.6b & 6.7b. Given a cluster's centre, K-means aims to maximise the similarity between that centre and the models of the segments assigned to the cluster. The performance of K-means is heavily influenced by the way it is initialised, hence, different initialisation criteria are investigated.

The segments used within the initialisation methodologies are the same ones that the system uses in the actual diarization process. Those are uniform with a coverage of one second and an overlap of one second with the proceeding and the following segments. Hence, the total segment length is three seconds.

Algorithm 2 The K-means Based Framework

Input: **X**, C ▷ **X** are the features vectors of a conversation
▷ C is the number of initial clusters

Split **X** into 3s segments with 1s overlap between adjacent segments
Derive a model for each segment and store it in SModels
▷ SModels are binary keys or cumulative vectors

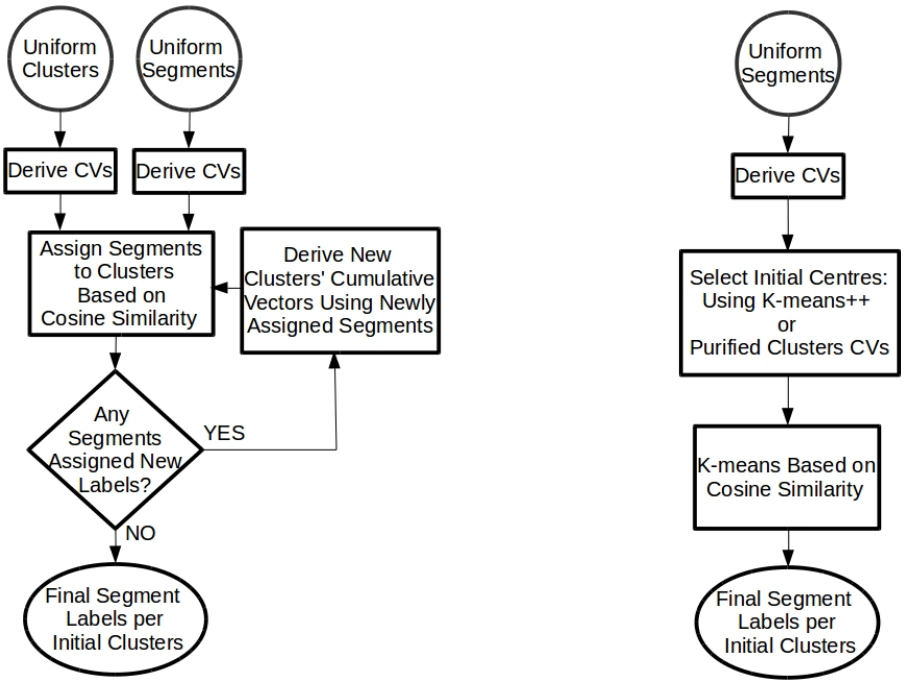
Select CCentres using one of the methods below
▷ CCentres are the initial cluster centres
CCentres = *random* (SModels, C) ▷ select C random segments models
CCentres = *K-means*++ (SModels, C) ▷ select using the K-means++ algorithm
CCentres = CModels ▷ use the cluster models obtained by cluster purification

Initialise a vector Labels with size S to zeros ▷ S is the number of segments
for $k = 1$ to N **do** ▷ N is the number of K-means iterations
  for $i = 1$ to C **do**
    for $j = 1$ to S **do**
      Similarities (i, j) = *metric* (CCentres(i), SModels(j))
      ▷ *metric* is the cosine similarity or the Jaccard coefficient
    end for
  end for
  for $i = 1$ to C **do**
    $\forall j \in S$ find the set of segments, S_i, with maximum scores to cluster i in Similarities (i, j)
    $\forall s \in S_i$, Labels(s) = i
    Calculate new cluster centre: CCentre = mean(SModels(s))
    CCentres(i) = CCentre
  end for
end for
Output: Labels

6.2.1 Cumulative Vector Based Initialisation

The cosine similarity is used as a similarity measure within the methodologies presented for this case. In cluster purification, feature vectors of a conversation are first uniformly divided into a set of preliminary clusters. Cumulative vectors are derived for the uniform segments and clusters. The cosine similarity is then used to assign segments to clusters and new cluster cumulative vectors are obtained from the corresponding speech frames of the newly assigned segments. The process is repeated until no segment is assigned a new label, see Fig. 6.6a.

Using the K-means algorithm, Fig. 6.6b, only the cumulative vectors of the segments are required. K-means is iterated for a maximum of 100 times. The K-means++ algorithm (Arthur & Vassilvitskii, 2007) is adopted here to find the initial centres for the K-means method instead of using arbitrary centres. K-means++ is an improved algorithm in comparison to K-means, mainly, in terms of speed. The cluster models obtained by the pre-described cluster purification (Fig. 6.6a) are also adopted as initial clusters' centres for the K-means algorithm.

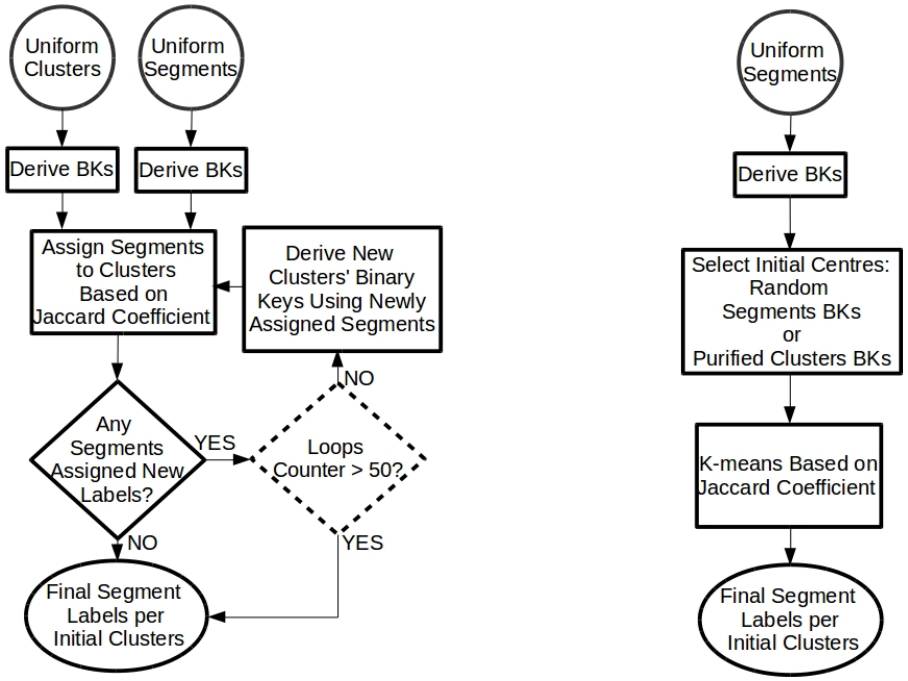


(a) Cluster Purification Framework (b) K-means based Framework

Fig. 6.6 Cumulative vectors and cosine similarity based initialisation.

6.2.2 Binary Key Based Initialisation

The methodology presented here uses the Jaccard coefficient as a similarity measure. The main difference in cluster purification here is the necessity of using a counter, see Fig. 6.7a. This is because, from empirical observation, cluster purification requires many iterations for the segment re-allocation process to converge.



(a) Cluster Purification Framework (b) K-means based Framework

Fig. 6.7 Binary keys and Jaccard coefficient based initialisation.

K-means based on binary keys is different since a straightforward mean calculation from the relevant segments' binary keys provides a vector of non-binary values. The mean is determined here as follows; given a group of segments clustered based on the Jaccard coefficient, the mean of their cluster is determined by averaging their corresponding cumulative vectors. Afterwards, the binary domain cluster centre is derived from that cluster's mean cumulative vector. As illustrated in Fig 6.7b, K-means is initialised with cluster purification or by having the initial centres as the binary keys of randomly selected segments.

6.2.3 Evaluation

This subsection provides an evaluation and examination of the initialisation methodologies presented. The datasets and acoustic features sourcing (from beamformed signals and selected

channels) of sections 5.4.2 and 5.4.3 are used here with the same system parameters given in Section 5.4.2. The main focus is put on identifying a robust initialisation method among the proposed ones which can therefore be reliably deployed as an alternative to uniform initialisation. The evaluation parameters used is DER and the Clustering Error which can be defined as

$$\text{Clustering Error } (\%) = \text{E}_{\text{FA}}(\%) + \text{E}_{\text{MISS}}(\%), \tag{6.7}$$

where the calculations of $\text{E}_{\text{FA}}(\%)$ and $\text{E}_{\text{MISS}}(\%)$ were given in equations (2.32) and (2.33), respectively.

A simultaneous low DER and Clustering Error might imply a good initialisation is obtained. This might also considered to imply that purer clusters were obtained for the final re-segmentation process. The clustering error indicates the system's ability to identify the correct number of speakers present in a conversation which can be greatly affected by the initialisation. In the results shown shortly, cluster purification with Cumulative Vectors is referred to as Purification (CV) and that with Binary Keys as Purification (BK). Similarly, K-means initialised with cluster purification are referred to as Kmeans-P (CV) and Kmeans-P (BK). Also, Kmeans-K (CV) is cumulative vector based K-means initialised with K-means++; and Kmeans-R (BK) is binary key based K-means initialised randomly[1].

6.2.3.1 The case when MFCC is Acquired from Beamformed Signals

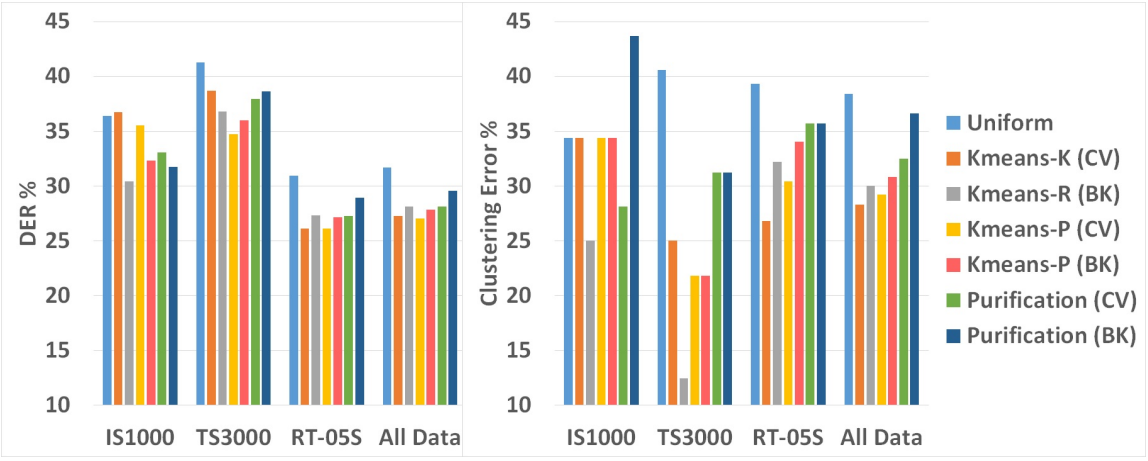
Figures 6.8a and 6.8b illustrate the system performance in DER and Clustering Error in light of the methodologies presented and in comparison to the uniform initialisation case.

In those figures, the system uses MFCC features extracted from the beamformed signals of each conversation. One can notice that on average, i.e. the case of all data, all the methodologies presented outperform the case of uniform initialisation.

For the individual sets, Purification (BK) results in an increase in the Clustering Error for the IS1000 dataset. Also, Kmeans-K (CV) slightly increase the DER for the same dataset. On the other hand, it did not provide an improvement in the Clustering Error for the same case. The rest of the methodologies provided a reduction in DER in all cases and similar (to uniform) or lower Clustering Error at the same time. On average, Kmeans-P (CV) results in the best DER in addition to significant reduction in the clustering error.

Diarization initialisation based on the K-means algorithm is clearly useful, however, its performance is influenced by its own initialisation. This can be investigated by further testing as in the following section.

[1] In case of grey scale print, the bars from left to right are inline with the legends from top to bottom.

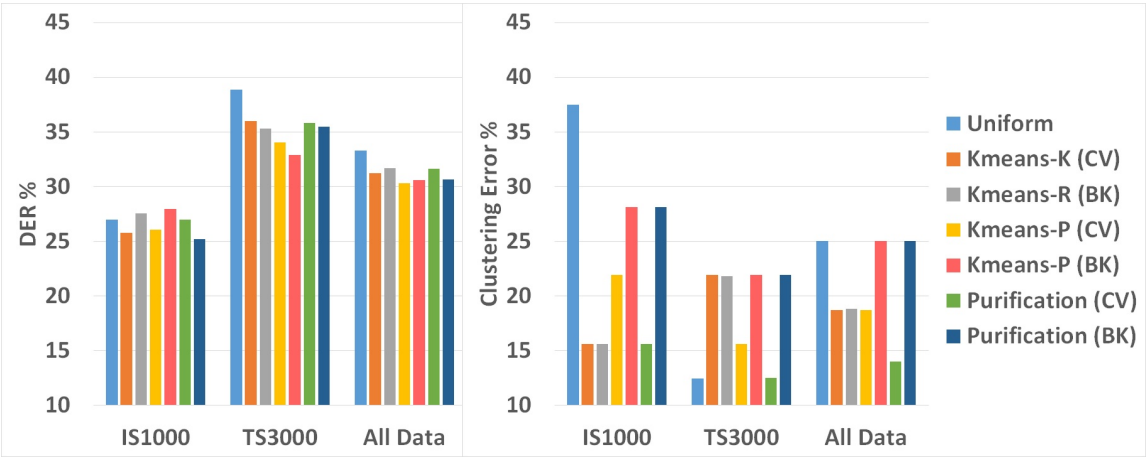


(a) DER (b) Clustering Error

Fig. 6.8 System performance in terms of DER and Clustering Error using 16 initial clusters with MFCC extracted from beamformed signals.

6.2.3.2 The case when MFCC is Acquired from Distant and Best Quality Channels

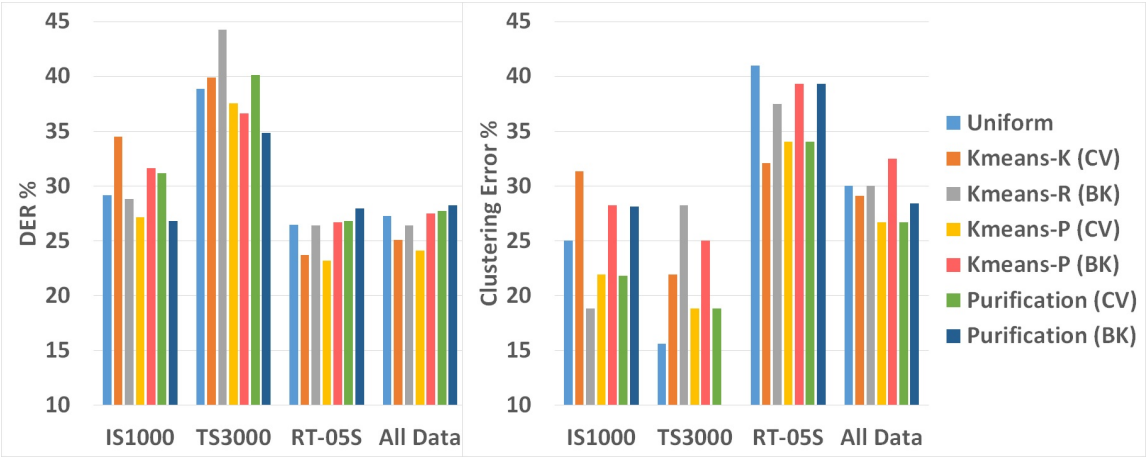
Figures 6.9a and 6.9b show the DER and Clustering Error, respectively, for the case when MFCC features are extracted from distant channels[1]. Additionally, figures 6.10a and 6.10b show the DER and Clustering Error for MFCC features extracted from best channels[2].



(a) DER (b) Clustering Error

Fig. 6.9 System performance in terms of DER and Clustering Error using 16 initial clusters with MFCC extracted from distant channels.

[1]The reader may refer to Section 5.4.3.1 and the results of Table 5.7.
[2]The reader may refer to Section 5.4.3.2 and the results of Table 5.8.



(a) DER (b) Clustering Error

Fig. 6.10 System performance in terms of DER and Clustering Error using 16 initial clusters with MFCC extracted from best channels.

From those figures, it can be observed that, again, Kmeans-P (CV) appears to have the best initialisation effect. One can infer that cluster purification provide good starting centres for the K-means algorithm. Also, in general, Kmeans-K (CV) presented slightly worse performance compared to Kmeans-P (CV). Similar performance is not achieved using binary key based methods, Kmeans-P (BK) and Kmeans-R (BK), possibly because a cluster mean is not being accurately calculated by the procedure described earlier at the end of Section 6.2.2.

6.2.3.3 Effect of the Number of Initial Clusters on Initialisation

The effect of the number of initial clusters is also of interest and is therefore investigated. A separate part of the AMI corpus collected at the University of Edinburgh is used for this purpose. It is referred to as ES2000 and includes the meetings: ES2002a, ES2003a, ES2004a, ES2006a, ES2007a, ES2009a, ES2011a and ES2012a. The conversations include four speakers each. MFCC features extracted from beamformed signals are used. Using this dataset, Fig. 6.11 shows the system's DER for the cases of 12, 16 and 20 initial clusters.

From Fig. 6.11, the proposed methods provide better performance than uniform initialisation. Except for Kmeans-K (CV) and Kmeans-R (BK), the methods exhibit a similar pattern of performance when varying the number of initial clusters. More specifically, Kmeans-P (BK) provides more consistent performance when the number of initial clusters vary. Purification (BK) presents similar behaviour with a small tendency to provide better performance for a smaller number of initial clusters. Kmeans-P (CV) and Purification (CV) have a positive effect on the performance for a smaller number of initial clusters which accentuates the

importance of the initial clusters purification idea presented here. This is because, especially for the Purification (CV) approach, more homogeneous segments result at the output of the purification process which is important before any cluster merging takes place.

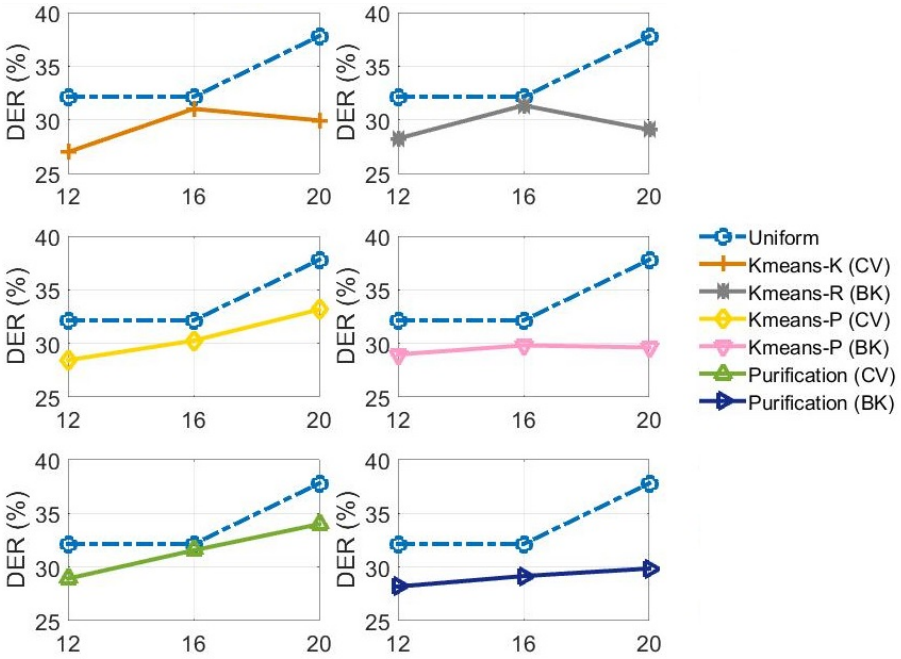
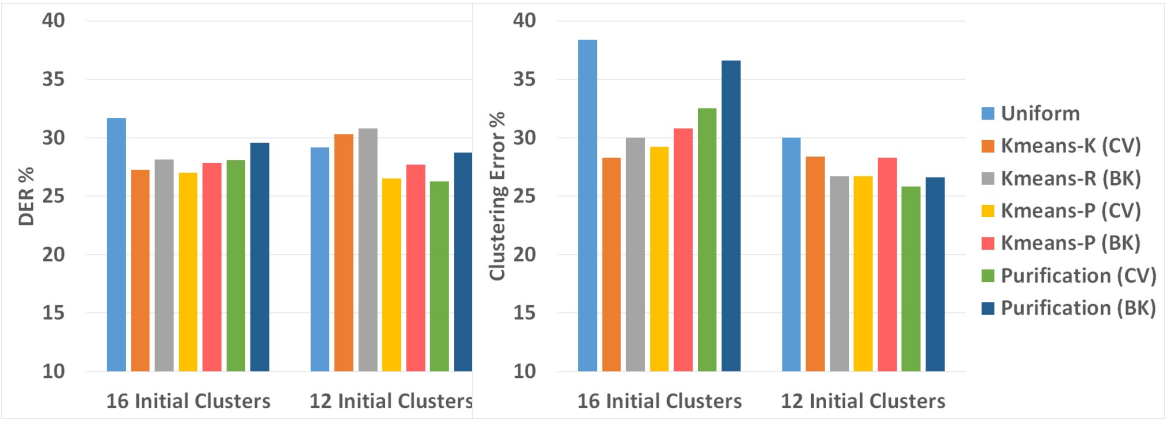


Fig. 6.11 DER for ES2000 dataset illustrates system performance in relation to the initial number of clusters with different initialisation methodologies. The horizontal axis represent the number of initial clusters.

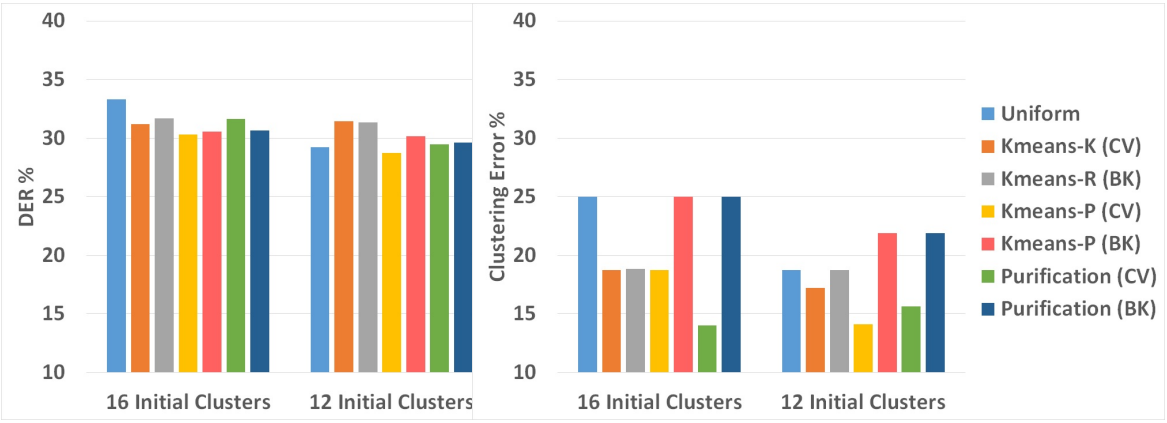
These observations of the the results in Fig. 6.11 were tested on the IS1000, TS3000 and RT-05 datasets. Comparisons between the cases of 16 and 12 initial clusters can be seen in figures 6.12, 6.13 and 6.14 with MFCC features extracted from beamformed signals, distant and best channels, respectively. In terms of DER, the performances with Kmeans-K (CV) and Kmeans-R (BK) are degraded which confirms the observation from the results shown in Fig. 6.11 that those methods do not follow a stable pattern when the number of initial clusters is changed. On the other hand, Kmeans-R (BK) interestingly delivers a reasonably stable performance. For the case of fewer number of initial clusters, the rest of the proposed methodologies provide better performance as expected from the results of Fig. 6.11.

It can also be noticed from figures 6.12, 6.13 and 6.14, that Kmeans-P (CV) and Purification (CV) present better performance than uniform initialisation in most of the cases. Despite the fact that DER with Purification (CV) is slightly worse than the uniform method as shown in Fig. 6.13 (for the case of 12 initial clusters), the Clustering Error provided by that method is noticeably smaller. This can be understood by the fact that lower clustering



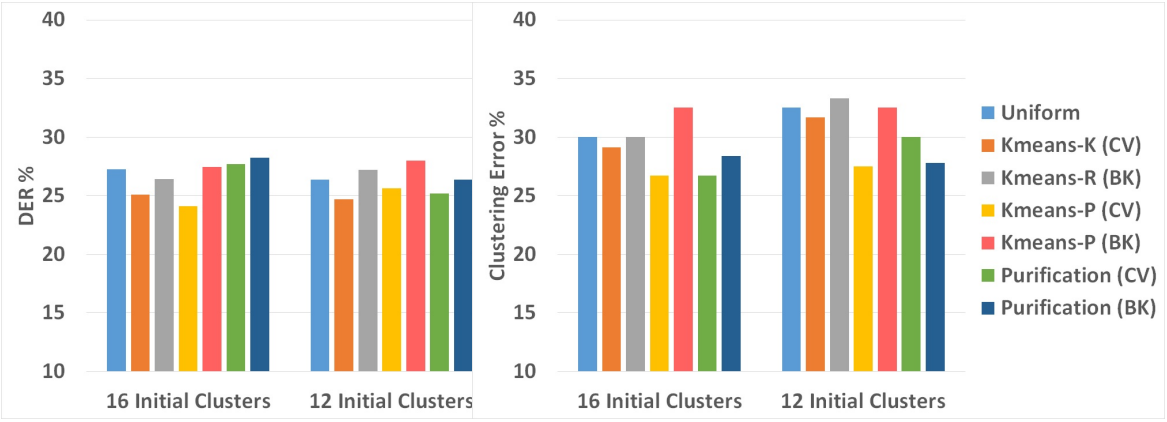
(a) DER (b) Clustering Error

Fig. 6.12 Effect of initial clusters number on system performance with MFCC extracted from beamformed signals for the combination of IS1000, TS3000 and RT-05S datasets.



(a) DER (b) Clustering Error

Fig. 6.13 Effect of initial clusters number on system performance with MFCC extracted from distant channels for the combination of IS1000 and TS3000 datasets.



(a) DER (b) Clustering Error

Fig. 6.14 Effect of initial clusters number on system performance with MFCC extracted from best channels for the combination of IS1000, TS3000 and RT-05S datasets.

errors indicate higher speaker detection accuracy which includes detecting those speakers with a proportionally low contribution to the conversation. In such cases the GMM models trained for those speakers in the final re-segmentation process may not be very reliable, given the small data, which increases the final segmentation errors which would be reflected in the DER.

Number of Initial Clusters	Unifrom	Kmeans-K (CV)	Kmeans-R (BK)	Kmeans-P (CV)	Kmeans-P (BK)	Purification (CV)	Purification (BK)
16	0.0227	0.0231	0.0300	0.0372	0.0384	0.0257	0.0317
12	0.0203	0.0209	0.0269	0.0310	0.0341	0.0231	0.0282

Table 6.8 System computation time in $\times$RT with various initialisation methods for the 16 and 12 initial clusters cases.

Finally, it is important to demonstrate the computational load incurred by the initialisation methods investigated. The choice of the number of initial clusters can vary depending on initial estimates of the number of speakers present in a conversation. With uniform initialisation, it is common to over-cluster a conversation several times higher than the expected number of speakers. However, with reliable initialisation methods, like Kmeans-P (CV), one might not need to over cluster as in the case of uniform initialisation. This can in turn reduce the overall computational load of the system. For the cases of 16 and 12 initial clusters, Table 6.8 reports the overall system computation time in terms of $\times$RT. The data used is the combination of IS1000, TS300 and RT-05S datasets with MFCC features extracted from beamformed signals. From Table 6.8, the simple uniform initialisation provided the

fastest performance. However, the proposed methodologies appear to incur relatively low computational load.

6.2.4 Discussion

Putting aside the way in which K-means is started, its effect on initialisation could have also been affected by concatenating features from different sources (whether distant or best channels) which may, however, confound the process of determining the clusters' centres. This might help to explain why the proposed initialisation methods performed very well for the case of a single source for feature extraction (the beamformed signal case). Features concatenated from different channels can be fused using PCA. In that case, it is anticipated that the proposed methodologies will present similar performance to the case of features extracted from the beamformed signal. This is based on the fact that the fused features (using PCA) are linear combination of all the original features which can remove the confusion in centres determination for the K-means based methodologies.

In terms of DER, Kmeans-P (CV) has proven to be the most robust initialisation method (among the ones proposed) that outperformed uniform initialisation through all the experiments presented. Regarding the Clustering Error, uniform initialisation outperformed Kmeans-P (CV) with the TS3000 dataset in particular for two cases only, see figures 6.9b and 6.10b. The robustness of Kmeans-P (CV) appears to come from two main aspects: determining clusters' means in the cumulative vectors domain appears to be more accurate than the methodology used for achieving the means of binary keys; and having the initial centres as models of purified clusters such that K-means will then serve as a fine tuner for the cluster boundaries.

6.3 Summary

This chapter tackled an interesting aspect for speaker diarization in general. In the presence of multiple microphones, acoustic feature sourcing was shown to improve by making a more appropriate use of the available signals as opposed to combining them in a beamformed signal. Beamforming Anguera et al. (2007) has limited consideration of channels' qualities. For the reverberation problem in particular, MFCC features extracted from selected channels' subbands was found to outperform those extracted from a beamformed signal. On simulated reverberation effects, the Average Joined Gradient (AJG) presented for spectral assessment was found to have direct relation with the degree of reverberation. On practical data, it

is anticipated that AJG provided successful assessment of spectral distortions given the improved performance shown in Table 6.5.

The chapter also addressed an issue related to binary key based diarization in particular which is the system's initialisation. Six methods were presented for binary key based initialisation. Three methods are based on Cumulative Vectors: Purification (CV), Kmeans-K (CV) and Kmeans-P (CV). The other three are based on Binary Keys: Purification (BK), Kmeans-R (BK) and Kmeans-P (BK). In terms of computational load, the complexity of the methods was shown to be accommodating with the fast performance of the binary key based system since they added slight computational time (refer to Table 6.8). The initialisation methods presented outperformed uniform initialisation in the majority of the experiments. Among these methods, Kmeans-P (CV) appeared to be the most robust one.

Chapter 7

Conclusions and Future Work

This chapter summarises the work carried out and the methodologies presented. Due to the different aspects of the contributions, this chapter allocates a separate section for interrelated subjects. Accordingly, the relevant potential future works are included at the end of each section.

7.1 Acoustic Feature Extraction

Acoustic features are a fundamental input to speaker recognition systems. It has been shown in this work that, in particular, the performance of the speaker diarization system investigated is largely influenced by the methodology used to extract the MFCC features. This thesis has presented a reliable improvement on the extraction of MFCC feature which is widely used in speaker recognition systems.

The Discrete Cosine Transform (DCT) used in MFCC extraction approximates the basis function of Karhunen-Loeve transform when the correlation matrix of the input resembles that of a highly correlated Markov-I process (Sahidullah & Saha, 2012). This is commonly considered important for reducing the losses in the DCT transformation. The proposed paradigm for odd-even MFCC (OE-MFCC) highlights the role of the filter bank as a 'transformation' of the speech spectrum as opposed to the DCT transformation of the filters' output. It was shown that using non-overlapped filters in spectrum decomposition results in having lower residual correlation in the correlation matrix of the filters' output. However, this is considered to be more important since the filter bank is designed according to the perceptual mechanism of speech. The application of the DCT to the outputs of the odd and even subsets of filters separately have two main advantages: each subset consists of non-overlapping filters

and the combination of both subsets captures the entire spectrum as it is done in traditional MFCC.

OE-MFCC presented a maximum absolute improvement of $\sim$6% in terms of DER in speaker diarization. For speaker verification, it provided an absolute improvement of $\sim$1% in terms of EER. As investigated in the framework of speaker verification, the system's accuracy using OE-MFCC features does not decrease when the number of filters increases unlike the case with traditional MFCC. This implies that OE-MFCC is more robust to such variations. It is important to note that OE-MFCC depends on the multitaper spectrum estimation where it has provided the best performance in speaker verification and diarization using this spectrum estimation method. The multitaper method provides smoothened spectral estimates which OE-MFCC appears to efficiently use in its lower residual filter bank transformation of the spectrum.

Given the enhancement provided by multitaper spectrum estimation in MFCC and OE-MFCC features extraction, it was fitted in the extraction of LPCC features in this work. The methodology presented for this purpose seamlessly approximated the autocorrelation function required to determine the coefficients of the linear prediction model by taking the inverse of the multitaper spectrum based on the Wiener-Khinchin theorem. Multitaper-Fitted LPCC outperforms traditional LPCC as it was shown in the i-vector based speaker verification framework. It is interesting that Multitaper-fitted LPCC provides the best results using four multipeak tapers. This is because MFCC features have previously provided the best results in (Kinnunen et al., 2010) using the same type and number of tapers.

Since the effects of noise were not part of the scope of this work, it did not include a study of the spectral envelop provided by the linear prediction coefficients which are estimated here using the inverse-multitaper-spectrum approximation of the autocorrelation function. In a future work, it will be interesting to investigate the impact of this approximation on the spectral envelop under the effect of various types of noise. It will also be interesting to investigate the effect of the proposed approximation of the autocorrelation function on sharp peaks in the spectrum of feminine speech. Such specifics were not investigated since the framework of this research is gender-independent.

Regarding OE-MFCC, future work can look into the possibility of having further reduction in the residual correlation of the correlation matrix of the filters output. One might consider a different decimation of the filters other than the odd-even criterion. Also, this work has used triangular shaped filters and it will be interesting to investigate how the methodology might perform using different filter shapes such as rectangular or Hamming filters.

7.2 RNN based Weighted PCA

The imperfect performance of speaker recognition systems can be attributed to a number of factors. One of those factors is that none of the existing features are capable of completely representing discriminative speaker-dependent characteristics. A plausible remedy for this issue is to pool together a number of different features. However, it is required to establish efficient methods for feature combination that can improve the performance with a minimal increment in a system's complexity. This work presented a weighted PCA method for this purpose.

The proposed method provides principal component analysis that allows altering the significance of the feature frames or the different features. By doing so, weighted PCA has two advantages over classical PCA. It is robust to outlying and corrupted feature frames as it was shown in the experiments for speaker verification where it provided a maximum absolute improvement over classical PCA of $\sim$0.5% in terms of EER for feature fusion (MFCC+LPCC & OE-MFCC+LPCC). Additionally, it enables the weighting of the features to be combined as it was shown in the fusion of MFCC and TTDOA features in speaker diarization. In this latter case, it was shown that weighted PCA noticeably outperforms classical PCA by $\sim$8% in terms of DER.

The choice of having the PCA based on the covariance or the correlation matrix may be influenced by the type of features and the speaker recognition modality. The choice of the correlation matrix (equivalently variance normalisation) is safer when there is uncertainty about the features' variances in order to avoid the risk of having non-significant feature attributes of high variance to dominate the principal components. On the other hand, the choice of the covariance matrix could be preferable when it is desired to let some types of features to have higher dominance on the principal components. This can be especially useful when it is needed to have the feature frames weighted since it could be complicated to assign weights to the individual features at the same time.

The RNN framework of extracting the dominant principal component is a fast and an uncomplicated iterative process. It is further developed here to solve for the entire set of the principal components provided that the correlation or covariance matrix remains symmetrical after subtracting the variance represented by a preceding principal component. The RNN framework is found to have relatively high convergence rate. Using prior principal components to start the iterative process is a plausible alternative to using arbitrary initial vectors. In the work of i-vector based speaker verification, the principal components retained by the SVD technique were used to help the network convergence. However, this is not

mandatory for the outcome of the principal component analysis performed by the RNN. Nonetheless, this can be a good choice in some conditions when there exists some prior principal components and one wishes to update them using additional data.

This latter situation can be studied in future work with particular modalities of speaker recognition systems that may require a constant update of the principal components. This work did not consider the fusion case where MFCC and LPCC features are assigned different weights. In a future work, it will be interesting to consider such weighted fusion as well as the fusion of more than two features by giving different weights to each feature. This can be complicated to calibrate but the prospective outcome is expected to be worthwhile. One can also investigate different weighting criterion of the feature frames other than the single Gaussian model and log-likelihood values.

7.3 Spatial Feature Transformation

It is reasonable that a difficult task, like speaker diarization, requires extra resources. Having a multi-speaker conversation, like a meeting, recorded by multiple microphones is useful for speaker diarization. For one aspect, one can exploit the availability of multiple speech signals to help the diarization process by measuring the time delay of arrival (TDOA) of those signals which indicates speakers' locations. The binary key based diarization system was incapable of using TDOA features in the diarization process as it is the case with other diarization systems, see (Martínez-González et al., 2017).

This work conducted an analysis within the framework of binary key based diarization which revealed that the problem resides in the positively skewed distribution of TDOA features. In order to derive discriminative binary keys, it was also found that the normality of the features' distribution is necessary in order to have the Gaussian models (forming the KBM) to be positioned so that the peak of a Gaussian suitably coincides with the centre of the feature density population. In this work, the distribution of TDOA features is assumed to fit a skew-normal distribution which considers a shape parameter that is related to the skewness. The skewness of the distribution of these features seems to be caused by the Generalised Cross Correlation with the Phase Transformation (GPHAT) algorithm used to estimate the delays. However, this work showed that the severity of the skewness can be influenced by other factors like the locations of the speakers which cannot be controlled in practice.

Since the GPHAT algorithm is widely used to estimate TDOA features, this work considered normalising their distribution which is found to enable binary key based diarization using such features. Among the normalising operations investigated, the Box-Cox power

transformation is found to incur the best normalising effect on the distribution of TDOA features as it was measured by high order moments (skewness and kurtosis). This enabled the integration of these spatial features (transformed TDOA) with MFCC features in binary key based diarization which provided 5-10% improvement in terms of DER for three different datasets. The diarization performance using TTDOA features alone is outperformed by the case of only using MFCC features which is plausible since acoustic features are the primary input to diarization systems (Pardo et al., 2007).

The standard Box-Cox transformation is also found to outperform the proposed modified Box-Cox transformation presented in this work. This is because the objective of the proposed modification was to normalise the distribution of individual speakers' features which affected the normality of the entire stream of TTDOA features. However, that outcome confirms the finding in this work about the necessity of providing the binary key based system with normally distributed features. This work considered TDOA features estimated between a central microphone (as a reference) and the rest of the available microphones. It is possible that the degree of skewness of TDOA features distribution is affected by the choice of the reference microphone or any other pairs of microphones to be used to estimate the delays. However, this was not investigated here since any findings are unlikely to generalise to different meeting scenarios, setups or venues.

Alternatively, future work could consider TDOA feature selection depending on their degree of normality before or after their transformation. In such a case, it can be useful to estimate the delays between all possible pairs of microphones in order to provide a broader choice of feature streams. This can, however, be time consuming when there exists a large number of microphones but the selection could be done only once since it is expected to generalise to different meetings given a fixed venue and setup.

Despite the efforts made here, the distribution of transformed TDOA features remains somewhat imperfectly normal. Future work could investigate the possibility of developing a specialised normalisation framework for this purpose. It will also be interesting to consider other algorithms to estimate delays that could potentially be more normally distributed than those provided by the GPHAT algorithm.

7.4 Channel and Channel's Subband Selection

Recording a conversation using multiple microphones (channels) also provides a richer resource for acoustic feature extraction in speaker diarization. Nonetheless, it is computationally expensive to use speech features extracted from all of the channels especially when

there is a plethora of channels. Combining all channels using the beamforming technique is one way to exploit all channels. Acoustic features are then extracted from the combined signal and used in the diarization system. One type of acoustic feature was considered in this diarization framework which is MFCC features. The objective of beamforming is to direct the spatial signal's beam towards the talking speaker. Alternatively, the work presented here investigated the idea of having the acoustic features extracted from selected channels or channels' subbands.

The methodologies proposed here, for different channel selections, provides acoustic features in which their concatenation presents better accuracy in speaker diarization than the features extracted from a beamformed signal. One way to interpret this outcome is that the beamforming technique does not optimally exploit the available channels. The alternatives presented here, in particular, provide the necessary enhancements to improve the performance of binary key based diarization systems. Additionally, the relatively high processing speed of this system makes the resultant increase in feature dimensionality a minor concern. The system's processing time using features extracted from a beamformed signal is around $0.025\times$RT which only increases to a maximum of $0.05\times$RT using a concatenation of features from up to six channels with a noticeable increase in the accuracy. The improvements to be stated here are in relation to using acoustic features extracted from a beamformed signal in the diarization process.

Two methodologies for channel selection have been described. One methodology considers the selection of spatially diverse channels (distant microphones) in an attempt to capture equivalent acoustics to those recorded by all microphones. The acoustic features of the channels selected using this methodology provides an improvement of 3-9% in terms of DER. Two different datasets were used in the evaluation which indicates that the methodology is operational in different scenarios. However, the choice of the optimal number of selected channels may vary from one meeting setup to another. To address this, the methodology considers the selection of the central microphone and those close to it as well as the microphones that are distant from the central microphone. Given the achieved outcome, such configuration appears to provide a diversity among the selected channels and constrains the number of selected channels.

The objective of the second channel selection methodology is to identify best quality channels based on the cepstral distance measure and the beamformed signal as a reference. In the beamforming technique, the speech is assumed to be strengthened and the noise is assumed to be weakened. When it is used as a reference, it is found to provide good discrimination between the quality of the channels. With a maximum of five best quality

channels, this methodology presents an improvement of 2-7% in terms of DER for three different datasets. The relative improvement obtained using this method can be attributed to the fact that, in beamforming, all of the channels are combined irrespective of their qualities. However, using five selected channels may not always be the best choice. For example, when there exists a total of 7-10 channels, selecting five may include bad quality channels. In order to find an improvement over the case of the beamformed signal, one might have to opt for a relatively small number of selected channels.

The channel's subband selection methodology proposed in this work tackled a particular aspect of speech signal's quality which is related to reverberation effects. The selection framework accounts for the variability of the reverberation degree across the speech spectrum that is potentially caused by factors like the microphone location and the surrounding objects. The average joined gradient measure presented is found to be successful in differentiating the amount of reverberation time affecting the speech spectrum. Having this measure determined using a threshold that is calculated from the signals to be compared is particularly useful for its feasibility in different meeting scenarios. The combination of the acoustic features extracted from selected channels' subbands provides an improvement of around 4% in terms of DER for three different datasets. In comparison to using features extracted from the entire spectrum of selected channels, this selection methodology can be particularly useful for slower diarization systems that cannot afford an increase in feature dimensionality.

The reverberation degree can be influenced by speakers' locations. Since each speaker's turn in a conversation is initially unknown, future work could consider channels' subbands selection over segments of the speech signal. It will likely be a good methodology to conduct this selection over segments specified by adjacent silences since there is a good chance that the speakers' turns are changed between the silences. This can additionally apply to best quality channel selection. Future work could also investigate vastly distant microphone selections based on delays estimated between all possible pairs of microphones. Such a framework may also reduce the selection down to two groups of distant microphones but they would potentially be different from the groups selected in this work which could also form the basis of even further work.

7.5 Verification and Diarization Systems Studied

The establishment of the i-vector based verification system requires a large separate bank of data for development purposes to enable the modelling of session and channel variability in particular. It was shown here that the performance of i-vector based speaker verification

noticeably degrades when such external data is insufficient. This work presented a data augmentation method by having copies of the available development data which is then extended with the addition of Gaussian noise in order to increase the number of development utterances. This is found to enable the operation of the i-vector based system with reasonable performance as it was shown with telephone speech evaluation data (e.g the SRE 2002 dataset). However, it must be noted that if the system is already established with sufficient external data, then data augmentation may not present a noticeable impact on the performance. Nonetheless, in a future work, it will be interesting to investigate the performance of data augmentation when more than one copy of each development utterance is made by adding Gaussian noise at different SNRs. Alternatively, the use of a mixture of noise types, such as babble noise and Gaussian noise, can be appropriate in a multi-condition training framework.

The contributions made in the speaker diarization framework are particularly useful for the binary key based system. In this system, the integration of TTDOA features and MFCC features extracted from selected best quality channels have achieved a DER of 24.87% on the RT-05S dataset at 0.051×RT speed. This represents a 19.51% relative improvement on this system's accuracy and slightly outperforms the accuracy of the BIC based diarization system (Anguera & Bonastre, 2011). This is important for speaker diarization because a fast system allows real time application of further processes like a speech recogniser which is one of the objectives of diarization. It is found here that the performance of binary key based diarization is influenced by the initialisation method used. Six initialisation methods were presented in this work to provide suitably selected initial clusters. The experiments show that the commonly used uniform initialisation method can be reliably replaced by the Kmeans-P (CV) method. This method can be particularly important to improve the performance when a conversation is recorded by a single microphone since the resources to improve the systems' front-end are limited in such a case.

In a future work, the selection of the best clustering structure could benefit from innovative configurations. This can be particularly necessary for the case of acoustic and spatial features integration. Also, the system performance can be optimised by calibrating the fusion weights in compliance with specific feature types, dimensionality and meeting conditions. The most computationally expensive process in the binary key based system is calculating the log-likelihoods of all feature vectors to the KBM's Gaussian models. Some experimentation showed that down-sampling the feature vectors by a ratio of 1/2 expedites this process and approximately maintains the performance. This could be further investigated, analysed and justified in a future work.

7.6 Summary

This work introduced a number of enhancements on the front-end of speaker recognition systems that are developed to perform two different tasks: speaker verification and speaker diarization. The achieved enhancements are found to boost the accuracy of binary key based diarization in particular. This is important because a diarization system, especially a fast one like the binary key based, can itself be a 'front-end' for speech processing related applications.

Speaker recognition applications have good potentials to configure a cost effective monitoring system, for example, in a building. This is based on the fact that the speech signal propagates in all directions and this can be exploited to gain information that may not be discernible from a video. Binary key based diarization can form the basis for a monitoring system in an uncontrolled environment. One can make use of the proposed Kmeans-P (CV) initialisation method to start the system. Multiple distant microphones can be deployed to flexibly record the conversations around the building. Thus, an appropriate selection of the microphones' signals can be performed based on one of the methods proposed to secure reliable resources for acoustic feature extraction. Since multiple microphones would be available, spatial features can also be extracted, transformed and integrated in the system with the acoustic features. Acoustic features extracted from multiple microphones can be fused based on weighted PCA. The use of OE-MFCC and multitaper fitted LPCC speech features is recommended.

A small number of microphones might exist in some places which incur a limitation on channel selection. In such case, feature extraction based on least reverberated subband selection is a robust solution. For speaker tracking, speakers' clusters resulting from the diarization performed in a number of places in the building can be compared and matched. When two clusters are found to be matching, the decision can be verified using the i-vector based verification system. The performance of the verification system can be enhanced by retrieving the original speech signals and extracting a number of acoustic features including OE-MFCC. The features extracted can be fused with weighted PCA and used in the verification system. Verified speakers' utterances can be used as annotated data to re-develop the i-vector based system with more utterances.

References

Ahmad, K. S., Thosar, A. S., Nirmal, J. H., & Pande, V. S. (2015). A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network. In *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)* (pp. 1–6). Kolkata, India.

Ajmera, J., & Wooters, C. (2003). A robust speaker clustering algorithm. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)* (pp. 411–416). St Thomas, VI, USA, USA.

Alam, M. J., Kinnunen, T., Kenny, P., Ouellet, P., & O'Shaughnessy, D. (2013). Multitaper MFCC and PLP features for speaker verification using i-vectors. *Speech communication*, *55*, 237–251.

Alku, P., & Saeidi, R. (2017). The linear predictive modeling of speech from higher-lag autocorrelation coefficients applied to noise-robust speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*, 1606–1617.

Anguera, X. (2006). *Robust speaker diarization for meetings*. Universitat Politècnica de Catalunya.

Anguera, X., Aguilo, M., Wooters, C., Nadeu, C., & Hernando, J. (2006a). Hybrid speech/non-speech detector applied to speaker diarization of meetings. In *2006 IEEE Odyssey - The Speaker and Language Recognition Workshop* (pp. 1–6).

Anguera, X., & Bonastre, J.-F. (2010). A novel speaker binary key derived from anchor models. In *Eleventh Annual Conference of the International Speech Communication Association* (pp. 2118–2121). Makuhari, Chiba, Japan.

Anguera, X., & Bonastre, J.-F. (2011). Fast speaker diarization based on binary keys. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (pp. 4428–4431). IEEE.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*, 356–370.

Anguera, X., Woofers, C., & Hernando, J. (2005). Speaker diarization for multi-party meetings using acoustic fusion. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* (pp. 426–431). San Juan, Puerto Rico.

Anguera, X., Wooters, C., & Hernando, J. (2006b). Friends and enemies: A novel initialization for speaker diarization. In *Ninth International Conference on Spoken Language Processing* (pp. 689–692). Barcelona, Spain.

Anguera, X., Wooters, C., & Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*, 2011–2022.

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* SODA '07 (pp. 1027–1035). New Orleans, Louisiana: Society for Industrial and Applied Mathematics.

Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, *64*, 460–475.

Azzalini, A. (2013). *The skew-normal and related families* volume 3. Cambridge University Press.

Bailey, S. (2012). Principal component analysis with noisy and/or missing data. *Publications of the Astronomical Society of the Pacific*, *124*, 1015.

Beigi, H. (2011). *Fundamentals of speaker recognition*. Springer Science & Business Media.

Benesty, J., Sondhi, M. M., & Huang, Y. (2007). *Springer handbook of speech processing*. Springer Science & Business Media.

Besacier, L., & Bonastre, J.-F. (2000). Subband architecture for automatic speaker recognition. *Signal Processing*, *80*, 1245–1259.

Boesch, D., Laboratory, C. E. R., & of Marine Science, V. I. (1977). *Application of numerical classification in ecological investigations of water pollution*. Ecological research series ; EPA-600/3-77-033. Environmental Protection Agency, Office of Research and Development, Corvallis Environmental Research Laboratory.

Bosworth, B. T., Bernecky, W. R., Nickila, J. D., Adal, B., & Carter, G. C. (2008). Estimating signal-to-noise ratio (SNR). *IEEE Journal of Oceanic Engineering*, *33*, 414–418.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 211–252).

Brandstein, M. S., & Silverman, H. F. (1997). A robust method for speech signal time-delay estimation in reverberant rooms. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 375–378). Munich, Germany volume 1.

Broersen, P. M. (2006). *Automatic autocorrelation and spectral analysis*. Springer Science & Business Media.

Campbell, J., & Higgins, A. (1994). YOHO speaker verification corpus ldc94s16. *Available at the LDC website: http://www. ldc. upenn. edu*, .

Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*, *85*, 1437–1462.

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006a). Support vector machines for speaker and language recognition. *Computer Speech & Language*, *20*, 210–229.

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006b). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*, 308–311.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2006). The AMI meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction* MLMI'05 (pp. 28–39). Edinburgh, UK: Springer-Verlag.

Chakroborty, S., Roy, A., & Saha, G. (2007). Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. *International Journal of Signal Processing*, *4*, 114–122.

Chen, J., K. Paliwal, K., & Nakamura, S. (2000). A block cosine transform and its application in speech recognition. In *2000 Interspeech on Spoken Language Processing* (pp. 117–120). volume 4.

Chibelushi, C. C., Mason, J. S., & Deravi, F. (1997). Feature-level data fusion for bimodal person recognition. *IET Conference Proceedings*, (pp. 399–403(4)).

Chiu, C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., & Bacchiani, M. (2018). State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4774–4778). Calgary, AB, Canada.

Church, K., Zhu, W., Vopicka, J., Pelecanos, J., Dimitriadis, D., & Fousek, P. (2017). Speaker diarization: A perspective on challenges and opportunities from theory to practice. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4950–4954). New Orleans, LA, USA.

Cover, T., & Thomas, J. (2012). *Elements of Information Theory*. Wiley.

Cumani, S., & Laface, P. (2017). Nonlinear i-vector transformations for PLDA-based speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*, 908–919.

Dautrich, B., Rabiner, L., & Martin, T. (1983). On the effects of varying filter bank parameters on isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *31*, 793–807.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*, 357–366.

De Vries, N. J., Davel, M. H., Badenhorst, J., Basson, W. D., De Wet, F., Barnard, E., & De Waal, A. (2014). A smartphone-based ASR data collection tool for under-resourced languages. *Speech communication*, *56*, 119–131.

Deco, G., & Obradovic, D. (2012). *An information-theoretic approach to neural computing*. Springer Science & Business Media.

Dehak, N. (2009). *Discriminative and Generative Approaches for Long- and Short-term Speaker Characteristics Modeling: Application to Speaker Verification*. Ph.D. thesis École de technologie supérieure.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*, 788–798.

Delchambre, L. (2014). Weighted principal component analysis: a weighted covariance eigendecomposition approach. *Monthly Notices of the Royal Astronomical Society*, *446*, 3545–3555.

Delgado, H., Anguera, X., Fredouille, C., & Serrano, J. (2015a). Fast single-and cross-show speaker diarization using binary key speaker modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*, 2286–2297.

Delgado, H., Anguera, X., Fredouille, C., & Serrano, J. (2015b). Improved binary key speaker diarization system. In *2015 23rd European Signal Processing Conference (EUSIPCO)* (pp. 2087–2091). Nice, France.

Dmochowski, J., Benesty, J., & Affes, S. (2007). Direction of arrival estimation using the parameterized spatial correlation matrix. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*, 1327–1339.

Doddington, G. (2001). Speaker recognition based on idiolectal differences between speakers. In *Seventh European Conference on Speech Communication and Technology*. Scandinavia.

Dupuy, G., Meignier, S., Deléglise, P., & Esteve, Y. (2014). Recent improvements on ILP-based clustering for broadcast news speaker diarization. In *Odyssey 2014: The Speaker and Language Recognition Workshop*. Zadar, Croatia.

Dupuy, G., Rouvier, M., Meignier, S., & Esteve, Y. (2012). I-vectors and ILP clustering adapted to cross-show speaker diarization. In *Interspeech* (pp. 2174–2177). Portland, OR, USA.

Durbin, J. (1960). The fitting of time-series models. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *28*, 233–244.

Ekman, L. A., Kleijn, W. B., & Murthi, M. N. (2008). Regularized linear prediction of speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*, 65–73.

Ellis, D. P., & Liu, J. (2004). Speaker turn segmentation based on between-channel differences. In *Proceedings of NIST Meeting Recognition Workshop* (p. 12).

Evers, C., Rafaely, B., & Naylor, P. A. (2017). Speaker tracking in reverberant environments using multiple directions of arrival. In *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)* (pp. 91–95). San Francisco, CA, USA.

Falk, T. H., Zheng, C., & Chan, W. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*, 1766–1774.

Fan, Y., & Wang, Q. (2013). Robot. US Patent App. 29/431,926.

Feng, X., Zhang, Y., & Glass, J. (2014). Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1759–1763). Florence, Italy.

Fiscus, J. G., Ajot, J., Michel, M., & Garofolo, J. S. (2006). The rich transcription 2006 spring meeting recognition evaluation. In *International Workshop on Machine Learning for Multimodal Interaction* (pp. 309–322). Bethesda, MD, USA: Springer.

Fiscus, J. G., Radde, N., Garofolo, J. S., Le, A., Ajot, J., & Laprun, C. (2005). The rich transcription 2005 spring meeting recognition evaluation. In *International Workshop on Machine Learning for Multimodal Interaction* (pp. 369–389). Springer.

Flores, C. G., Tryfou, G., & Omologo, M. (2018). Cepstral distance based channel selection for distant speech recognition. *Computer Speech & Language*, *47*, 314–332.

Fredouille, C., & Senay, G. (2006). Technical improvements of the E-HMM based speaker diarization system for meeting records. In *International Workshop on Machine Learning for Multimodal Interaction* (pp. 359–370). Springer.

Gacula Jr, M. C. (2013). *Statistical methods in food and consumer research*. Elsevier.

Garau, G., & Bourlard, H. (2010). Using audio and visual cues for speaker diarisation initialisation. In *International Conference on Acoustics, Speech and Signal Processing* LIDIAP-CONF-2010-001 (pp. 4942–4945). Dallas, Texas, USA.

Garcia-Romero, D., & Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech* (pp. 249–252). volume 2011.

Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., & McCree, A. (2017). Speaker diarization using deep neural network embeddings. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4930–4934). New Orleans, LA, USA.

Garcia-Romero, D., Zhou, X., & Espy-Wilson, C. Y. (2012). Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4257–4260).

Gaydecki, P., & of Electrical Engineers, I. (2004). *Foundations of Digital Signal Processing: Theory, Algorithms and Hardware Design*. IEE circuits and systems series: Institution of Electrical Engineers. Institution of Engineering and Technology.

Giri, R., Seltzer, M. L., Droppo, J., & Yu, D. (2015). Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5014–5018). Brisbane, QLD, Australia.

Gonina, E., Friedland, G., Cook, H., & Keutzer, K. (2011). Fast speaker diarization using a high-level scripting language. In *2011 IEEE Workshop on Automatic Speech Recognition Understanding* (pp. 553–558). Waikoloa, HI, USA.

González, B. M., Muñoz, J. M. P., Correa, J. D. E., Pinto, J. Á. V., & Chicote, R. B. (2012). Selection of TDOA parameters for MDM speaker diarization. In *InterSpeech 2012, 13th Annual Conference of the International Speech Communication Association* (pp. 1–4). Portland, Oregon, USA.

Guerrero, C., Tryfou, G., & Omologo, M. (2016). Channel selection for distant speech recognition exploiting cepstral distance. In *INTERSPEECH* (pp. 1986–1990). San Francisco, USA.

Guizzo, E. (2018). Anki's vector is a little AI-powered robot now on kickstarter for $200. *IEEE Spectrum*, . URL: https://spectrum.ieee.org/automaton/robotics/home-robots/anki-vector-is-a-little-ai-powered-robot-now-on-kickstarter.

Hanilci, C., Kinnunen, T., Ertas, F., Saeidi, R., Pohjalainen, J., & Alku, P. (2012). Regularized all-pole models for speaker verification under noisy environments. *IEEE Signal Processing Letters*, *19*, 163–166.

Hartmann, W. M. (2004). *Signals, sound, and sensation*. Springer Science & Business Media.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, *87*, 1738–1752.

Himawan, I., Motlicek, P., Sridharan, S., Dean, D., & Tjondronegoro, D. (2015). Channel selection in the short-time modulation domain for distant speech recognition. In *Proceedings of Interspeech - Annual Conference of the International Speech Communication Association* (pp. 741–745). Dresden, Germany.

Houdré, C., Mason, D., Reynaud-Bouret, P., & Rosiński, J. (2016). *High Dimensional Probability VII: The Cargèse Volume*. Progress in Probability. Springer International Publishing.

Houtgast, T., & Steeneken, H. J. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, *77*, 1069–1077.

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*. Sage.

Hu, H.-T. (1998). Robust linear prediction of speech signals based on orthogonal framework. *Electronics Letters*, *34*, 1385–1386.

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, *47*, 64–79.

Imseng, D., & Friedland, G. (2009). Robust speaker diarization for short speech recordings. In *2009 IEEE Workshop on Automatic Speech Recognition Understanding* (pp. 432–437). Merano, Italy.

Imseng, D., & Friedland, G. (2010). An adaptive initialization method for speaker diarization based on prosodic features. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4946–4949). Dallas, Texas, USA.

Ismail, M. R. (2013). A parametric investigation of the acoustical performance of contemporary mosques. *Frontiers of Architectural Research*, *2*, 30 – 41.

Jiang, Y., Wang, D., Liu, R., & Feng, Z. (2014). Binaural classification for reverberant speech segregation using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*, 2112–2121.

Jo, J., Yoo, H., & Park, I.-C. (2016). Energy-efficient floating-point MFCC extraction architecture for speech recognition systems. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *24*, 754–758.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer.

Jones, M. C., & Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society. Series A (General)*, (pp. 1–37).

Joshi, A., Kumar, M., & Das, P. K. (2016). Speaker diarization: A review. In *2016 International Conference on Signal Processing and Communication (ICSC)* (pp. 191–196). Noida, India.

Kajarekar, S. S. (2005). Four weightings and a fusion: a cepstral-SVM system for speaker recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* (pp. 17–22). San Juan, Puerto Rico.

Kantz, H., & Schreiber, T. (2004). *Nonlinear Time Series Analysis*. Cambridge nonlinear science series. Cambridge University Press.

Kendall, M. G. (1949). The estimation of parameters in linear autoregressive time series. *Econometrica: Journal of the Econometric Society*, *17*, 44–57.

Kenny, P. (2006). Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13*, *14*, 28–29.

Kenny, P. (2010). Bayesian speaker verification with heavy-tailed priors. In *Odyssey* (p. 14). Brno, Czech Republic.

Kenny, P., Boulianne, G., & Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, *13*, 345–354.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2004). Speaker adaptation using an eigenphone basis. *IEEE Transactions on Speech and Audio Processing*, *12*, 579–589.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*, 1435–1447.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., & Dumouchel, P. (2008). A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*, 980–988.

Kenny, P., Stafylakis, T., Ouellet, P., Alam, M. J., & Dumouchel, P. (2013). PLDA for speaker verification with utterances of arbitrary duration. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7649–7653).

Kheder, W. B., Matrouf, D., Ajili, M., & Bonastre, J.-F. (2016). Probabilistic approach using joint long and short session i-vectors modeling to deal with short utterances for speaker recognition. In *Interspeech* (pp. 1830–1834). San Francisco, USA.

Khosravani, A., & Homayounpour, M. M. (2018). Nonparametrically trained PLDA for short duration i-vector speaker verification. *Computer Speech & Language*, *52*, 105–122.

Kim, S., Ji, M., & Kim, H. (2008). Noise-robust speaker recognition using subband likelihoods and reliable-feature selection. *ETRI journal*, *30*, 89–100.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, *52*, 12–40.

Kinnunen, T., Saeidi, R., Sandberg, J., & Hansson-Sandsten, M. (2010). What else is new than the Hamming window? Robust MFCCs for speaker recognition via multitapering. In *Eleventh Annual Conference of the International Speech Communication Association* (pp. 2734–2737). Japan.

Kinnunen, T., Wu, Z., Lee, K. A., Sedlak, F., Chng, E. S., & Li, H. (2012). Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4401–4404).

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., & Maas, R. (2013). The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on* (pp. 1–4). New Paltz, NY, USA: IEEE.

Kitawaki, N., Itoh, K., Honda, M., & Kakehi, K. (1982). Comparison of objective speech quality measures for voiceband codecs. In *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1000–1003). Paris, France volume 7. doi:10.1109/ICASSP.1982.1171566.

Kitawaki, N., Nagabuchi, H., & Itoh, K. (1988). Objective quality evaluation for low-bit-rate speech coding systems. *IEEE Journal on Selected Areas in Communications*, *6*, 242–248.

Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*, 320–327.

Koh, E. C. W., Sun, H., Nwe, T. L., Nguyen, T. H., Ma, B., Chng, E.-S., Li, H., & Rahardja, S. (2008). 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). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kumar, P., Jakhanwal, N., Bhowmick, A., & Chandra, M. (2011). Gender classification using pitch and formants. In *Proceedings of the 2011 International Conference on Communication, Computing & Security* ICCCS '11 (pp. 319–324). Rourkela, Odisha, India: ACM.

Kwok, J. T., Mak, B., & Ho, S. (2004). Eigenvoice speaker adaptation via composite kernel principal component analysis. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16* (pp. 1401–1408). MIT Press.

Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, *13*, 293–303.

Lee, Y., Lee, J., & Lee, K. (2002). GMM based on local robust PCA for speaker identification. In *NINTH AUSTRALIAN INTERNATIONAL CONFERENCE ON SPEECH SCIENCE AND TECHNOLOGY* (pp. 273–278). Melbourne, Australia.

van Leeuwen, D. A. (2006). The TNO speaker diarization system for NIST RT05S meeting data. In *Machine Learning for Multimodal Interaction* (pp. 440–449). Berlin, Heidelberg: Springer Berlin Heidelberg.

van Leeuwen, D. A., & Huijbregts, M. (2006). The AMI speaker diarization system for NIST RT06S meeting data. In *Machine Learning for Multimodal Interaction* (pp. 371–384). Berlin, Heidelberg: Springer Berlin Heidelberg.

van Leeuwen, D. A., & Konečnỳ, M. (2008). Progress in the AMIDA speaker diarization system for meeting data. In *Multimodal Technologies for Perception of Humans* (pp. 475–483). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lehiste, I. (1976). Chapter 7 - suprasegmental features of speech. In N. J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 225 – 239). Academic Press.

Li, M., Han, K. J., & Narayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, *27*, 151 – 167. Special issue on Paralinguistics in Naturalistic Speech and Language.

Liu, Y., Qian, Y., Chen, N., Fu, T., Zhang, Y., & Yu, K. (2015). Deep feature for text-dependent speaker verification. *Speech Communication*, *73*, 1 – 13.

Luque, J., Segura, C., & Hernando, J. (2008). Clustering initialization based on spatial information for speaker diarization of meetings. In *Ninth Annual Conference of the International Speech Communication Association* (pp. 383–386). Brisbane, Australia.

Ma, C., Kamp, Y., & Willems, L. (1993). Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, *12*, 69 – 81.

Madikeri, S., Himawan, I., Motlicek, P., & Ferras, M. (2015). Integrating online i-vector extractor with information bottleneck based speaker diarization system. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Mak, M.-W., Pang, X., & Chien, J.-T. (2016). Mixture of PLDA for noise robust i-vector speaker verification. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *24*, 130–142.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*, 561–580.

Makhoul, J., & Cosell, L. (1976). LPCW: An LPC vocoder with linear predictive spectral warping. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'76*. (pp. 466–469). IEEE volume 1.

Malik, H., & Farid, H. (2010). Audio forensics from acoustic reverberation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1710–1713). Dallas, TX, USA.

Markel, J., Oshika, B., & Gray, A. (1977). Long-term feature averaging for speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *25*, 330–337.

Martin, A., & Mark, P. (2004). 2002 NIST speaker recognition evaluation LDC2004S04. URL: https://catalog.ldc.upenn.edu/LDC2004S04.

Martin, A. F., & Greenberg, C. S. (2010). The NIST 2010 speaker recognition evaluation. In *Eleventh Annual Conference of the International Speech Communication Association* (pp. 2726–2729). Japan.

Martinez, D., Bürget, L., Stafylakis, T., Lei, Y., Kenny, P., & Lleida, E. (2014). Unscented transform for ivector-based noisy speaker recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4042–4046). Florence, Italy.

Martínez-González, B., Pardo, J. M., Echeverry-Correa, J. D., & San-Segundo, R. (2017). Spatial features selection for unsupervised speaker segmentation and clustering. *Expert Systems with Applications*, *73*, 27–42.

McLaren, M., & Lei, Y. (2015). Improved speaker recognition using DCT coefficients as features. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4430–4434). Brisbane, QLD, Australia.

Meriem, F., Farid, H., Messaoud, B., & Abderrahmene, A. (2017). New front end based on multitaper and Gammatone filters for robust speaker verification. In *Recent Advances in Electrical Engineering and Control Applications* (pp. 344–354). Cham: Springer International Publishing.

Milner, B., & Shao, X. (2007). Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*, 24–33. doi:10.1109/TASL.2006.876880.

Moattar, M., & Homayounpour, M. (2012). A review on speaker diarization systems and approaches. *Speech Communication*, *54*, 1065 – 1103.

Neustein, A., & Patil, H. A. (2012). *Forensic speaker recognition*. Springer.

Novoselov, S., Pekhovsky, T., Kudashev, O., Mendelev, V. S., & Prudnikov, A. (2015). Non-linear PLDA for i-vector speaker verification. In *Sixteenth Annual Conference of the International Speech Communication Association*. Dresden, Germany.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, *15*, 267–273.

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks*, *5*, 927 – 935.

Omar, N. M., & El-Hawary, M. E. (2017). Feature fusion techniques based training MLP for speaker identification system. In *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)* (pp. 1–6).

Oo, Z., Wang, L., Phapatanaburi, K., Iwahashi, M., Nakagawa, S., & Dang, J. (2018). Phase and reverberation aware DNN for distant-talking speech enhancement. *Multimedia Tools and Applications*, *77*, 18865–18880.

Pal, M., & Saha, G. (2017). Spectral mapping using prior re-estimation of i-vectors and system fusion for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*, 2071–2084.

Parada, P. P., Sharma, D., van Waterschoot, T., & Naylor, P. A. (2017). Robust statistical processing of TDOA estimates for distant speaker diarization. In *Signal Processing Conference (EUSIPCO), 2017 25th European* (pp. 86–90). IEEE.

Pardo, J., Anguera, X., & Wooters, C. (2007). Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Transactions on Computers*, *56*, 1212–1224.

Pardo, J. M., Anguera, X., & Wooters, C. (2006). Speaker diarization for multi-microphone meetings using only between-channel differences. In *Machine Learning for Multimodal Interaction* (pp. 257–264). Berlin, Heidelberg: Springer Berlin Heidelberg.

Paul, D., Pal, M., & Saha, G. (2017). Spectral features for synthetic speech detection. *IEEE Journal of Selected Topics in Signal Processing*, *11*, 605–617.

Peso, P. (2016). Spatial features of reverberant speech: estimation and application to recognition and diarization, .

Plapous, C., Marro, C., & Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 2098–2108.

Pohjalainen, J., & Alku, P. (2013). Extended weighted linear prediction using the autocorrelation snapshot-a robust speech analysis method and its application to recognition of vocal emotions. In *14thAnnual Conference of the International Speech Communication Association, INTERSPEECH 2013* (pp. 1931–1935). France.

Poularikas, A. D. (2010). *Transforms and applications handbook*. CRC press.

Prieto, G., Parker, R., Thomson, D., Vernon, F., & Graham, R. (2007). Reducing the bias of multitaper spectrum estimates. *Geophysical Journal International*, *171*, 1269–1281.

Prince, S. J. D., & Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision* (pp. 1–8).

Pruzansky, S. (1963). Pattern-matching procedure for automatic talker recognition. *The Journal of the Acoustical Society of America*, *35*, 354–358.

Przybocki, M. A. (2011 accessed December 13, 2018). *Speaker Recognition*. https://www.nist.gov/itl/iad/mig/speaker-recognition.

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). "Alexa is my new bff": Social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* CHI EA '17 (pp. 2853–2859). Denver, Colorado, USA: ACM.

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

Rabiner, L. R., & Juang, B.-H. (1990). *Fundamentals of speech recognition*. Prentice Hall Signal Processing Series. New Jersey: PTR Prentice Hall.

Rabiner, L. R., Juang, B.-H., & Rutledge, J. C. (1993). *Fundamentals of speech recognition* volume 14. PTR Prentice Hall Englewood Cliffs.

Rajan, P., Afanasyev, A., Hautamäki, V., & Kinnunen, T. (2014). From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification. *Digital Signal Processing*, *31*, 93 – 101.

Rajan, P., Kinnunen, T., & Hautamäki, V. (2013). Effect of multicondition training on i-vector PLDA configurations for speaker recognition. In *Interspeech* (pp. 3694–3697). Lyon, France: Citeseer.

Rajasekaran, S., & Pai, G. V. (2002). Recurrent neural dynamic models for equilibrium and eigenvalue problems. *Mathematical and computer modelling*, *35*, 229–240.

Rao, S. V., Shah, N. J., & Patil, H. A. (2016). Novel pre-processing using outlier removal in voice conversion. In *9th ISCA Speech Synthesis Workshop* (pp. 147–152). Sunnyvale, CA, USA.

Rao, W., & Mak, M.-W. (2013). Boosting the performance of i-vector based speaker verification via utterance partitioning. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*, 1012–1022.

Rencher, A. C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician*, *46*, 217–225.

Reyes-Galaviz, O. F., & García, C. A. R. (2009). Fuzzy relational compression applied on feature vectors for infant cry recognition. In *MICAI* (pp. 420–431). Springer.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, *10*, 19–41.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on Speech and Audio Processing*, *3*, 72–83.

Ribas, D., Vincent, E., & Calvo, J. R. (2015). Full multicondition training for robust i-vector based speaker recognition. In *Interspeech* (p. 1057–1061). Dresden, Germany.

Rosenberg, A. E. (1976). Automatic speaker verification: A review. *Proceedings of the IEEE*, *64*, 475–487.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, *79*, 871–880.

Rouvier, M., & Meignier, S. (2012). A global optimization framework for speaker diarization. In *Odyssey 2012*. Singapore.

Roweis, S. T. (1998). EM algorithms for PCA and SPCA. In *Advances in neural information processing systems* (pp. 626–632).

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, *47*, 69–76.

Sadjadi, S. O., Slaney, M., & Heck, L. (2013). MSR identity toolbox v1. 0: A MATLAB toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter*, *1*, 1–32.

Safavi, S., Russell, M., & Jančovič, P. (2018). Automatic speaker, age-group and gender identification from children's speech. *Computer Speech & Language*, *50*, 141 – 156.

Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Communication*, *54*, 543–565.

Sarkar, A. K., Do, C.-T., Le, V.-B., & Barras, C. (2014). Combination of cepstral and phonetically discriminative features for speaker verification. *IEEE Signal Processing Letters*, *21*, 1040–1044.

Sarria-Paja, M., & Falk, T. H. (2018). Fusion of bottleneck, spectral and modulation spectral features for improved speaker verification of neutral and whispered speech. *Speech Communication*, . URL: http://www.sciencedirect.com/science/article/pii/S0167639317304703.

Sarria-Paja, M., Senoussaoui, M., O'Shaughnessy, D., & Falk, T. H. (2016). Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5480–5484). Shanghai, China.

Sell, G., & Garcia-Romero, D. (2014). Speaker diarization with PLDA i-vector scoring and unsupervised calibration. In *2014 IEEE Spoken Language Technology Workshop (SLT)* (pp. 413–417). South Lake Tahoe, NV, USA.

Seo, C., Youn, J., Kim, Y., Sim, K., Ko, J., Zhao, M., Kim, J., Ko, H., Kim, E., & Lim, Y. (2009). A global covariance matrix based principal component analysis for speaker identification. In *15th Asia-Pacific Conference on Communications* (pp. 245–248).

Shannon, B. J., & Paliwal, K. K. (2006). Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition. *Speech Communication*, *48*, 1458 – 1485. Robustness Issues for Conversational Interaction.

Sharma, A. (2005). *Text book of correlations and regression*. DPH mathematics series. Discovery Publishing House.

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. crc Press.

Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D., & Glass, J. (2011). Exploiting intra-conversation variability for speaker diarization. In *Twelfth Annual Conference of the International Speech Communication Association*. Florence, Italy.

Siegler, M. A., Jain, U., Raj, B., & Stern, R. M. (1997). Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA speech recognition workshop*. volume 1997.

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In *Interspeech* (pp. 999–1003). Stockholm, Sweden.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *ICASSP*. Calgary.

So, S., & Paliwal, K. K. (2008). Quantization of speech features: Source coding. In *Automatic Speech Recognition on Mobile Devices and over Communication Networks* (pp. 131–161). London: Springer London.

Song, Y., Jiang, B., Bao, Y., Wei, S., & Dai, L. (2013). I-vector representation based on bottleneck features for language identification. *Electronics Letters*, *49*, 1569–1570.

Soong, F., Rosenberg, A., Rabiner, L., & Juang, B. (1985). A vector quantization approach to speaker recognition. In *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 387–390). Tampa, FL, USA volume 10.

Sturim, D. E., Reynolds, D. A., Singer, E., & Campbell, J. P. (2001). Speaker indexing in large audio databases using anchor models. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)* (pp. 429–432 vol.1). Salt Lake City, UT, USA volume 1.

Teunen, R., Shahshahani, B., & Heck, L. (2000). A model-based transformational approach to robust speaker recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. ISCA.

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, *70*, 1055–1096.

Tibrewala, S., & Hermansky, H. (1997). Sub-band based recognition of noisy speech. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1255–1258). volume 2.

Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, *90*, 250 – 271.

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

Tu, Y.-H., Du, J., Wang, Q., Bao, X., Dai, L.-R., & Lee, C.-H. (2017). An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech. *Computer Speech & Language*, *46*, 517–534.

Valin, J. ., & Lefebvre, R. (2000). Bandwidth extension of narrowband speech for low bit-rate wideband coding. In *2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat. No.00EX421)* (pp. 130–132). Delavan, WI, USA, USA.

Vijayasenan, D., & Valente, F. (2012). Speaker diarization of meetings based on large TDOA feature vectors. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4173–4176). Kyoto, Japan.

Vijayasenan, D., Valente, F., & Bourlard, H. (2007). Agglomerative information bottleneck for speaker diarization of meetings data. In *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)* (pp. 250–255). Kyoto, Japan.

Vijayasenan, D., Valente, F., & Bourlard, H. (2008). Integration of TDOA features in information bottleneck framework for fast speaker diarization. In *Ninth Annual Conference of the International Speech Communication Association*.

Vijayasenan, D., Valente, F., & Bourlard, H. (2011a). An information theoretic combination of MFCC and TDOA features for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*, 431–438.

Vijayasenan, D., Valente, F., & Motlicek, P. (2011b). Multistream speaker diarization through information bottleneck system outputs combination. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4420–4423). Prague, Czech Republic.

Villalba, J., & Lleida, E. (2013). Handling i-vectors from different recording conditions using multi-channel simplified PLDA in speaker recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6763–6767). Vancouver, BC, Canada.

de Vries, N. J., Davel, M. H., Badenhorst, J., Basson, W. D., de Wet, F., Barnard, E., & de Waal, A. (2014). A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication*, *56*, 119 – 131.

Vu, N.-V., Whittington, J., Ye, H., & Devlin, J. (2010). Implementation of the MFCC front-end for low-cost speech recognition systems. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on* (pp. 2334–2337). France: IEEE.

Weinberg, S. L., & Abramowitz, S. K. (2008). *Statistics using SPSS: An integrative approach*. Cambridge University Press.

Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, *51*, 2044–2056.

Wolf, M., & Nadeu, C. (2010). On the potential of channel selection for recognition of reverberated speech with multiple microphones. In *Eleventh Annual Conference of the International Speech Communication Association* (pp. 574–577). Makuhari, Chiba, Japan.

Wolf, M., & Nadeu, C. (2014). Channel selection measures for multi-microphone speech recognition. *Speech Communication*, *57*, 170 – 180.

Woubie, A., Luque, J., & Hernando, J. (2015). Using voice-quality measurements with prosodic and spectral features for speaker diarization. In *Sixteenth Annual Conference of the International Speech Communication Association* (pp. 3100–3104). Dresden, Germany.

Wu, Q., Merchant, F., & Castleman, K. (2010). *Microscope image processing*. Academic press.

Wu, Z., Xiao, X., Chng, E. S., & Li, H. (2013). Synthetic speech detection using temporal modulation feature. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7234–7238). Vancouver, BC, Canada.

Yegnanarayana, B., & Kishore, S. P. (2002). AANN: an alternative to GMM for pattern recognition. *Neural Networks*, *15*, 459–469.

Yi, Z., Fu, Y., & Tang, H. J. (2004). Neural networks based approach for computing eigenvectors and eigenvalues of symmetric matrix. *Computers & Mathematics with Applications*, *47*, 1155–1164.

Yu, D., & Deng, L. (2014). *Automatic Speech Recognition: A Deep Learning Approach*. Signals and Communication Technology. London: Springer.

Zeinali, H., Sameti, H., & Burget, L. (2017a). HMM-based phrase-independent i-vector extractor for text-dependent speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*, 1421–1435.

Zeinali, H., Sameti, H., Burget, L. et al. (2017b). Text-dependent speaker verification based on i-vectors, Neural Networks and Hidden Markov Models. *Computer Speech & Language*, *46*, 53 – 71.

Zhang, C., Li, X., Li, W., Lu, P., & Zhang, W. (2016). A novel i-vector framework using multiple features and PCA for speaker recognition in short speech condition. In *2016 International Conference on Audio, Language and Image Processing (ICALIP)* (pp. 499–503). China: IEEE.

Zhang, Y., Chuangsuwanich, E., & Glass, J. (2014). Extracting deep neural network bottleneck features using low-rank matrix factorization. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 185–189). Florence, Italy.

Zhang, Y., Xu, J., Yan, Z.-J., & Huo, Q. (2011). An i-vector based approach to training data clustering for improved speech recognition. In *Twelfth Annual Conference of the International Speech Communication Association*. Florence, Italy.

Zhou, G., & Mikhael, W. B. (2006). Speaker identification based on adaptive discriminative vector quantisation. *IEE Proceedings - Vision, Image and Signal Processing*, *153*, 754–760.

Appendix A

A.1 Baum-Welch Statistics

The foundation work that led to the introduction of the i-vector representation of speech utterances considered a Gaussian Mixture Model based Hidden Markov Model (GMM-HMM) as a Universal Background Model (UBM) to represent the global acoustic space, see e.g Kenny et al. (2005). The Baum-Welch statistics, also known as the forward-backward algorithm, is the maximum likelihood estimation algorithm for an HMM Beigi (2011).

In the specific problem of text-independent speaker recognition, there is no textual structure to be modelled by the transition probabilities between an HMM states. Therefore, the GMM-UBM model introduced by Reynolds et al. (2000) was widely considered to represent the global acoustic space. The forward-backward algorithm is an application of the Expectation Maximisation (EM) algorithm. It was used in the extraction of the i-vector in Dehak et al. (2011) to adapt the speaker-and-channel independent supervector obtained from the GMM-UBM to an utterance's feature vectors to produce an utterance-dependent supervector.

A.2 The i-vector Extraction

This appendix section provides the derivation of equation (2.26) which is used to extract the i-vector. Let $\mathbf{Y}^u = (\mathbf{y}_1^u, \mathbf{y}_2^u, ..., \mathbf{y}_T^u)$ be a sequence of feature vectors of an utterance u. To extract the i-vector for this utterance, recall the following factor analysis model

$$\mathbf{m}^u = \mathbf{m} + \mathbf{T}\mathbf{w}^u, \tag{A.1}$$

where $\mathbf{T}$ is a low rank rectangular matrix of dimensions $C \times D$ by F. C is the number of components in the GMM-UBM model Λ and D is the feature vectors dimension. $\mathbf{w}^u$ is an

F-dimensional random vector having a prior distribution of a standard normal distribution $\mathcal{N}(.;0,\mathbf{I})$. Each mixture component c, for $c = 1,2,...,C$, of Λ has a D dimensional mean vector, $\boldsymbol{\mu}_c$, a D by D covariance matrix, $\boldsymbol{\Sigma}_c$, and a weight w_c. $\mathbf{m}$ is a $C \times D$ speaker-and-utterance independent supervector obtained by concatenating all $\boldsymbol{\mu}_c$ of Λ. $\mathbf{m}^u$ is an utterance dependent supervector.

Let $\boldsymbol{\Sigma}$ be a $C \times D$ by $C \times D$ matrix of diagonal blocks $\boldsymbol{\Sigma}_c$ for $c = 1,2,...,C$. Given $\mathbf{Y}^u$, $\mathbf{m}^u$ and $\boldsymbol{\Sigma}$, the i-vector is the *MAP* solution of the following problem

$$\begin{aligned} \hat{\mathbf{w}}^u &= \arg\max_{\mathbf{w}^u} \mathscr{P}(\mathbf{w}^u|\mathbf{Y}^u) \\ &= \arg\max_{\mathbf{w}^u} \mathscr{P}(\mathbf{Y}^u|\mathbf{w}^u)\mathscr{P}(\mathbf{w}^u), \end{aligned} \tag{A.2}$$

where

$$\mathscr{P}(\mathbf{Y}^u|\mathbf{w}^u) = \prod_{t=1}^{T}\sum_{c=1}^{C} w_c \mathcal{N}(\mathbf{y}_t^u;\mathbf{m}_c^u,\boldsymbol{\Sigma}_c), \tag{A.3}$$

where $\mathbf{m}_c^u$ is the c^{th} D-dimensional sub-vector of $\mathbf{m}^u$. Equation (A.3) is intractable because of the summation element, hence, the solution of (A.2) becomes a solution of the following problem (Zhang et al., 2011)

$$\hat{\mathbf{w}}^u = \arg\max_{\mathbf{w}^u} \prod_{t=1}^{T}\prod_{c=1}^{C} \mathcal{N}(\mathbf{y}_t^u;\mathbf{m}_c^u,\boldsymbol{\Sigma}_c)^{\mathscr{P}(c|\mathbf{y}_t^u,\Lambda)}\, \mathscr{P}(\mathbf{w}^u), \tag{A.4}$$

where

$$\mathscr{P}(c|\mathbf{y}_t^u,\Lambda) = \frac{w_c \mathcal{N}(\mathbf{y}_t^u;\mathbf{m}_c,\boldsymbol{\Sigma}_c)}{\sum_{l=1}^{C} w_l \mathcal{N}(\mathbf{y}_t^u;\mathbf{m}_l,\boldsymbol{\Sigma}_l)}. \tag{A.5}$$

By comparing equations (A.4) and (A.2), the solution for $\mathscr{P}(\mathbf{Y}^u|\mathbf{w}^u)$ is

$$\log \mathscr{P}(\mathbf{Y}^u|\mathbf{w}^u) = \sum_{t=1}^{T}\sum_{c=1}^{C} \mathscr{P}(c|\mathbf{y}_t^u,\Lambda) \log \mathcal{N}(\mathbf{y}_t^u;\mathbf{m}_c^u,\boldsymbol{\Sigma}_c). \tag{A.6}$$

The right hand side of (A.1) is used to substitute $\mathbf{m}_c^u$, thus

$$\begin{aligned}\log \mathscr{P}(\mathbf{Y}^u|\mathbf{w}^u) &= \sum_{t=1}^{T}\sum_{c=1}^{C} \mathscr{P}(c|\mathbf{y}_t^u,\Lambda)\Big[\log\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_c|^{1/2}} \\ &\quad -\frac{1}{2}(\mathbf{y}_t^u-\mathbf{m}_c-\mathbf{T}_c\mathbf{w}^u)^T\mathbf{\Sigma}_c^{-1}(\mathbf{y}_t^u-\mathbf{m}_c-\mathbf{T}_c\mathbf{w}^u)\Big] \\ &= \sum_{t=1}^{T}\sum_{c=1}^{C} \mathscr{P}(c|\mathbf{y}_t^u,\Lambda)\Big[\log\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_c|^{1/2}} - \frac{1}{2}(\mathbf{y}_t^u-\mathbf{m}_c)^T\mathbf{\Sigma}_c^{-1}(\mathbf{y}_t^u-\mathbf{m}_c) \\ &\quad +(\mathbf{w}^u)^T\mathbf{T}_c^T\mathbf{\Sigma}_c^{-1}(\mathbf{y}_t^u-\mathbf{m}_c) - \frac{1}{2}(\mathbf{w}^u)^T\mathbf{T}_c^T\mathbf{\Sigma}_c^{-1}\mathbf{T}_c\mathbf{w}^u\Big],\end{aligned} \tag{A.7}$$

where $\mathbf{T}_c$ is a sub-matrix of $\mathbf{T}$ associated with mixture component c of Λ. Note that $\mathbf{T}$ (of (A.1)) comprises C component matrices stacked up column wise such that $\mathbf{T} = [\mathbf{T}_1^T,\mathbf{T}_2^T,...,\mathbf{T}_C^T]^T$. Also note that $\mathbf{m}_c = \boldsymbol{\mu}_c$. Only the last two terms of (A.7) are related to $\mathbf{w}^u$ and one can define

$$\mathscr{H}^u = \sum_{t=1}^{T}\sum_{c=1}^{C} \mathscr{P}(c|\mathbf{y}_t^u,\Lambda)\Big[(\mathbf{w}^u)^T\mathbf{T}_c^T\mathbf{\Sigma}_c^{-1}(\mathbf{y}_t^u-\mathbf{m}_c) - \frac{1}{2}(\mathbf{w}^u)^T\mathbf{T}_c^T\mathbf{\Sigma}_c^{-1}\mathbf{T}_c\mathbf{w}^u\Big]. \tag{A.8}$$

The calculations in (A.8) include, for each mixture component c, the computations of the 0^{th} order Baum-Welch statistics

$$\mathbf{n}_c^u = \sum_{t=1}^{T} \mathscr{P}(c|\mathbf{y}_t^u,\Lambda), \tag{A.9}$$

and the centralised 1^{st} order Baum-Welch statistics

$$\tilde{\mathbf{f}}_c^u = \sum_{t=1}^{T} \mathscr{P}(c|\mathbf{y}_t^u,\Lambda)(\mathbf{y}_t^u-\mathbf{m}_c). \tag{A.10}$$

Hence, One can re-write (A.8) as

$$\mathscr{H}^u = (\mathbf{w}^u)^T\mathbf{T}\mathbf{\Sigma}^{-1}\tilde{\mathbf{f}}^u - \frac{1}{2}(\mathbf{w}^u)^T\mathbf{T}^T\mathbf{N}^u\mathbf{\Sigma}^{-1}\mathbf{T}\mathbf{w}^u. \tag{A.11}$$

Using (A.11), one can write

$$\begin{aligned}\mathscr{P}(\mathbf{w}^u|\mathbf{Y}^u) &\propto \mathscr{P}(\mathbf{Y}^u|\mathbf{w}^u)\mathscr{P}(\mathbf{w}^u)\\ &\propto \exp\left((\mathbf{w}^u)^T\mathbf{T}\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}}^u - \frac{1}{2}(\mathbf{w}^u)^T\mathbf{T}^T\mathbf{N}^u\boldsymbol{\Sigma}^{-1}\mathbf{T}\mathbf{w}^u\right)\exp\left(-\frac{1}{2}(\mathbf{w}^u)^T\mathbf{w}^u\right)\\ &= \exp\left((\mathbf{w}^u)^T\mathbf{T}\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}}^u - \frac{1}{2}(\mathbf{w}^u)^T\left[\mathbf{T}^T\mathbf{N}^u\boldsymbol{\Sigma}^{-1}\mathbf{T}+\mathbf{I}\right]\mathbf{w}^u\right)\\ &\propto \left(-\frac{1}{2}\left(\mathbf{w}^u - \left[\mathbf{T}^T\mathbf{N}^u\boldsymbol{\Sigma}^{-1}\mathbf{T}+\mathbf{I}\right]^{-1}\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}}^u\right)^T\left[\mathbf{T}^T\mathbf{N}^u\boldsymbol{\Sigma}^{-1}\mathbf{T}+\mathbf{I}\right]\right.\\ &\qquad\left.\left(\mathbf{w}^u - \left[\mathbf{T}^T\mathbf{N}^u\boldsymbol{\Sigma}^{-1}\mathbf{T}+\mathbf{I}\right]^{-1}\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}}^u\right)\right)\end{aligned} \tag{A.12}$$

Therefore, the posterior distribution of $\mathbf{w}^u$ is Gaussian with mean $\left[\mathbf{T}^T\mathbf{N}^u\boldsymbol{\Sigma}^{-1}\mathbf{T}+\mathbf{I}\right]^{-1}\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}}^u$, determined in equation (2.26), and covariance $\left[\mathbf{T}^T\mathbf{N}^u\boldsymbol{\Sigma}^{-1}\mathbf{T}+\mathbf{I}\right]^{-1}$.

A.3 Multitaper Methods

This appendix section provides the bases for the multitaper methods investigated in this work, see Alam et al. (2013).

Thomson Method

In this method, a set of M orthonormal tapers is specified from the Slepian sequences. These are the solutions of the following eigenvalue problem

$$\mathbf{A}\boldsymbol{\lambda}_j^p = v^p\boldsymbol{\lambda}_n^p, \tag{A.13}$$

where $0 \le n \le N-1$, $0 \le j \le N-1$, $\mathbf{A}$ is a real symmetric matrix and $0 < v^p \le 1$ is the p^{th} eigenvalue associated with p^{th} eigenvector, $\boldsymbol{\lambda}_n^p$, known as the Slepian taper. N is the length of the speech frame. The elements of the matrix $\mathbf{A}$ are given by

$$a_{nj} = \frac{\sin 2\pi W(n-j)}{\pi(n-j)}, \tag{A.14}$$

where W is the half-frequency bandwidth. Each taper weight is $w_p = 1/M$.

Sine Multitaper

In this method, a taper is given by

$$\lambda_p(j) = \sqrt{\frac{2}{N+1}} \sin\left(\frac{\pi p(j+1)}{N+1}\right) \quad \text{for} \quad j = 0,1,2,...,N-1, \tag{A.15}$$

where p, $1 \leq p \leq M$, is the order of a taper. The taper weight is determined as in the following

$$w_p = \frac{\cos\left(\frac{2\pi(p-1)}{M/2}\right)+1}{\sum_{p=1}^{M}\left(\cos\left(\frac{2\pi(p-1)}{M/2}\right)+1\right)}. \tag{A.16}$$

Multipeak Multitaper

These are peak matched multiple windows. The multipeak tapers are obtained by solving the following generalised eigenvalue problem

$$\mathbf{R}_B\lambda_j = v_j\mathbf{R}_z\lambda_j \quad \text{for} \quad j = 1,2,...,N, \tag{A.17}$$

where $\mathbf{R}_B$ is the N by N Toeplitz covariance matrix of the following assumed spectrum model

$$S(f) = \begin{cases} e^{\frac{2L|f|}{10B\log_{10}(e)}} & \text{for} \quad |f| \leq B/2 \\ 0 & \text{for} \quad |f| > B/2, \end{cases} \tag{A.18}$$

with $L = 20$ dB and a predetermined interval of width B outside of which spectral leakage is to be prevented. $\mathbf{R}_z$ is the Toeplitz matrix, chosen for decreasing the leakage from the sidelobes of the tapers, of the following frequency penalty function

$$S_Z(f) = \begin{cases} G & \text{for} \quad |f| > B/2 \\ 1 & \text{for} \quad |f| \leq B/2, \end{cases} \tag{A.19}$$

where $G = 30$ dB. The eigenvectors corresponding to the M largest eigenvalues of (A.17) comprise the set of multipeak multitaper. Using the eigenvalues, each taper's weight is determined as

$$w_p = \frac{v_p}{\sum_{p=1}^{M} v_p} \quad \text{for} \quad p = 1,2,...,M. \tag{A.20}$$

A.4 Cross Correlation

The cross-correlation indicates the matching between two signals and it is determined as the summation of the product of the signals (Gaydecki & of Electrical Engineers, 2004). It is usually used to estimate the similarity between two measurements where the higher the cross-correlation the more similar the measurements. The cross-correlation between two signals is calculated as

$$y[t] = \sum_{n=0}^{N} x_1[n]x_2[t+n] \tag{A.21}$$

A.5 High Order Moments of Distributions

High order moments of distributions assess the departure of a distribution from normality (Hoyle, 1995) and are used to identify a feasible transformation. Skewness is the third order moment which characterises the symmetry of the distribution. The fourth order moment, kurtosis, characterises how much the height of the distribution is different from that of a normal distribution. In other words, it characterises the peakedness of the distribution, thus, it is also used to measure the unimodality or bimodality of the distribution (Gacula Jr, 2013). When the value of the kurtosis is 3 and the skewness is 0, the distribution is said to be normal and it can be sufficiently described only by its mean and variance (the first and second order moments, respectively) (Hoyle, 1995).

A.6 The Skew-Normal Distribution

In the skew-normal distribution (Azzalini, 2013), the probability density function (PDF) of a continuous random variable z_n is expressed as

$$\mathscr{G}_{SN} = 2\phi_{SN}(z_n)\hat{\phi}_{SN}(\alpha z_n), \tag{A.22}$$

where

$$\phi_{SN}(z_n) = \frac{e^{(-z_n^2/2)}}{\sqrt{2\pi}}, \tag{A.23}$$

and

$$\hat{\phi}_{SN}(\alpha z_n) = \int_{-\infty}^{\alpha z_n} \phi_{SN}(z_n)dt, \tag{A.24}$$

where α is called the shape parameter. The variable z_n and its probability density function $\mathscr{G}_{SN}$ are basic components of the skew-normal distribution. In the formulation of the skew-normal

distribution (Azzalini, 2013), a practical measurement, let it be q_n, is a linear transformation of z_n

$$q_n = \tilde{\mu} + \tilde{\sigma} z_n, \tag{A.25}$$

where $\tilde{\mu}$ and $\tilde{\sigma}$ are called the location and scale parameters, respectively. Note that a linear transformation does not change the shape of the distribution so α remains constant.

A.7 Least-Squares Scoring on the Principal Components

The power iteration method for weighted PCA was presented in (Delchambre, 2014). In that work the authors described, for comparison proposes, the implementation of an Expectation-Maximisation (EM) approach for weighted PCA. That EM based approach viewed the eigendecomposition as a minimisation problem of the following formula

$$\mathscr{X}^2 = \sum_{ij} \mathbf{W}_{ij}^2 \left(\mathbf{X}_{ij} - [\mathbf{P}_w \hat{\mathbf{X}}]_{ij} \right)^2 \tag{A.26}$$

where i and j indicate the rows and columns indices, respectively. $\hat{\mathbf{X}}$ is the projection (scores) of $\mathbf{X}$ on $\mathbf{P}_w$, where $\mathbf{X}$ is the feature vectors matrix and $\mathbf{P}_w$ is the entire set of principal components. $\mathbf{W}$ is the weights matrix.

While $\mathbf{P}_w$ is held fixed, the expectation step retrieves $\hat{\mathbf{X}}$ that optimises (A.26). Since each feature vectors of $\mathbf{X}$ is a linear combination of the principal components, the solution for $\hat{\mathbf{X}}$ is given by

$$\hat{\mathbf{X}}_j = (\mathbf{P}_w^T \mathbf{w}^2 \mathbf{P}_w)^{-1} \mathbf{P}_w^T \mathbf{w}^2 \mathbf{X}_j \tag{A.27}$$

where $\hat{\mathbf{X}}_j$ is a column of $\hat{\mathbf{X}}$, $\mathbf{X}_j$ is a column of $\mathbf{X}$ and $\mathbf{w} = \text{diag}(\mathbf{W}_j)$.

The maximisation step, on the other hand, retrieves $\mathbf{P}_w$ that optimises (A.26), for a given $\hat{\mathbf{X}}$, as in the following

$$\mathbf{P}_{w,i} = \mathbf{X}_i \mathbf{w}^2 \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \mathbf{w}^2 \hat{\mathbf{X}}^T) \tag{A.28}$$

where $\mathbf{P}_{w,i}$ is a row of $\mathbf{P}_w$ and $\mathbf{X}_i$ is a row of $\mathbf{X}$.

In the RNN based framework presented in this work, the scores (projection) of $\mathbf{X}$ on $\mathbf{P}_w$ can be achieved using (A.27). However, since (A.27) is based on the formulation in (A.26), the use of (A.27) may only be valid when $\mathbf{P}_w$ is estimated from $\mathbf{X}$.

Appendix B

Research Ethics Review Checklist (Form UPR16) and Certificate

FORM UPR16

Research Ethics Review Checklist

Please include this completed form as an appendix to your thesis (see the Postgraduate Research Student Handbook for more information

Postgraduate Research Student (PGRS) Information			**Student ID:**	444944
PGRS Name:	AHMED AHMED			
Department:	School of Energy and Electronic Engineering	**First Supervisor:**	Dr John Chiverton	
Start Date: (or progression date for Prof Doc students)	**02/02/2015**			
Study Mode and Route:	Part-time ☐ Full-time ☒	MPhil ☐ PhD ☒	MD ☐ Professional Doctorate ☐	

Title of Thesis:	Enhancing the Front-End of Speaker Recognition Systems
Thesis Word Count: (excluding ancillary data)	~ 45000

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES ☒ NO ☐
b) Have all contributions to knowledge been acknowledged?	YES ☒ NO ☐
c) Have you complied with all agreements relating to intellectual property, publication and authorship?	YES ☒ NO ☐
d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES ☒ NO ☐
e) Does your research comply with all legal, ethical, and contractual requirements?	YES ☒ NO ☐

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	ETHIC-2019-181

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

Signed *(PGRS)*:		**Date:** 17/01/2019

Certificate of Ethics Review

Project Title: Enhancing the Front-End of Speaker Recognition Systems

Name: Ahmed Ahmed **User ID:** 566637 **Application Date:** 17-Jan-2019 19:53 **ER Number:** ETHIC-2019-181

You must download your certificate, print a copy and keep it as a record of this review.

It is your responsibility to adhere to the University Ethics Policy and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers and University Health and Safety Policy.

It is also your responsibility to follow University guidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

You are reminded that as a University of Portsmouth Researcher you are bound by the UKRIO Code of Practice for Research; any breach of this code could lead to action being taken following the University's Procedure for the Investigation of Allegations of Misconduct in Research.

Any changes in the answers to the questions reflecting the design, management or conduct of the research over the course of the project must be notified to the Faculty Ethics Committee. **Any changes that affect the answers given in the questionnaire, not reported to the Faculty Ethics Committee, will invalidate this certificate.**

This ethical review should not be used to infer any comment on the academic merits or methodology of the project. If you have not already done so, you are advised to develop a clear protocol/proposal and ensure that it is independently reviewed by peers or others of appropriate standing. A favourable ethical opinion should not be perceived as permission to proceed with the research; there might be other matters of governance which require further consideration including the agreement of any organisation hosting the research.

(A1) Please briefly describe your project:: **A number of enhancements on the front-end of speaker verification and diarization systems are introduced in this project. This is achieved by tackling the methods of acoustic feature extraction and feature combination and by proposing a source selection of the speech signal and spatial feature transformation for speaker diarization.**
(A2) What faculty do you belong to?: **Technology**
(A3) I am sure that my project requires ethical review by my Faculty Ethics Committee because it includes at least one material ethical issue.: **No**
(A5) Has your project already been externally reviewed?: **No**
(B1) Is the study likely to involve human participants?: **No**
(B2) Are you certain that your project will not involve human subjects or participants?: **Yes**
(C6) Is there any risk to the health & safety of the researcher or members of the research team beyond those that have already been risk assessed?: **No**
(D2) Are there risks of damage to physical and/or ecological environmental features?: **No**
(D4) Are there risks of damage to features of historical or cultural heritage (e.g. impacts of study techniques, taking of samples)?: **No**
(E1) Will the study involve the investigator and/or any participants in activities that could be considered contentious, unacceptable, or illegal, or in any other way harmful to the reputation of the University of Portsmouth?: **No**
(E2) Are there any potentially socially or culturally sensitive issues involved? (e.g. sexual, political, legal/criminal or financial): **No**
(F1) Does the project involve animals in any way?: **No**
(F2) Could the research outputs potentially be harmful to third parties?: **No**
(G1) Please confirm that you have read the University Ethics Policy and have considered the implications for your project.: **Confirmed**
(G2) Please confirm that you have read the UK RIO Code of Practice for Research and will conduct your project in accordance with it.: **Confirmed**
(G3) The University is committed to The Concordat to Support Research Integrity.: **Confirmed**
(G4) Submitting false or incorrect information is a breach of the University Ethics Policy and may be considered as misconduct and be subject to disciplinary action. Please confirm you understand this and agree that the information you have entered is correct.: **Confirmed**