Improving Automatic Speaker Verification Using Front-end and Back-end diversity

Jia Min Karen Kua

A thesis submitted in fulfilment
of the requirement for the degree of
Doctor of Philosophy

The University of New South Wales
School of Electrical Engineering and Telecommunications
Sydney, Australia

March 2012

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Kua

First name: Jia Min Karen

Other name/s:

Abbreviation for degree as given in the University calendar: PhD

School: Electrical Engineering & Telecommunication

Faculty: Engineering

Title: Improving Automatic Speaker Verification Using Front-end and Back-end diversity

Abstract 350 words maximum: (PLEASE TYPE)

Technologies that exploit biometrics can potentially be applied to the identification and verification of individuals for controlling access to secured areas or materials. Among these technologies, automatic speaker verification systems are of growing interest, as they are the least invasive and they allow recognition via any type of communication network over long distances. The overall goal of this thesis is to improve the performance of automatic speaker verification systems by investigating novel features and classification methods that complement current state-of-the-art systems.

At the feature level, novel log-compressed least squares group delay and spectral centroid features are proposed. The log-compression and least squares regularisation are shown to reduce the dynamic range of modified group delay features and outperform other existing group delay extraction methods. The proposed spectral centroid features provide a better characterisation of spectral energy distribution and experimental results show that the detailed spectral characterisation significantly improves performance.

A diverse front-end involving multiple features would improve both phonetic (acoustic) and speaker modelling. In this regard, the relative contributions of the acoustic and speaker modelling 'stages' on the speaker recognition performance across different features are investigated. The investigation conducted through the use of clustering comparison measures suggests that front-end diversity, and hence improved performance from fused systems, can be achieved purely through different 'partitioning' of the acoustic space. Built on the finding, a novel universal background model (UBM) data/utterance selection algorithm that increases stability of the acoustic modelling is proposed.

Finally, at the classification level, the use of the sparse representation classification (SRC) using Gaussian mixture model supervectors (GMM-SRC) is proposed and is found to perform comparably to Gaussian mixture model-support vector machines (GMM-SVM). However, GMM-SRC results in a slower verification process. In order to increase the computation efficiency, the large dimensional supervectors are replaced with speaker factors resulting in the joint factor analysis-sparse representation classification (JFA-SRC). In addition, a novel dictionary composition technique to further improve the computation efficiency is developed. Results demonstrate that the refined dictionary provide comparable performance over the use of the complete dataset and generalises well to the evaluation on other databases. Notably, a detailed comparison of the proposed JFA-SRC across various state-of-the-art classifiers on the NIST 2010 databases showed that the proposed JFA-SRC achieved the best Minimum Detection Cost Function (minDCF), highlighting the usefulness of the SRC-based systems.

ORIGINALITY STATEMENT

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed ………………………………………………..............

Date ………………………………………………..............

COPYRIGHT STATEMENT

AUTHENTICITY STATEMENT

Abstract

Technologies that exploit biometrics can potentially be applied to the identification and verification of individuals for controlling access to secured areas or materials. Among these technologies, automatic speaker verification systems are of growing interest, as they are the least invasive and they allow recognition via any type of communication network over long distances. The overall goal of this thesis is to improve the performance of automatic speaker verification systems by investigating novel features and classification methods that complement current state-of-the-art systems.

At the feature level, novel log-compressed least squares group delay and spectral centroid features are proposed. The log-compression and least squares regularisation are shown to reduce the dynamic range of modified group delay features and outperform other existing group delay extraction methods. The proposed spectral centroid features provide a better characterisation of spectral energy distribution and experimental results show that the detailed spectral characterisation significantly improves performance.

A diverse front-end involving multiple features would improve both phonetic (acoustic) and speaker modelling. In this regard, the relative contributions of the acoustic and speaker modelling ‘stages’ on the speaker recognition performance across different features are investigated. The investigation conducted through the use of clustering comparison measures suggests that front-end diversity, and hence improved performance from fused systems, can be achieved purely through different ‘partitioning’ of the acoustic space. Built on the finding, a novel universal background model (UBM)

data/utterance selection algorithm that increases stability of the acoustic modelling is proposed.

Finally, at the classification level, the use of the sparse representation classification (SRC) using Gaussian mixture model supervectors (GMM-SRC) is proposed and is found to perform comparably to Gaussian mixture model-support vector machines (GMM-SVM). However, GMM-SRC results in a slower verification process. In order to increase the computation efficiency, the large dimensional supervectors are replaced with speaker factors resulting in the joint factor analysis-sparse representation classification (JFA-SRC). In addition, a novel dictionary composition technique to further improve the computation efficiency is developed. Results demonstrate that the refined dictionary provide comparable performance over the use of the complete dataset and generalises well to the evaluation on other databases. Notably, a detailed comparison of the proposed JFA-SRC across various state-of-the-art classifiers on the NIST 2010 databases showed that the proposed JFA-SRC achieved the best Minimum Detection Cost Function (minDCF), highlighting the usefulness of the SRC-based systems.

Acknowledgements

This thesis would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study. First and foremost, my utmost gratitude to my supervisor, Professor Eliathamby Ambikairajah, for his unfailing supervision, guidance and advice. Above all and the most needed, for lifting me up whenever I am on the brink of giving up. Also, special thanks must go to my joint-supervisor, Dr. Julien Epps, for his motivation, enthusiasm, numerous suggestions and ideas, and specifically for his patient proofreading of all my technical writings. I would also like to thank my co-supervisor, Dr. Eric Choi, for his guidance.

I am indebted to my many colleagues at the UNSW Signal Processing research group for providing a stimulating and fun environment in which to learn and grow. I am especially grateful to Dr. Hadis Nosratighods, Dr. Thiruvaran Tharmarajah, Dr. Phu Ngoc Le, Dr. Bo Yin, Dr. Ronny Kurniawan, Dr. Ning Wang, Dr. Reji Mathew, Dr. Aous Naman, Dr. Wenliang Lu, Phyu Phyu Khing, Xi Chen, Pega Zarjam, Siyuan Chen, Stefanie Brown, Tet Fei Yap, Qingqing Meng, Shuai Wang, Kun Yu, Chee Cheun Huang, Nicholas Cummins, Jonathan Gan and Tom Millet. Special thanks go to Dr. Vidhyasaharan Sethu and Raji Ambikairajah for their emotional support, company and proofreading this thesis.

I also wish to acknowledge National ICT Australia for awarding me the NICTA International Postgraduate Award which enabled me to pursue my PhD at UNSW and the

School of Electrical Engineering and Telecommunications, UNSW, for supporting me during this period.

Lastly, and most importantly, I wish to express my love and gratitude to my family, who has always been my pillar of strength. To my parents and my brothers, thanks for the unconditional love and support that you have given me all these years. To them I dedicate this thesis.

Acronyms and Abbreviations

AM	**A**mplitude **M**odulation
ASCM	**A**nti-**S**pectral **C**entroid **M**agnitude
ASV	**A**utomatic **S**peaker **V**erification
CDS	**C**osine **D**istance **S**coring
CMS	**C**epstral **M**ean **S**ubtraction
DARPA	**D**efense **A**dvanced **R**esearch **P**rojects **A**gency
DCT	**D**iscrete **C**osine **T**ransform
DET	**D**etection **E**rror **T**rade-off
DFT	**D**iscrete **F**ourier **T**ransform
EARS	**E**ffective, **A**ffordable, **R**eusable **S**peech-to-text
EER	**E**qual **E**rror **R**ate
EM	**E**xpectation **M**aximisation
F	**F**ormant
FFT	**F**ast **F**ourier **T**ransform
FM	**F**requency **M**odulation
FoCal	**F**usi**o**n and **Cal**ibration
GD	**G**roup **D**elay
GLDS	**G**eneralised **L**inear **D**iscriminant **S**equence

GMM	**G**aussian **M**ixture **M**odel
HTK	**H**idden markov model **T**ool**K**it
JFA	**J**oint **F**actor **A**nalysis
KL	**K**ullback**-L**eibler
LDA	**L**inear **D**iscriminant **A**nalysis
LDC	**L**inguistic **D**ata **C**onsortium
LLR	**L**og **L**ikelihood **R**atio
LogLSGD	**Log-**Compressed **L**east **S**quares **G**roup **D**elay
LP	**L**inear **P**rediction
LPCC	**L**inear **P**rediction **C**epstral **C**oefficients
MAP	**M**aximum ***A-P****osteriori*
MFCC	**M**el **F**requency **C**epstral **C**oefficients
MI	**M**utual **I**nformation
minDCF	**Min**imum **D**etection **C**ost **F**unction
MLE	**M**aximum **L**ikelihood **E**stimation
MLLR	**M**aximum **L**ikelihood **L**inear **R**egression
MODGD	**MOD**ified **G**roup **D**elay
MSE	**M**ean **S**quared **E**rror
NAP	**N**uisance **A**ttribute **P**rojection
NID	**N**ormalised **I**nformation **D**istance
NIST	**N**ational **I**nstitute of **S**cience and **T**echnology
PDF	**P**robability **D**ensity **F**unction

PLP	**P**erceptual **L**inear **P**rediction
RASTA	**R**el**A**tive **S**pec**T**r**A**ls
SCF	**S**pectral **C**entroid **F**requency
SCM	**S**pectral **C**entroid **M**agnitude
SRC	**S**parse **R**epresentation **C**lassification
SRE	**S**peaker **R**ecognition **E**valuation
SSC	**S**ubband **S**pectral **C**entroid
SVM	**S**upport **V**ector **M**achines
T-norm	**T**est-**n**ormalisation
UBM	**U**niversal **B**ackground **M**odel
VAD	**V**oice **A**ctivity **D**etector
WCCN	**W**ithin-**C**lass **C**ovariance **N**ormalisation
Z-norm	**Z**ero-**n**ormalisation

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Overview

Biometrics is the study of automated methods for uniquely recognising humans based upon one or more intrinsic physical or behavioral traits [1]. With the rapidly expanding research in the area of biometric technologies, biometric authentication is preferred over traditional methods that involve the use of passwords or cards that may be forgotten, stolen or lost. Currently on the market, there are numerous types of biometrics recognition systems such as fingerprints, eye retinas, facial patterns and voice [2]. Among them, speech as a biometric, often referred to as voice biometric or speaker recognition, has become an attractive biometric modality for two key reasons. The first reason is the non-invasive nature of acquiring speech for authentication which does not require direct contact with the individual, unlike fingerprint recognition systems. The other reason is the ability to provide speech samples for authentication remotely and conveniently through telephony-based technologies – both wired and unwired without the need for specialised and expensive equipment which is useful for industrial applications such as telephone direct banking services [3].

Generally, speaker recognition is classified into two specific tasks: identification and verification. Automatic speaker identification is the process of determining which registered speaker provides a given speech and automatic speaker verification is the process of accepting or rejecting the identity claimed by a speaker. Speaker verification is the overarching problem studied in this thesis.

The operation of a speaker verification system typically involves two distinct phases, enrollment and verification as shown in Figure 1.1. Common to both phases is the feature extraction stage which involves the transformation of the raw speech signal into feature vectors which carry speaker discriminative information. In the enrollment phase, a speaker model is trained for each target speaker using its feature vectors. In the verification phase, the feature vectors extracted from an unknown person's utterance are compared against the model in the system database to give a similarity score. The decision module uses this similarity score to make the final decision.

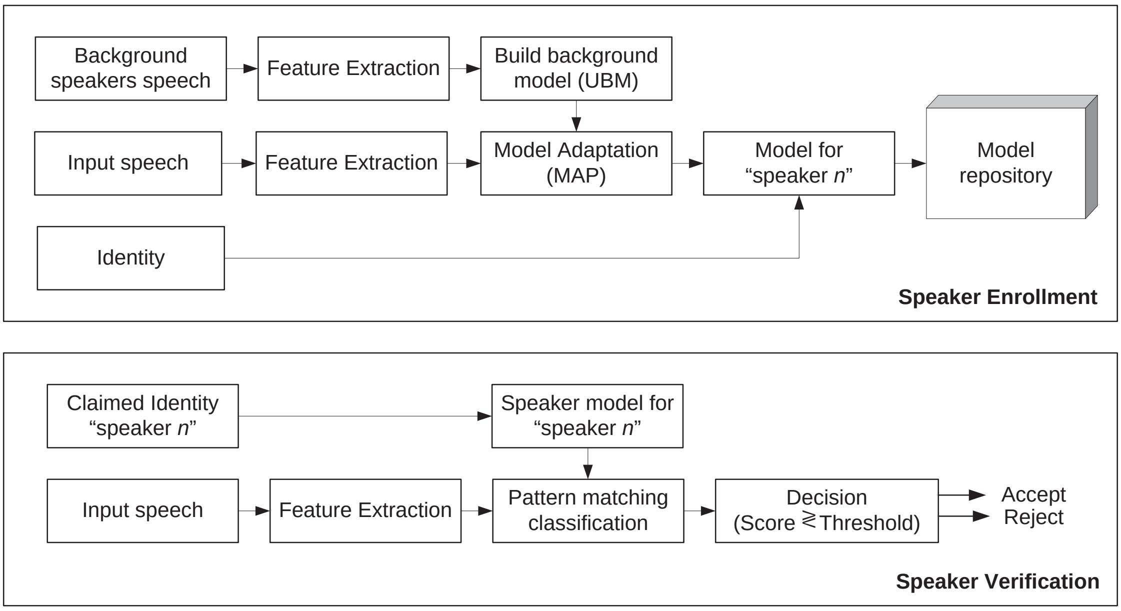


Figure 1.1 Components of a typical automatic speaker verification system.

In recent years, the majority of the state-of-the-art speaker verification systems have utilised mel-frequency cepstral coefficients (MFCC) as features and a Gaussian mixture model (GMM) based approach for speaker modelling [4-6], where each speaker is represented by a GMM which involves the modelling of the speaker's features probability distribution as a weighted sum of Gaussian mixture components. Although many distinct techniques have been developed in the area of speaker recognition [4, 7], the use of GMMs for modelling acoustic features has become almost exclusive. The most commonly used GMM-based speaker recognition methods include the classical maximum a-posteriori (MAP) adaptation of universal background model parameters (GMM-UBM) [8] and support vector machine (SVM) classification of GMM supervectors (GMM-SVM) [9]. The MAP adaptation framework provides a way of incorporating prior information in the training process by adapting the GMM parameters from a generalised speaker model (UBM), which is particularly useful for dealing with problems posed by sparse training data. An SVM is basically a two-class classifier that uses a non-linear function to map data on to a higher (maybe infinite) dimensional space and finds the best hyperplane separating the two classes in this space.

While to date, high accuracy speaker verification has been achieved under ideal conditions and is suitable for practical implementation under matched channel conditions, the performance degrades significantly under mismatched conditions. Mismatch can occur due to the use of different telephone handsets or different acoustic environments between the acquired training and test speech samples. The difficulties associated with compensating for these differences have presented an active research topic for the speaker verification field in recent years and some of the state-of-the-art channel compensation schemes include joint factor analysis (JFA) [10], *i-vectors* [11] or nuisance attribute projection (NAP) [12]. The basic idea of JFA is to decompose a speaker's supervector into speaker independent, speaker dependent, channel dependent and residual components, whereby the channel dependent component will be removed during verification (as a form of channel compensation). On the other hand, *i-vectors* involve the decomposition of the speaker's supervector into two components, speaker independent and total variability, where channel compensation is performed through Linear

Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) on the *i-vectors* after decomposition. NAP works by removing dimensions from the SVM expansion space that are irrelevant to the speaker recognition problem. These schemes have been used in state-of-the-art systems as can be seen from the recent 2010 National Institute of Science and Technology (NIST) speaker recognition evaluations (SRE) [5, 13-15].

In an effort to advance the current state-of-the-art systems in the last few years, a significant amount of work has centred around variations of JFA and NAP [4, 16-18]. However, another possible avenue for improving system performance exists - using the combination of information from different sources of evidence, termed as fusion. Within the context of speaker verification, fusion refers to the combination of scores based on different models trained for a speaker [19] and these models could be trained using different speech features and/or classifiers [20-23]. In the former case, phase-based and frequency-based features [24-26] which have recently shown to be successful as complementary features for magnitude-based features (MFCCs) suggest a potential area of research [21].

On the other hand, while many speaker recognition researchers have been motivated to investigate features derived from different sources of information in speech (e.g. frequency, phase, modulation energy), little has been done to determine the relative contributions of the acoustic and speaker modelling ‘stages’ (by the UBM and MAP adaptation blocks as shown in Figure 1.1 respectively) and the benefits brought about by fusing different acoustic features. Knowledge of the relative contribution of each stage on the entire speaker recognition system is useful, as it provides a framework to understand and discuss the classification performance of various features.

In regards to the complementary properties of classification techniques, recent developments in the theory of compressive sensing [27] has evolved into a new classification method, Sparse Representation Classification (SRC). Sparse representation classification has shown interesting results in face recognition [28], speaker identification [29] and various other applications [30-32]. Furthermore, most experimental results to date indicate that SRC can achieve a generalisation performance that is greater than or equal to that of other classifiers, in particularly the SVM [28, 29, 31-34], and has the benefit of not requiring a training phase, resulting in a much simpler training process and also the opening of a new paradigm. Despite the mentioned benefits and its application in speaker identification [29], SRC hasn't been employed in the context of speaker verification making it another potential area of research.

1.2 Thesis Objectives

Given the limitations identified in the previous section, the principal objective of this thesis is to investigate alternative features to MFCCs and speaker modelling/classification. This broad objective may be expressed in terms of a number of aims:

- To investigate and develop novel features as a complementary front-end to conventional magnitude-based (i.e. MFCCs) systems. In particular, features that captures information about the phase or dominant frequencies in the speech signal, which are not used explicity in standard feature extraction methods.
- To *separately* investigate the acoustic and speaker modelling 'stages' of the GMM-UBM based systems, towards determining the contributions of each

stage (acoustic and speaker modelling stage) to the speaker recognition performance across different features.

- To analyse the fusion of multiple classification systems and its performance against a single classifier approach.
- To investigate the use of sparse representation classification in the context of speaker verification systems.

1.3 Organization of the Thesis

The remainder of the thesis is organised as follows:

Chapter 2 provides an overview of the speaker specific aspects of speech production systems, feature extraction, speaker modelling techniques, decision making and evaluation measures that are applied to automatic speaker recognition systems. It also identifies some of the problems to be addressed in current automatic speaker recognition systems.

Chapter 3 proposes methods to address the difficulties identified in Chapter 2 for group delay and frequency modulation features when applying them in speech based classification problems. These difficulties include the presence of artifacts due to ill-conditioning and computationally inefficient feature extraction respectively. Furthermore, it will show the complementary characteristics of group delay and spectral centroid features when compared to a magnitude-based system.

Chapter 4 investigates the relative contributions of the acoustic and speaker modelling ‘stages’ and the benefits brought about by fusing different acoustic features using clustering comparison measures.

Chapter 5 starts with the evaluation of the GMM-SRC on a speaker verification task. Then it introduces the use of speaker factors from the JFA approach as an alternative to GMM supervectors due to its excellent discriminative capability and small dimensionality.

Chapter 6 concludes the thesis with a summary of the contributions of this research and presents potential future work to follow up from this thesis.

1.4 Major Contributions

This research described in this thesis provides original contributions to the field of automatic speaker verification system. The major contributions can be summarised as follows:

Feature Extraction: Proposed log compressed least squares group delay (LogLSGD) features

An alternative group delay (GD) feature extraction method is proposed in order to reduce the dynamic range and variability of the modified group delay (MODGDF) features [24] using least squares regularisation. This method of extraction is simpler when compared with MODGD because the LogLSGD feature extraction algorithm does not depend on any empirical parameters and it is data independent.

Feature Extraction: Proposed spectral centroid features

The characterisation of subband energy (i.e. MFCCs) as a two dimensional feature, comprising spectral centroid magnitude (SCM) and spectral centroid frequency (SCF) is proposed. Compared with conventional MFCCs, the proposed combination of SCM and SCF produces better recognition performance and both features fuse effectively with MFCCs. Furthermore, SCF is also shown to perform significantly better than the previously proposed subband spectral centroid [35] and frame-averaged FM features [36] for speaker recognition.

Front-end Diversity: Fused subsystems based on different MFCC-variant features

Research reported in machine learning [37] literature demonstrated that classifier ensembles are well established as a method for obtaining highly accurate classification by combining (fusing) less accurate individual classifiers. In an attempt to utilise this concept for speaker recognition, the work in this thesis experimented with some possible variations to the computation of the de facto standard feature for speaker recognition, MFCCs, and showed that the fusion of suboptimal systems based on features comprising essentially of the same information as that contained in MFCCs outperforms an individual MFCC based system, and given appropriate design choices, this improvement can be significant.

Front-end Diversity: Proposed clustering comparison measures and UBM data selection

The clustering comparison measures are utilised to investigate the acoustic and speaker modelling aspects of the speaker recognition task separately and demonstrate that front-end diversity can be achieved purely through different 'partitioning' of the acoustic space. Furthermore, features that exhibit good 'stability' with respect to repeated clustering are shown to give good equal-error-rate (EER) performance in speaker recognition. Then, a novel utterance selection algorithm for training a compact "stable" UBM is presented and evaluated on the NIST 2006 database. Results show that using Normalised Information Distance (NID)-based resampling to select utterances during UBM training can improve speaker recognition performance despite employing a smaller set of training data.

Classification: Proposed sparse representation classifier for speaker verification

The discriminative nature of a sparse representation classifier (SRC) for a speaker verification task is investigated. Experimental results demonstrated that supervector-based SRC classifiers (GMM-SRC) are able to achieve comparable performance to the current state-of-the-art modelling/classification techniques (UBM-GMM and GMM-SVM). Building on the concept of GMM-SRC and joint factor analysis-support vector machines (JFA-SVM), a speaker factor-based SRC as an alternative classifier to GMM-SVM, producing an approach this thesis terms JFA-SRC, was discussed. Furthermore, motivated by background speaker selection for the SVM-based system [38], a novel SRC background dataset selection based on column vector frequency is presented, allowing a faster verification process with a smaller dictionary.

1.5 List of Publications

Journal publication

- **Kua, J. M. K.**, Epps, J. and Ambikairajah, E., "Joint factor analysis with sparse representation classification for speaker verification", *Speech Communication* (Submitted January 2012).

Conference publications

- **Kua, J. M. K.**, Tharmarajah, T. and Ambikairajah, E., "A Non-Uniform Filterbank for Speaker Recognition" to be published in *Proceedings of Annual Conference of the International Speech Communication Association, INTERSPEECH 2012.*
- **Kua, J. M. K.**, Ambikairajah, E., Epps, J. and Togneri, R., "Speaker verification using sparse representation classification", in *Proceedings of IEEE International*

Conference on Acoustics, Speech and Signal Processing, ICASSP 2011, pp. 4548-4551.

- **Kua, J. M. K.**, Epps, J., Nosratighods, M., Ambikairajah, E. and Choi, E. H. C., “Using clustering comparison measures for speaker recognition”, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011*, pp. 5452-5455.
- **Kua, J. M. K.**, Epps, J., Ambikairajah, E. and Nosratighods, M., "Front-end Diversity in Fused Speaker Recognition Systems," in *Proceedings of Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2010*, pp. 59-63.
- **Kua, J. M. K.**, Tharmarajah , T., Nosratighods, M., Ambikairajah, E. and Epps, J., "Investigation of spectral centroid magnitude and frequency for speaker recognition," in *Proceedings of Odyssey Speaker and Language Recognition Workshop 2010*, pp. 34-39.
- **Kua, J. M. K.**, Epps, J., Ambikairajah, E. and Choi, E. H. C., "LS regularization of group delay features for speaker recognition," in *Proceedings of Annual Conference of the International Speech Communication Association, INTERSPEECH 2009*, pp. 2887–2890.

Chapter 2

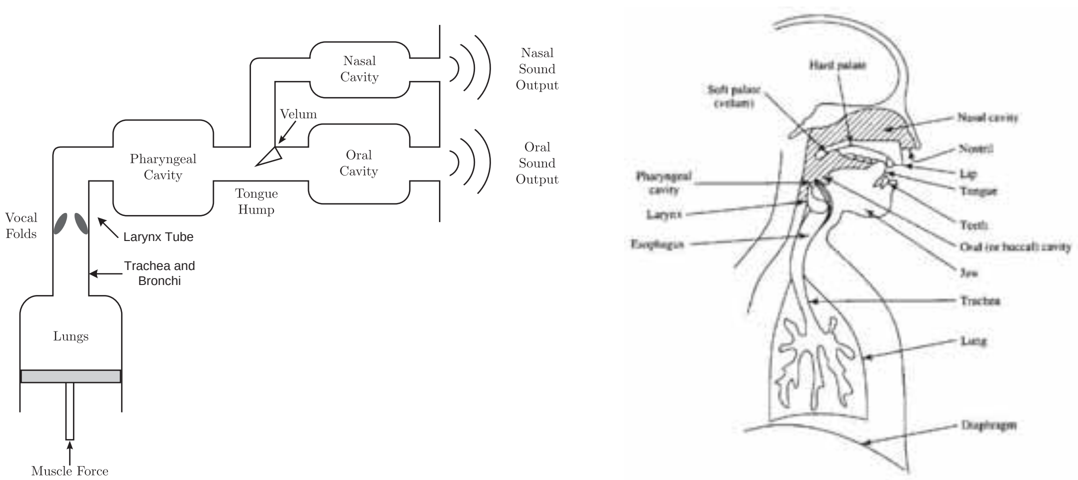
An Overview of Speaker Recognition Systems

This chapter starts with a discussion on the mechanisms involved in human speech production and how some of the speaker-specific characteristics are identified within this model. It then provides a brief background to automatic speaker recognition systems. This includes a description of the main components that make up a speaker verification system. Specifically, it elaborates on the different types of features, modelling and classification methods, normalisation techniques and performance measures in state-of-the-art speaker verification systems. Furthermore, the databases used to evaluate the techniques described in this thesis are presented, with focus on the National Insitute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) corpora.

2.1 Human Speech Production System

Spoken language is the most natural mode used by humans to communicate information. The speech signal conveys not only what is being said but also embodies individual characteristics of the speaker. Speaker specific characteristics are governed by both the physiological and behavioural characteristics of the speaker. Prominent physiological features are the shape and length of the vocal tract and how the physical speech apparatus is manipulated to produce speech.

To identify speaker-specific characteristics in speech, for later use in performing automated speaker verification, it is essential to recognise the process by which speech is generated and the acoustic effect of inter-speaker variability. An understanding of the nature of speech production will assist in determining more effective techniques for isolating speaker specific properties. Figure 2.1(a) shows a schematic representation of the human speech production system, and Figure 2.1(b) shows the human speech production organs. From a physiological perspective, speech is driven by an excitation source - air flow from the lungs through the trachea and vocal folds. For voiced speech, the vibrating vocal folds modulate the air flowing from the lungs into quasi-periodic pulses. The fundamental frequency[1] of these pulses corresponds to the rate at which the vocal folds are vibrating. Alternatively for unvoiced speech, the vocal folds do not vibrate and air flows through a narrow opening, typically created by the position of the vocal folds, tongue, and/or lips, resulting in turbulent airflow with noise-like characteristics.



(a) Schematic diagram

(b) Speech production mechanism

Figure 2.1 Human speech production system [39]

The pulsed or turbulent air stream, which corresponds to the source of excitation for voiced and unvoiced speech respectively, then excites the vocal tract causing it to

[1] The fundamental frequency is often used interchangeably with pitch, even though pitch is technically the

resonate at its characteristic frequencies (formant frequencies). The vocal tract starts at the opening of the vocal folds and ends at the lips and it comprises three main cavities, the pharyngeal, oral and nasal cavities, and a velum that controls the amount of airflow to the nasal cavity (as shown in Figure 2.1). The formant frequencies are determined by the shape of the vocal tract, which in turn is determined by the positions of articulators, such as tongue, lips, jaw and velum. This allows humans to control the resonance characteristics and consequently the speech sound being produced.

2.1.1 Speaker-Specific Properties in Speech

Generally, recognising speakers from speech relies on how the speech is influenced by the organs of the human speech production system, which is thought to have two components: individual characteristics of the laryngeal source, and those of the supralaryngeal vocal tract as discussed above. While the physical properties of the vocal folds are mainly determined by the speaker's age and gender, variations in vocal tract shape produce strong differences in the spectrum of speech that distinguish one speaker from another [40].

Honda [41] showed a correlation between geometrical measures of the vocal tract and the lower formant frequencies, and Fitch et al. [42] reported a high positive correlation between vocal tract length and body size. Because vocal tract length influences formant frequencies, variation in speakers' body size can give rise to speaker specific characteristics in speech. In addition, it was shown in [40] that the variations of the third and fourth formants (F_3 and F_4) contain a significant amount of speaker-specific information (localized in the speech frequency spectrum around 2.5kHz as they were stable over vowel phonation). Investigation in [43] also revealed that the piriform fossa (two small pockets behind the larynx) as shown in Figure 2.2 causes troughs in the transfer function of the vocal tract around the frequency region between 4kHz and 5kHz, in addition to some global effect on the lower formants. From these analyses, it can be observed that speaker-specific information is distributed non-uniformly in different frequency bands of speech from the speech production point of view.

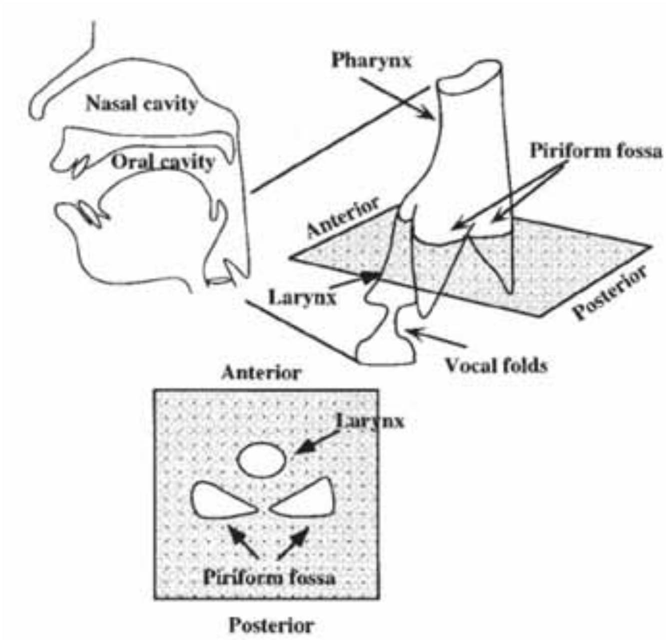


Figure 2.2 Diagram of the vocal tract in the vicinity of the piriform fossa [43]

2.2 Automatic Speaker Verification Systems

Automatic speaker verification is the process of determining to a specified level of confidence if a person is who he or she claims to be through the analysis of their speech. A speaker verification system can typically be broken down into three components as shown in Figure 2.3. The first one is dedicated to feature extraction, where raw speech is processed to obtain a set of speaker-discriminate features representing the characteristics of the speaker (section 2.3). The second component is the modelling module, where a speaker model is trained using the extracted features (section 2.4). The final component that makes up a speaker verification system is the scoring and decision making process (section 2.4). In the automatic speaker verification literature, the feature extraction stage is commonly known as the front-end and the modelling and score computation stages together are commonly referred to as the back-end as shown in Figure 2.3.

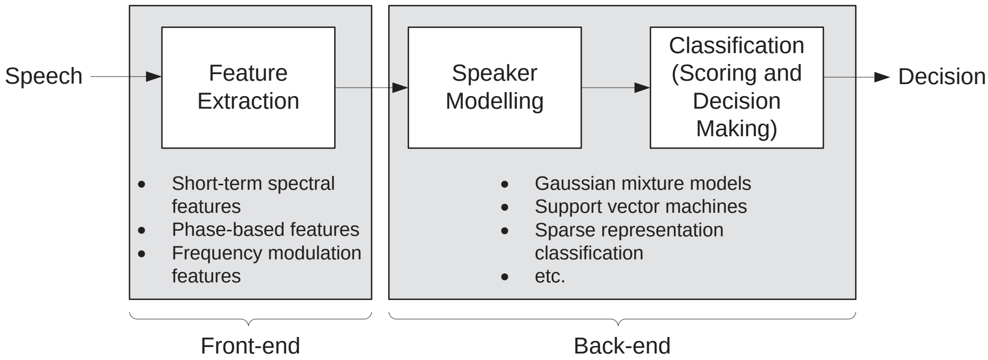


Figure 2.3 Overview of automatic speaker verification system

2.3 Feature Extraction

As stated above, feature extraction is the process of transforming the raw signal into some type of parametric representation of lower information rate, termed feature vectors, in which speaker-specific properties are emphasized and statistical redundancies suppressed. In speaker recognition, features can be broadly categorized into a hierarchy running from low-level information, such as the sound of a person's voice (related to physical traits of the vocal apparatus as discussed in section 2.1), to high-level information, such as particular word usage or idiolect (related to learned habits and style) as shown in Figure 2.4 [44]. While all of these levels convey useful speaker information, automatic speaker recognition systems have relied almost exclusively on short-term, low-level acoustic information, such as cepstral features because of their ability to extract speaker discriminative information whilst also retaining information regarding the linguistic content of the speech utterance [45, 46]. Although recent developments have investigated the potential benefits of high-level characteristics of speech[2] [47] on speaker verification

[2] Readers interested in the exploitation on high-level information for speaker recognition may refer to the SuperSID workshop (http://www.clsp.jhu.edu/ws2002/groups/supersid/) for further details.

task, unfortunately, in contemporary speaker verification applications, insufficient training data is available to model all of these levels of information. Hence, and in keeping with the trends in most state-of-the-art speaker recognition systems, the features reported in this thesis focus on capturing the low-level information via short-term spectral features.

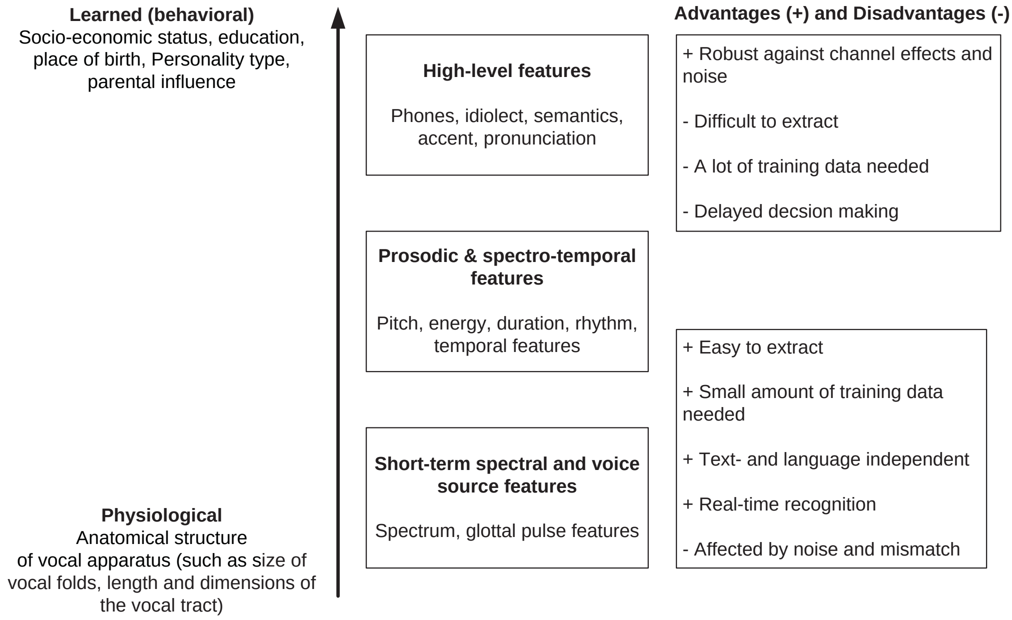


Figure 2.4 A summary of features from viewpoint of their physical interpretation. The choice of features has to be based on their discrimination, robustness, and practicality. Short-term spectral features are the simplest, yet most discriminative; prosodics and high-level features have received much attention at high computational cost. [4]

2.3.1 Short-term Spectral Features

To date, the short-term spectral features that are most commonly used in speaker verification are the mel frequency cepstral coefficients (MFCCs) [48], linear prediction cepstral coefficients (LPCCs) [49] and perceptual linear prediction coefficients (PLPs) [50]. In order to extract any of those three features, the continuously changing speech signal is first segmented into short frames of about 20-30 ms duration based on the assumption that the speech waveform is approximately stationary over short intervals. Then the frame of speech is pre-emphasised and multiplied by a window function prior to further processing. Pre-emphasis is carried out to enhance the high frequencies of the spectrum (or to balance the spectrum of voiced sounds that have a steeper roll-off in the high frequency region) [51] and window functions (usually Hamming or von Hann windows) are typically used to taper the original signal near the frame edges and thus reduce the edge effects (aliasing).

Mel frequency cepstral coefficients

To compute the MFCCs, the fast Fourier transform (FFT), a fast implementation of discrete Fourier Transform (DFT), is used to decompose each frame of speech into its frequency components [48]. The magnitude spectrum of each frame is then computed with the phase spectrum discarded based on the assumption that phase has little perceptual importance. Thereafter the magnitude spectrum is multiplied with a series of triangular bandpass filters (termed critical band filter bank analysis), which are equally spaced on a mel-frequency scale (equation 2.1) to approximate the behaviour of the human auditory system, giving more detail to the low frequencies when viewed in the linear scale (Hz). Finally, logarithmic compression of the filterbank energies and discrete cosine transform (DCT) are performed for reducing the correlation between pairs of feature components. An overview of MFCC computation is shown in Figure 2.5.

$$f_{MEL} = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{2.1}$$

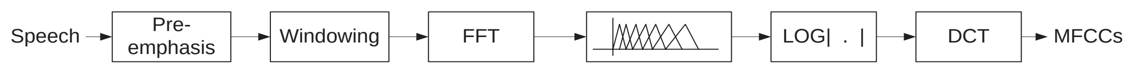


Figure 2.5 Overview of MFCC feature extraction

Linear Prediction Coefficients

The computation of linear prediction coefficients (LPC) are based on linear predictive analysis which attempts to describe a speech signal $\tilde{s}[n]$ at time n as a linear combination of P past signal values as follows

$$\tilde{s}[n] = \sum_{k=1}^{P} a_k s[n-k] \tag{2.2}$$

where a_k are the linear prediction coefficients. These coefficients a_k are determined by minimising the mean-squared prediction error between the speech sample, $s[n]$, and its linearly predicted value, $\tilde{s}[n]$ using the Levinson-Durbin algorithm [39, 51]. While the coefficients produced by the LP model form the fundamental feature set used in speaker recognition systems, they are generally transformed into a more suitable representation (i.e cepstral coefficients) for the purpose of speaker modelling and classification. Herein LPCCs are calculated as a Fourier or cosine transform from the log-magnitude spectrum that is estimated through the frequency response of the all-pole filter defined by the prediction coefficients [49]. In contrast to LPCCs, the extraction of PLPs [50] exploits selected psychoacoustic principles such as critical band analysis (Bark), equal loudness pre-emphasis and intensity-loudness relationship, before being fitted with an all-pole model and converted to cepstral coefficients as shown in Figure 2.6 [50].

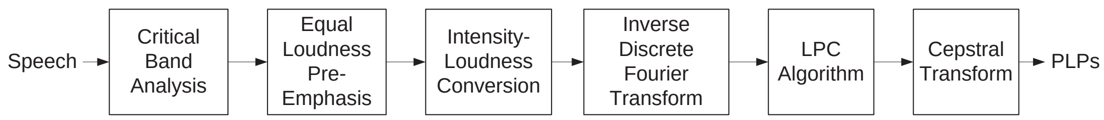


Figure 2.6 Overview of PLP feature extraction

2.3.2 Phase-based Features

In addition to the above mentioned features that are mainly based on the (FFT or LP) magnitude spectrum of the speech, alternative features based on the phase spectrum termed phase-based features have grown in interest. Previously phase information has normally been discarded because of the complexity of extracting features from the phase spectrum and earlier psychoacoustic experiments/human perception studies by Helmholtz [52] and Liu et al. [53] indicating that the human ear is almost insensitive to phase and/or that the short-time phase spectrum conveys no information about the intelligibility of speech for small window durations (20-40 ms) [52]. However, recent human perception experiments conducted by Paliwal and Alsteris [54] provides opposing evidence. The experiments are built on the basis of Liu's experimental procedures with a number of modifications. Their results indicate that the short-time phase spectrum (with window size of 32 ms) contributes to speech intelligibility as much as the corresponding power spectrum if the shape of the window function is properly selected [54, 55] and this is further supported with more listening tests in [56]. Furthermore Hedge et al. [24] and Thiruvaran et al. [25, 26, 36] have shown success in utilising phase information relating to the frequency-domain signal using group delay-based features and in the time-domain captured by using frequency modulation (FM) features for speaker recognition.

2.3.2.1 Group Delay Features

The group delay, $G(f)$, is the negative frequency derivative of the spectral phase (phase of the spectrum after Fourier transform), $\phi(f)$, as follows:

$$G(f) = -\frac{1}{2\pi}\frac{d\phi(f)}{df} \tag{2.3}$$

It has been shown in [57] that the group delay captures the formants information (which are related to the recognition of speakers as discussed in section 2.1.1). Specifically, the resonances (formants) of the speech signal which correspond to the peaks of the envelope in the short-time magnitude spectrum will appear as transitions in the short-time phase spectrum.

One of the problems with extracting the group delay according to equation (2.3) is that the phase spectrum of a signal is wrapped within $\pm\pi$ (shown in Figure 2.7 (a)), so unwrapping, which is a non-unique process, is required. Although phase unwrapping techniques have been proposed [58, 59], it is generally preferred to avoid unwrapping since in general phase unwrapping is a heuristic approach [60]. In order to circumvent unwrapping, the group delay defined in equation (2.3) can be estimated using the real and imaginary part of the Fourier transform spectrum, $S(f)$, as follows [61, 62].

$$G(f) = \frac{S_R(f)F_R\{ts(t)\} + S_I(f)F_I\{ts(t)\}}{|S(f)|^2} \tag{2.4}$$

where $s(t)$ denotes the speech signal in the time-domain (t), the subscript R and I denote the real and imaginary parts respectively and $F\{.\}$ denotes the Fourier transform. However the group delay is not an ideal feature in its original form (computed based on equation (2.4)) as it suffers from extreme estimates as shown in Figure 2.7 (b). The peaks are caused by zeros of the z-transform of the excitation components (being close to the unit circle) of the speech signal as illustrated in [63]. In an attempt to suppress the peaks, various modifications have been proposed, including replacement of the power spectrum with cepstrally smoothed power spectrum [63], inclusion of two empirical parameters [64], low pass filtering [61], and log compression [25] in the group delay calculation, which will be discussed next.

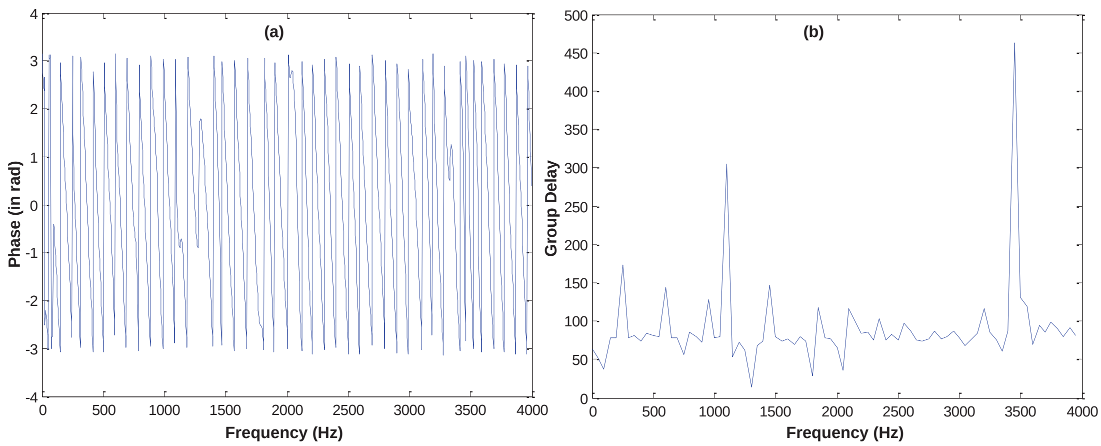


Figure 2.7 (a) Phase spectrum (wrapped within $\pm\pi$) (b) Group delay function (with high-amplitude peaks) for a frame of voiced speech taken from the NIST 2001 database

Cepstral Smoothing

As mentioned above, in order to extract a meaningful representation from the group delay function, it is necessary for the zeros of the transfer function of speech to be away from the unit circle in the *z* plane. In the context of speech, the poles are normally well within the unit circle, while the zeros could either be within or outside the unit circle. Moreover it is important to note that the denominator term $|S(f)|^2$ in equation (2.4) becomes zero at zeros of the excitation component that are located close to the unit circle [65]. Yegnanarayana and Murthy [65] proposed the multiplication of the group delay function by the source component of the power spectrum. This operation gives less weight to peaks in the group delay function, which are the result of excitation-induced zeros near the unit circle. This is equivalent to replacing the denominator in equation (2.4) with the system component of the power spectrum, $|\mathbb{S}(f)|^2$:

$$G_s(f) = \frac{S_R(f)F_R\{ts(t)\} + S_I(f)F_I\{ts(t)\}}{|\mathbb{S}(f)|^2} \tag{2.5}$$

where $|\mathbb{S}(f)|^2$ is the cepstrally smoothed spectrum of $|S(f)|$ [66]. In practice, the cepstral smoothing operation not only smoothes out zeros introduced by excitation, but also those contributed by noise and windowing [67, 68].

Modified Group Delay

In order to further suppress the peaks and to restore the dynamic range of the speech spectrum, two heuristic compression parameters, γ and $\beta \in [0,1]$ were introduced [64]. The resulting group delay function is termed the modified group delay (MODGD) as follows:

$$G_{mod}(f) = \left|\frac{S_R(f)F_R\{ts(t)\} + S_I(f)F_I\{ts(t)\}}{|\mathbb{S}(f)|^{2\gamma}}\right|^{\beta} \cdot sgn\big(G_s(f)\big) \tag{2.6}$$

where *sgn* is the signum function and $G_s(f)$ is as given by equation (2.5).

Log Compression

Recently, the log compression of group delay function was proposed for speaker verification systems [25], as shown in equation (2.7). As the peaks are mainly caused by the excitation source, which has speaker-specific information in addition to the effect of the vocal tract on the group delay, the authors believe that the peaks should be suppressed as opposed to eliminated. Although this approach is not applicable if the group delay has both positive and negative values, informal experiments in [25] revealed that the occurrence of sign change was less than 2%. Further, negative values for group delay are difficult to interpret, so preserving the sign does not seem to provide benefit. Thus, the absolute values were taken before log compression.

$$G_{log}(f) = \log\left\{\frac{|S_R(f)F_R\{ts(t)\} + S_I(f)F_I\{ts(t)\}|}{|\mathbb{S}(f)|^2}\right\} \quad (2.7)$$

Parameterising Group Delay Function

Finally, common to all group delay feature extraction algorithms in this thesis, the discrete cosine transform (DCT) [69] is used to convert the group delay spectra to cepstral features [62]. This is primarily to yield feature components that are decorrelated, which allows the use of diagonal covariance matrices in modelling the speech vector distribution (discussed in section 2.4.1) [24, 64].

2.3.2.2 Frequency Modulation Features

As mentioned previously, in addition to group delay, another phase-based feature that has shown promise for speaker recognition in recent years is the frequency modulation (FM) feature [4]. In particular, the frequency modulation feature is motivated by an AM-FM model [70-72] of the speech signal, in which the speech signal $s[n]$ is modelled as the sum of K AM-FM signals, one for each resonance represented in discrete form as follows:

$$s[n] = \sum_{k=1}^{K} r_k[n] \tag{2.8}$$

where each speech resonance, $r_k[n]$, is modelled as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) as follows:

$$r[n] = A[n]\, cos \left[\frac{2\pi f_c n}{f_s} + \frac{2\pi}{f_s} \sum_{r=1}^{n} q[r] \right] \tag{2.9}$$

where $A[n]$ is the time-varying amplitude component, $q[r]$ is the time-varying frequency component and f_c is the “center value” of the resonant frequency (formant frequency) and f_s is the sampling frequency.

The AM-FM modulation model is occasionally preferred over other models, such as the traditional source-filter model [73], because it describes the nonlinear and time-varying phenomena during speech production. In the source-filter model, the sound source is assumed to localised in the larynx, while the vocal tract acts as a convolution filter for the emitted sound (as discussed in section 2.1). Although this approach has led to great advances, it is known to neglect some structure present in the speech signal. Examples of phenomena not well-captured by the source–filter model include unstable airflow, turbulence, and nonlinearities arising from oscillators with time-varying masses [70, 74]. This is further supported in [75] indicating that a significant part of the acoustic information cannot be modelled by the linear source-filter acoustic model, and thus, the need for nonlinear features becomes apparent.

Recently, Thiruvaran has shown the complementary behaviour of FM and MFCC for speaker recognition, with detailed comparative experiments across different FM feature extraction methods [76]. The all-pole FM [26] outperformed other existing techniques such as discrete energy separation algorithm (DESA) [70], smoothed energy operator separation algorithm (SEOSA or SESA) [77], Hilbert transform based method [78] and frequency amplitude modulation encoding (FAME) [79, 80], in the context of speaker recognition.

All-pole Frequency Modulation Features

In the all-pole method, FM features are extracted from subbands of speech decomposed using a fixed filter bank as proposed in [70] for isolating individual resonances. The k^{th} bandpass filter output $p_k[n]$ can be represented according to the AM-FM model (shown in equation (2.9)) as

$$p_k[n] = a_k[n] \cos\left[\frac{2\pi f_{ck} n}{f_s} + \frac{2\pi}{f_s}\sum_{r=1}^{n} q_k[r]\right] \tag{2.10}$$

where f_{ck} is the center frequency of the k^{th} bandpass filter and $q_k[n]$ is the FM component. The FM component $q_k[n]$ is estimated by modelling the subband signals using second order all-pole resonators

$$q_k[n] = \theta_k \frac{f_s}{2\pi} - f_{ck} \tag{2.11}$$

where θ_k is the pole angle of the resonator. The filter coefficients of the resonator can be obtained from second order linear prediction analysis. Figure 2.8 summarises the all-pole FM feature extraction. The Gabor filter has usually been used as the filter bank for sub-band FM extraction because it gives optimally compact sensitivity in the time and frequency domain and its shape does not produce any large side lobes that would introduce spectral leakage [70].

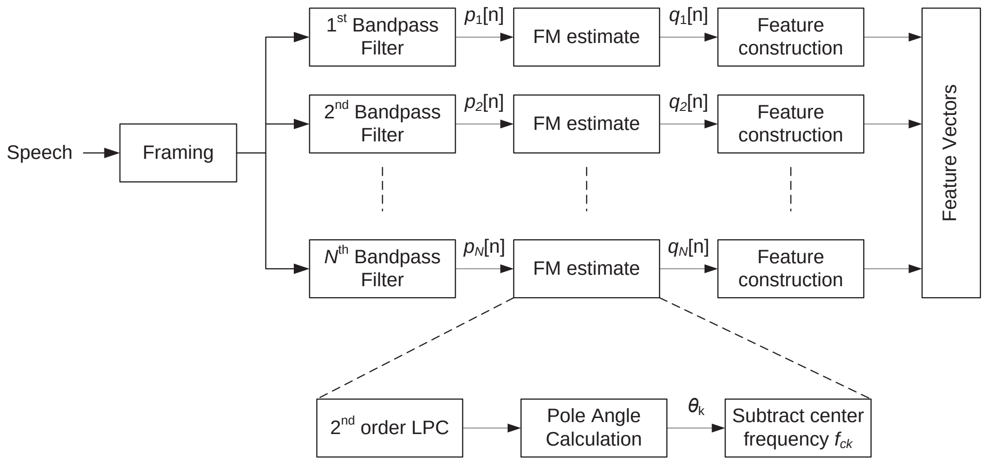


Figure 2.8 Schematic of all-pole FM extraction [76]

Subband Spectral Centroid Features

Furthermore, the dependency of all-pole FM on the resonance frequency motivated Thiruvaran to conduct a comparison (in [76]) with the subband spectral centroid (SSC) feature [35] which has shown success for speech recognition in noisy environment. Although similar time trajectories of FM and SSC features were observed (particularly in the lower frequency bands), suggesting both features carry similar information, SSCs were unable to achieve comparable speaker recognition performance to the all-pole FM.

Subband spectral centroid is the weighted average frequency for a given subband, where the weights are the normalised energy of each frequency component in that subband. Since this measure captures the ‘centre of gravity’ of each subband, it can detect the approximate location of formants if the bandwidth is wide enough; otherwise it finds harmonics [35, 81]. The m^{th} subband spectral centroid F_m is defined as follows [35]:

$$F_m = \frac{\sum_{k=l_m}^{u_m} k|S[k]w_m[k]|}{\sum_{k=l_m}^{u_m} |S[k]w_m[k]|} \tag{2.12}$$

where $|S[k]|$ represents the magnitude spectrum of a frame of speech. $|S[k]|$ is then divided into M sub-bands, $w_m[k]$, where each sub-band is defined by a lower frequency edge (l_m) and an upper frequency edge (u_m).

2.3.3 Open Questions on Feature Extraction

Given the wide variety of features in literature, which one should be used for speaker recognition and why are some of the common questions in the literature. Some comparisons can be found in [46, 82-84] addressing the former question and it has been observed in the biennial NIST speaker recognition evaluation (SRE)[3] that MFCC have been the dominant feature in most speaker recognition systems. Given this, the latter question of 'why this feature (MFCCs) should be used?' is yet to be explored in detail. Furthermore, recently Lu et al. [85] showed that the standard way of extracting MFCC features using an auditory (mel) scale is not an optimal scale for designing a speaker identification scale, as it does not take into account the distribution of speaker discriminative information across the spectrum (discussed in section 2.1) since more importance is given to the lower frequency region, gradually reducing the importance to higher frequency area. Consequently, all these unknowns raise doubts as to what makes a good feature for speaker recognition.

2.4 Speaker Modelling and Classification

Given a suitable set of speaker representative features, the next question would be how to effectively organise and exploit the speaker cues in the classifier design for the best performance. Addressing this issue, some of the conventional methods include Gaussian mixture model-universal background models (GMM-UBM) [8, 86] and support vector machines (SVM) [20, 87]. Recently, a new combination of GMM-UBM and SVM termed as Gaussian mixture model-support vector machines (GMM-SVM) [9] has been developed. This is a hybrid classifier where the GMM-UBM model is used for creating

[3] http://www.itl.nist.gov/iad/mig/tests/sre/

"feature vectors" for the SVM. Alternatively to GMM-SVM, one other promising classifier, the sparse representation classification computed by ℓ_1-minimisation (to approximate the ℓ_0-minimisation) has recently demonstrated its effectiveness in the close set speaker identification task [29]. In this section, we will discuss the main modelling techniques used in this thesis.

2.4.1 Gaussian Mixture Models

Gaussian mixture models are currently the de facto reference method for speaker recognition system because of their ability to provide a smooth approximation to any arbitrary probability distribution function (PDF) using a compact set of parameters [88]. For a *D*-dimensional feature vector, $\mathbf{x}$, the GMM mixture density, $p(\mathbf{x}|\boldsymbol{\lambda})$, is a weighted linear combination of *M* uni-modal Gaussian densities, $p_i(\mathbf{x})$, as follows:

$$p(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} w_i p_i(\mathbf{x}) \tag{2.13}$$

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathrm{i})^\mathrm{T}(\boldsymbol{\Sigma}_i)^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)} \tag{2.14}$$

where w_i is the weight of the i^{th} mixture component satisfying $\sum_{i=1}^{M} w_i = 1$, D is the feature dimension, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\mu}_i$ are the covariance matrix and mean vector of the i^{th} mixture component respectively and $\boldsymbol{\lambda} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ is the set of GMM parameters. Figure 2.9 illustrates how a single dimensional (*D* = 1) probability density function is approximated by a 3-component Gaussian mixture model (*M* = 3).

While the general model allows for full covariance matrices, diagonal covariance matrices are usually used (as well as in this work) since the parameter estimation of a full-covariance GMM in general requires more training data and is computationally expensive. In addition, empirical results have shown diagonal matrices outperforming full matrices for practical speaker recognition systems [8].

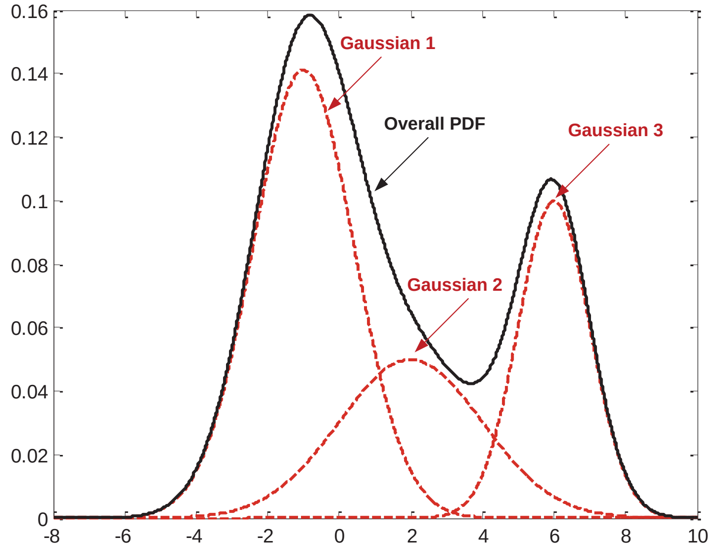


Figure 2.9 Probability density function approximated by a 3-component Gaussian mixture model

Training Speaker Models: Maximum A Posteriori Adaptation

For speaker recognition, each speaker is represented by a GMM model and will be referred to by its model parameters $\boldsymbol{\lambda}$ throughout this thesis, where the parameters of the GMM model are trained using Expectation Maximization (EM) algorithm, which iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors[4] [89]. However in most speaker verification systems, we do not have enough data to train speaker-specific GMMs using the EM algorithm. To overcome these difficulties, a speaker verification system based on Maximum A Posteriori (MAP) adaptation from the Universal Background Model (UBM), termed the Gaussian Mixture Model − Universal Background Model (GMM-UBM), was introduced in [8], under the assumption that the UBM will adequately describe the underlying characteristics of a large speaker population. Generally, the UBM is trained on a large and diverse set of speakers, and their identities are different from the target speaker. The speaker GMM model is then derived from the UBM by MAP adaptation

[4] In general, only local maximum likelihood estimates of the model parameters can be found, since the EM algorithm assures convergence to a local, rather than global optimum.

using the target speaker data as follows [8, 90]: Given a UBM and training data from a hypothesised speaker, $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ where $\mathbf{x}_t$ is the feature vector at time t, the probabilistic alignment of the training vectors for each i^{th} UBM mixture components is first determined as shown in equation (2.15).

$$\mathrm{P}(i|\mathbf{x}_t) = \frac{w_i p_i(\mathbf{x}_t)}{\sum_{j=1}^{M} w_j p_j(\mathbf{x}_t)} \tag{2.15}$$

Next, the new sufficient statistics for the weight, mean, and variance parameters are computed as follows:

$$n_i = \sum_{t=1}^{T} \mathrm{P}(i|\mathbf{x}_t) \tag{2.16}$$

$$E_i(\mathbf{x}) = \frac{1}{n_i} \sum_{t=1}^{T} \mathrm{P}(i|\mathbf{x}_t)\, \mathbf{x}_t \tag{2.17}$$

$$E_i(\mathbf{x}^2) = \frac{1}{n_i} \sum_{t=1}^{T} \mathrm{P}(i|\boldsymbol{x}_t)\, \mathbf{x}_t^2 \tag{2.18}$$

The adapted parameters for the i^{th} mixture component are created by updating the old UBM parameters using the new sufficient statistics:

$$\widehat{w}_i = \left[\alpha_i \frac{n_i}{T} + (1 - \alpha_i) w_i\right] \gamma \tag{2.19}$$

$$\widehat{\boldsymbol{\mu}}_i = \alpha_i E_i(\mathbf{x}) + (1 - \alpha_i) \boldsymbol{\mu}_i \tag{2.20}$$

$$\widehat{\boldsymbol{\sigma}}_i^2 = \alpha_i E_i(\mathbf{x}^2) + (1 - \alpha_i)(\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2) - \widehat{\boldsymbol{\mu}}_i^2 \tag{2.21}$$

where γ is a scale factor to ensure the adapted weights sum to unity and α_i are the per-mixture component adaptation coefficients controlling the balance between the old and new estimates for the weights, means and variances with a fixed data-dependent relevance factor, r, (typically set between 8 and 32) as follows:

$$\alpha_i = \frac{n_i}{n_i + r} \tag{2.22}$$

Unless otherwise stated, a relevance factor of $r = 10$ is used throughout this thesis. The basic idea behind adaptation is to utilise a well-trained model (UBM) as a basis to obtain better GMM models of different speakers [8]. The per-mixture component adaptation coefficient is designed in such a way that the model parameters are updated only when the training data from the hypothesised speaker are reliable (high counts of data available for mixture component *i*); otherwise, the parameters from the UBM are used instead. In practice, only mean vectors $\boldsymbol{\mu}_i$ are adapted since it has been shown that updating the weights and covariance matrices does not significantly impact system performance [8].

Log-Likelihood Ratio Scoring

Finally, the task of speaker verification is to ascertain whether a test set of speech frames $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ belongs to the claimed speaker. In the GMM paradigm, the aim is to test the following hypotheses:

- H_{tar}: **X** is uttered by claimed speaker.
- H_{non}: **X** is not uttered by claimed speaker.

The decision score is based on a likelihood ratio as follows:

$$\Lambda(\mathbf{X}) = \frac{\mathrm{P}(\mathbf{X}|H_{tar})}{\mathrm{P}(\boldsymbol{X}|H_{non})} \qquad \begin{Bmatrix} \geq \ \vartheta \ \Rightarrow H_{tar} \\ < \ \vartheta \ \Rightarrow H_{non} \end{Bmatrix} \tag{2.23}$$

where $\mathrm{P}(\mathbf{X}|H_{tar})$ and $P(\boldsymbol{X}|H_{non})$ are the likelihood of **X** being uttered and not being uttered by the claimed speaker respectively. In a speaker verification task, the likelihood ratio is compared with a threshold, ϑ, to accept/reject the claimed identity of a speaker. Often the log-likelihoods (log of the likelihood) are used in place of likelihood values to improve numerical precision as the likelihood values tend to be very small. In addition, H_{tar} and H_{non} usually refer to the speaker model $\boldsymbol{\lambda}_{spk}$ and the UBM model $\boldsymbol{\lambda}_{UBM}$

respectively which allows equation (2.23) to be expressed as equation (2.24). Figure 2.10 summarises all components of the GMM-UBM system.

$$\log\big(\Lambda(\boldsymbol{X})\big) = \log\Big(\mathrm{P}\big(\mathbf{X}|\boldsymbol{\lambda}_{spk}\big)\Big) - \log\big(\mathrm{P}(\boldsymbol{X}|\boldsymbol{\lambda}_{UBM})\big) \tag{2.24}$$

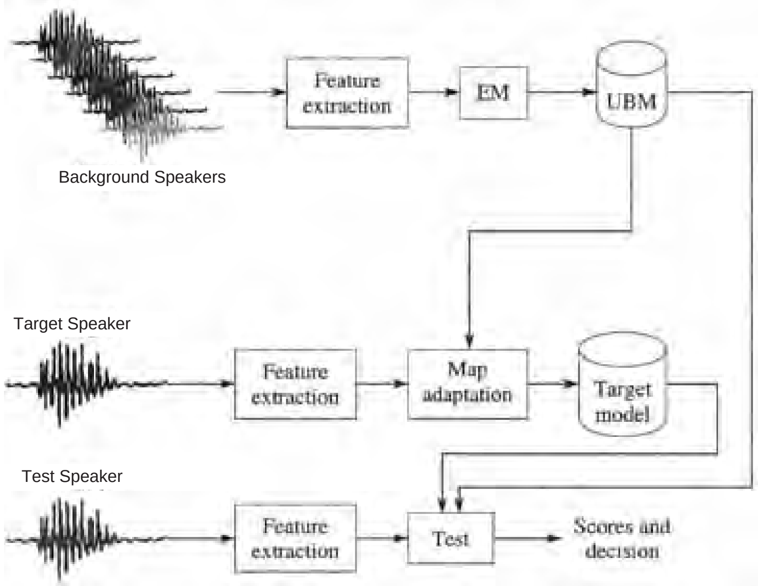


Figure 2.10 Architecture of the GMM-UBM system

Phonetically Structured GMMs

As mentioned, GMM-UBM has been one of the predominant modelling approaches, particularly in text-independent speaker recognition, despite the fact that the UBM is trained by “blindly” pooling all of the speech data together. This is because in the GMM-UBM framework, with enough mixture components (64 or more) the UBM component densities will represent the broad phonetic class distribution [91]. Given that English speech has no more than 45 or so distinct phones, one can surmise that with at least as many mixture components all the possible distinct ways a speaker can speak are modelled.

A recent variant of the traditional GMM approach proposed by Faltlhauser et al. [92] is the so-called “phonetically-structured” GMM (PGMM) method. This method trains smaller “granular” GMMs on separate phonetic classes for each speaker, then combines them into a larger single model which is used for recognition. By combining the various phonetic models using a globally determined weighting, this method is believed to be less sensitive to phonetic biases present in the enrolment data of individual speakers. Examples of the phonetic classes used are: vowels, strong fricatives, liquids, etc [93].

Ever since it has been introduced, many speaker recognition researchers have proposed different ways of building the phoneme-specific GMMs [92-95]. However experimental results so far indicate that generally PGMM is unable to perform as well as the phoneme independent GMM system (as in UBM-GMM). Hence, while the UBM performs an implicit (soft) acoustics ‘partitioning’ in the UBM-GMM paradigm and the MAP adaptation provides speaker discrimination [86], it seems that dividing the training data into phonetic classes to reduce the phonetic biases in the training data might not necessarily improve the system performance. This suggests that acoustic and speaker modelling are difficult to decouple in the current GMM paradigm (Ideally it would be nice to decouple them but it hasn’t been done). Furthermore, it raises the question of the relative importance of the acoustic and speaker discrimination in the speaker recognition problem.

2.4.2 Support Vector Machines

In recent years, support vector machine (SVM) has proven to be a new effective method for speaker verification [20, 87, 96]. The SVM is a binary classifier that makes its decisions by constructing a separating hyperplane that optimally separates the two classes as shown in Figure 2.11 [97]. Formally, it is constructed from sums of a kernel function $K(.,.)$

$$f(\mathbf{x}) = \sum_{i=1}^{L} \alpha_i t_i K(\mathbf{x}, \mathbf{x}_i) + d \tag{2.25}$$

where t_i are the class labels, d is a learned constant, $\sum_{i=1}^{L} \alpha_i t_i = 0$, $\mathbf{x}_i$ for which $\alpha_i > 0$ are the support vectors (nearest data points on each side of the hyperplane as shown in Figure 2.11) obtained from the training set using an optimisation procedure such as steepest-descent exact line search in SVMTorch [98, 99] and L is the number of support vectors. The class labels are either +1 or -1, depending upon whether the corresponding support vector is in class 0 or class 1, respectively. For classification, a class decision is based upon whether the value $f(\mathbf{x})$, is above or below a threshold.

The kernel $K(.,.)$ is constrained to satisfy the Mercer condition so that $K(.,.)$ can be expressed as

$$K(\mathbf{x}, \mathbf{x}_i) = b(\mathbf{x})^t b(\mathbf{x}_i) \tag{2.26}$$

where $b(\mathbf{x})$ is a mapping from the input space (where $\mathbf{x}$ lives) to a kernel feature space of much larger (possibly infinite) dimension. The kernel function allows computing inner products of two vectors in the kernel feature space [100].

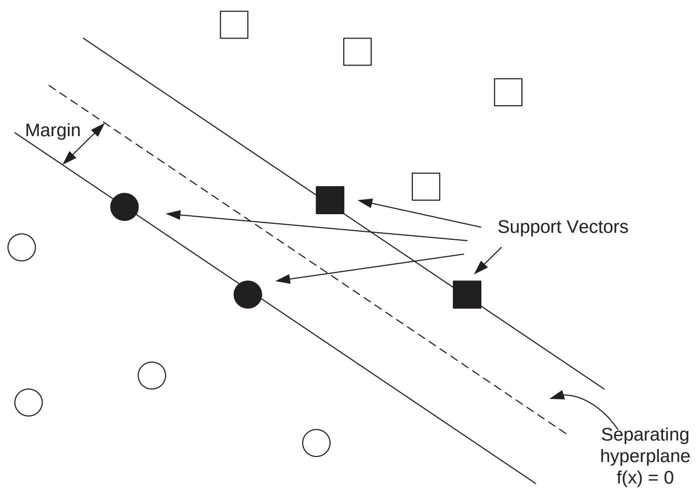


Figure 2.11 Support Vector Machine concept where ○/● and □/■ represent the training data from class 0 and 1 respectively.

Although SVMs have proven their effectiveness for speaker recognition tasks, a critical aspect of using SVMs successfully is the design of the kernel [100], the SVM cost parameter and kernel parameters [101-103]. Many researchers have committed considerable time to finding the optimum kernel functions for speaker recognition [9, 104-107] from a diverse set of available kernel functions. Although some comparative studies across different types of kernel have been conducted [4, 107], the choice of kernel for speaker recognition still remains as a trial-and-error procedure [16]. Among these, the generalised linear discriminant sequence (GLDS) kernels [104], Maximum likelihood linear regression (MLLR) kernels [106] and GMM supervectors [9] are some common kernels used, and GMM supervectors are employed in this thesis since they have been shown in [4] to perform better than the others. GMM supervectors, $\mathbb{M}$, are formed by concatenating all the mean vector elements ($\boldsymbol{\mu}_{i,j}$) normalised using the weights (w_i) and the diagonal covariance elements ($\sigma_{i,j}$) from the GMM-UBM model (discussed in section 2.4.1) as follows:

$$\mathbb{M} = \left[\frac{\sqrt{w_1}\mu_{1,1}}{\sqrt{\sigma_{1,1}}} \dots \frac{\sqrt{w_1}\mu_{1,D}}{\sqrt{\sigma_{1,D}}} \dots \frac{\sqrt{w_i}\mu_{i,1}}{\sqrt{\sigma_{i,1}}} \dots \frac{\sqrt{w_i}\mu_{i,j}}{\sqrt{\sigma_{i,j}}} \cdots \frac{\sqrt{w_i}\mu_{i,D}}{\sqrt{\sigma_{i,D}}} \dots \frac{\sqrt{w_M}\mu_{M,D}}{\sqrt{\sigma_{M,D}}} \dots \frac{\sqrt{w_M}\mu_{M,D}}{\sqrt{\sigma_{M,D}}}\right]^{\mathrm{T}} \tag{2.27}$$

where i is the index of the mixture component, j is the index of the feature dimension, M is the total number of mixture components and D is the number of dimensions of the feature vector. Since SVMs are not invariant to linear transformations in feature space, variance normalisation is performed so that some supervector dimensions do not dominate the inner product computations [4, 108]. Alternatively, the GMM supervector can be considered as a mapping from the spectral features of an utterance to a high-dimensional feature vector. The classification of GMM supervectors by SVMs is termed a Gaussian mixture model-support vector machines (GMM-SVM) system configuration.

Moreover, besides the factors as discussed above, it has recently been shown that the composition of speakers in the SVM background dataset has a significant impact on speaker verification performance [38, 109-111]. This is because the hyperplane that is trained using the target and background speakers' data tends to be biased towards the background dataset in a speaker verification task since the number of utterances from the target speaker (normally only one utterance) is usually much less than the background speaker (thousands of utterances). Therefore effective selection of the background dataset can improve the performance of a SVM-based speaker verification system. Researchers such as McLaren et al. [38, 110, 112] and Suh et al. [111] have extended their investigations of SVM for speaker recognition to SVM background speaker selection. In [38, 110], the support vector frequency was used to rank and select negative examples by evaluating the examples using the target SVM model, and then selecting the closest negative examples to the enrolment speaker as the background dataset. Their proposed technique results in a relative improvement of 10% in EER on NIST 2006 SRE database over a heuristically chosen set. Unless otherwise stated, the GMM-SVM is used as the baseline in this thesis.

2.4.3 Sparse Representation Classification

Sparse Representation

Widespread interest in sparse signal representations is a recent development in digital signal processing [28, 31, 113, 114]. The sparse representation paradigm, when it was originally developed, was not intended for classification purposes but instead for an efficient representation and compression of signals at a greatly reduced rate than the standard Shannon-Nyquist rate with respect to an overcomplete dictionary of base elements [115, 116]. Given a $K \times N$ matrix $\mathbf{D}$, where each column represents an individual vector from the overcomplete dictionary, with $N > K$ and usually $N >> K$, then the problem of identifying a sparse representation of a signal $\mathbf{S} \in \mathbb{R}^K$, becomes the problem of finding an $N \times 1$ coefficient vector $\boldsymbol{\gamma}$

$$\boldsymbol{\gamma} = \arg\min_{\boldsymbol{\gamma}'} \|\boldsymbol{\gamma}'\|_0 \qquad s.t. \quad \mathbf{S} = \mathbf{D}\boldsymbol{\gamma} \tag{2.28}$$

where $\|.\|_0$ denotes the ℓ_0-norm, which counts the number of nonzero entries in a vector. However, the problem of finding the sparsest solution of an underdetermined system of linear equations is NP-hard and difficult even to approximate [117]. Recent developments in sparse representation and compressive sensing [118-120] indicate that if the solution $\boldsymbol{\gamma}$ sought is sparse enough, the ℓ_0-norm in equation (2.28) can be replaced with an ℓ_1-norm as shown in equation (2.29), which can be efficiently solved by linear programming techniques.

$$\boldsymbol{\gamma} = \min_{\boldsymbol{\gamma}'} \|\boldsymbol{\gamma}'\|_1 \qquad s.t. \quad \mathbf{S} = \mathbf{D}\boldsymbol{\gamma} \tag{2.29}$$

Classification based on Sparse Representation

In classification problems, the main objective is to determine correctly the class of a test sample ($\mathbf{S}$) given a set of labelled training samples from L distinct classes. First, the l_i training samples $\mathbf{v}_{i,j} \in \mathbb{R}^K$ from the ith class are arranged as the columns of a matrix $\mathbf{D}_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \ldots, \mathbf{v}_{i,l_i}]$. If $\mathbf{S}$ is from class i, then $\mathbf{S}$ will approximately lie in the linear span of the training samples in $\mathbf{D}_i$ [28] and can be represented as follows

$$\mathbf{S} \approx \alpha_{i,1}\mathbf{v}_{i,1} + \alpha_{i,2}\mathbf{v}_{i,2} + \alpha_{i,j}\mathbf{v}_{i,j} + \cdots + \alpha_{i,l_i}\mathbf{v}_{i,l_i} \tag{2.30}$$

for some scalars, $\alpha_{i,j} \in \mathbb{R}, j = 1,2, \ldots, l_i$.

Since the knowledge of the membership i of the test sample is unknown during classification, a new matrix $\mathbf{D}$ is defined for the entire training set as the concatenation of the training samples of all L classes:

$$\begin{aligned} \mathbf{D} &= [\mathbf{D}_1, \mathbf{D}_2, \ldots, \boldsymbol{D}_L] \\ &= [\mathbf{v}_{1,1}, \mathbf{v}_{1,2}, \cdots, \mathbf{v}_{1,l_1}, \mathbf{v}_{2,1}, \mathbf{v}_{2,2}, \cdots, \mathbf{v}_{2,l_2}, \cdots, \mathbf{v}_{L,1}, \mathbf{v}_{L,1}, \cdots, \mathbf{v}_{L,l_L}] \end{aligned} \tag{2.31}$$

Then, the linear representation of $\mathbf{S}$ can be rewritten in terms of all training samples as

$$\mathbf{S} = \mathbf{D}\boldsymbol{\gamma} \tag{2.32}$$

where the coefficient vector termed sparse coefficients [28], $\boldsymbol{\gamma} = \left[0, \dots, 0, \alpha_{i,1}, \alpha_{i,2}, \dots \alpha_{i,l_i}, 0, \dots, 0\right]^T$ has entries that are mostly zero except those associated with the ith class after solving the linear system of equations $\mathbf{S} = \mathbf{D}\boldsymbol{\gamma}$ using equation (2.29). In this case, the indices of the sparse coefficients encode the identity of the test sample $\mathbf{S}$.

In recent years, sparse representation based classifiers have begun to emerge for various applications, and experimental results indicate that they can achieve a generalisation performance that is greater than or equal to that of other classifiers [28, 29, 31-34]. In the case of face recognition, Wright et al. cast the problem in terms of finding a sparse representation of the test image features with respect to the training set, whereby the sparse representation can be accurately and efficiently computed by ℓ_1-minimisation [28]. They exploit the following simple observation: if sufficient training data are available for each class, test samples are represented only as a linear combinations of the training samples from the same class, wherein the representation is sparse by excluding samples from other classes. They have shown an absolute accuracy gain of 0.4% and 7% over linear SVM and nearest neighbour methods respectively on the Extended Yale B database [121]. Further, in [29], Naseem et al. showed classification based on sparse representation to be a promising method for speaker identification. Their speaker identification experiments conducted on the TIMIT database [122] achieved better recognition accuracy using a sparse representation classifier (98.24%), as compared with GMM-SVM based (97.80%) and GMM-UBM based (96.93%) speaker identification systems. Although these initial investigations for the proposed sparse representation classifier for speaker identification were encouraging, the relatively small TIMIT database characterises an ideal speech acquisition environment and does not include e.g. reverberant noise and session variability.

Recently, a discriminative sparse representation classification, which focuses on achieving high discrimination between classes as opposed to the standard sparse representation that focuses on achieving small reconstruction error, was proposed specifically for classification tasks [31]. The results in [31] demonstrated that discriminative SRC is more robust to noise and occlusion than the standard SRC for signal classification. The discriminative approach works by incorporating an additional Fisher's discrimination power to the sparsity property in the standard sparse representation. Our initial investigation was unsuccessful since the discriminative SRC requires the computation of the Fisher F-ratio (ratio of between-class and within-class variances) [123] with multiple samples per class. However for the task of speaker verification (which is a two class problem) with only one sample for the target class, the within-class scatter for the target class always goes to zero.

2.5 Robustness and Channel Compensation

One of the major challenges to improving accuracy in state-of-the-art speaker recognition algorithms is reducing the impact of variation in transmission channel and handset that occurs between testing and training data and/or additive noise on system performance [124, 125]. For example, training data for an individual may be obtained via one channel (e.g. a carbon-button microphone telephone), and test data via another (e.g. cellphone). In this case, impostors using carbon-button handsets are more likely to match the target speaker than usual, and the target speaker on a cell telephone is more likely to be incorrectly rejected. During the past years, much research has been conducted towards reducing the effect of channel mismatch. Generally, robustness in speaker verification can be improved at all three stages: the feature extraction stage, the modelling stage and the classification stage as shown in Figure 2.12.

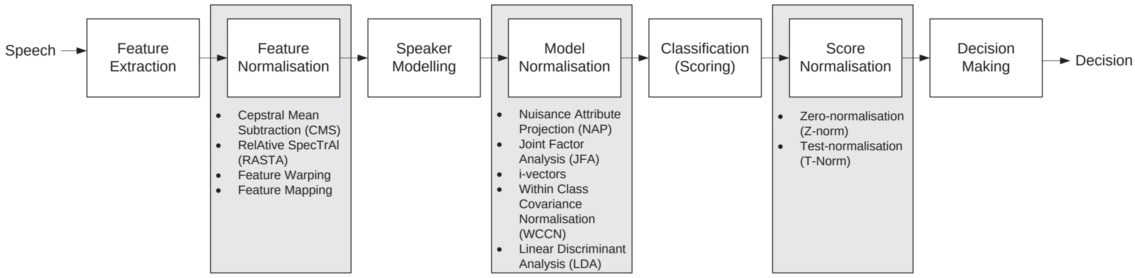


Figure 2.12 A schematic block diagram of a typical speaker verification system with normalisation techniques.

2.5.1 Feature-based Normalisation

In the feature space, Cepstral Mean Subtraction (CMS) [126] and RelAtive SpecTrAl (RASTA) filtering [127] are the simplest forms of normalisation in speaker recognition, used to remove slowly varying convolutive noise. In CMS, the mean is computed over the entire utterance and subtracted from each feature since convolutive channel noise becomes additive in the log-spectral or cepstral domain. On the other hand, RASTA works by band pass filtering the speech in the log-spectral or cepstral domain. The filter is applied along the temporal trajectory of each feature, and it suppresses modulation frequencies which are outside of typical speech signals. Although CMS and RASTA are effective at reducing the distortions, they have been shown to also remove beneficial speaker-specific information in [128] and [129] respectively.

The more recent and more successful channel compensation techniques at the feature level are feature warping [129] and feature mapping [130]. Feature warping maps the cumulative distribution of a cepstral feature stream to a standardised distribution, for example a normal distribution, over a specified time interval. This transformation is performed using a table that establishes the correspondence between the acoustic feature's cumulative distribution and the cumulative normal distribution. Feature mapping is a supervised normalisation method which transforms the features obtained from different channel conditions into a channel-independent feature space such that channel variability is reduced. This is achieved with a set of channel-dependent GMMs adapted from a channel-independent root model. In the training or operational phase, the most likely channel (highest GMM likelihood) is detected, and the relationship between the root model and the channel-dependent model is used for mapping the vectors into a channel-independent space. In this thesis, unless otherwise mentioned, feature warping is used on all features over feature mapping because feature warping does not require channel-labelled training data.

2.5.2 Model-based Normalisation

In the model space, Nuisance Attribute Projection (NAP) [124, 125], Joint Factor Analysis (JFA) [10, 131] and *i-vectors* [132] are amongst the main channel compensation techniques commonly used.

Nuisance Attribute Projection

NAP [124, 125] is a successful method for compensating SVM supervectors. It attempts to remove the effects of channel variability or nuisance by projecting the data in the SVM kernel space onto a subspace less prone to variation as follows:

$$\widehat{\mathbb{M}} = \mathbb{M} - \mathbf{U}(\mathbf{U}^{\mathbf{T}}\mathbb{M}) \tag{2.33}$$

where $\widehat{\mathbb{M}}$ is the nuisance removed supervectors and $\mathbf{U}$ is the eigenchannel matrix trained using a development dataset with a large number of speakers, each having several training utterances with several variability and all possible nuisances. The training set is prepared by subtracting the mean of the supervector within each speaker and pooling all the supervectors from different speakers together forming a matrix **A**. This is assumed to remove most of the between-speaker variability and preserve the session and channel variability. Then, by performing eigen-analysis on **A**, the principal directions of the nuisance subspace corresponding to the most dominant within-speaker variability are identified. Then *K* eigenvectors corresponding to the *K* largest eigenvalues are used to form the eigenchannel matrix, **U** as shown in equation (2.33).

Joint Factor Analysis

JFA is a method used for modelling inter-speaker variability and compensating channel/session variability in the context of GMM based classifiers. In the JFA framework [133, 134], a speaker and channel-dependent supervector **M** is viewed as a combination of two components: a speaker supervector **s** and a channel supervector **c**.

$$\mathbb{M} = \mathbf{s} + \mathbf{c} \tag{2.34}$$

The speaker and channel supervectors are represented as

$$\mathbf{s} = \mathbf{m} + \mathbf{Dz} + \mathbf{Vy} \tag{2.35}$$

$$\mathbf{c} = \mathbf{Ux} \tag{2.36}$$

where m is the speaker- and channel-independent supervector (generally from the UBM), V and U are rectangular matrices of low rank (typically 300 and 100 respectively) representing the principal direction of the speaker and channel variability, D is a diagonal matrix modelling the residual variability and y, z, x are independent random vectors having standard normal distributions. The components of y, z, x are termed the speaker, common and channel factors in the JFA model respectively.

Collectively, the matrices **U**, **V** and **D** are called the hyperparameters of the JFA model and are usually estimated beforehand on large labelled development datasets. For a given training sample, the latent factors **x** and **y** are jointly estimated, followed by the estimation of $\mathbf{z}$. Then, the channel supervector $\mathbf{c}$ is discarded and the speaker supervector $\mathbf{s}$ is used as the speaker model. By doing so, channel compensation is accomplished via the explicit modelling of the channel component during training.

i-vectors

The above classical joint factor analysis modelling based on speaker and channel factors consists in defining two distinct spaces: the speaker space defined by the eigenvoice matrix $\mathbf{V}$ and the channel space defined by the eigenchannel matrix $\mathbf{U}$. Recently, Dehak defined a new space, termed the "total variability space", which contains the speaker and channel variabilities simultaneously [16]. In the new model, no distinction between the speaker effects and the channel effects in GMM supervector space is made because experimental work carried out in [16] shows that channel factors estimated using JFA, which are supposed to model only channel effects, also contained information about speakers. Therefore given an utterance, the new speaker- and channel-dependent GMM supervector defined in (2.34) – (2.36) is rewritten as follows:

$$\mathbb{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{2.37}$$

where $\mathbf{T}$ is a rectangular matrix of low rank and $\mathbf{w}$ is a random vector having a standard normal distribution. The components of the vector $\mathbf{w}$ are the total factors referred to as the identity vectors or *i-vectors* for short. Channel compensation is then performed in the total factor space using within class covariance normalisation (WCCN) [135], and/or linear discriminant analysis (LDA) [132]. In the total factor space, a new classification method based on cosine distance, termed the Cosine Distance Scoring (CDS) classifier, as shown in equation (2.38) is then used for classification, where $\mathbf{w}_{test}$ and $\mathbf{w}_{target}$ are the test and target speaker's i-vectors respectively and $\langle .,. \rangle$ denotes the inner product.

$$\text{score}(\mathbf{w}_{test}, \mathbf{w}_{target}) = \frac{\langle \mathbf{w}_{test}, \mathbf{w}_{target} \rangle}{\|\mathbf{w}_{test}\| \|\mathbf{w}_{target}\|} \tag{2.38}$$

Within Class Covariance Normalisation

WCCN was introduced in [135] for minimising the expected error rate of false acceptances and false rejections during the SVM training step. However, in the context of *i-vectors*, WCCN is used for normalising the direction of the total factor components, without removing any nuisance direction [16]. The WCC matrix is computed as

$$W = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{n_c}\sum_{i=1}^{n_c}(\mathbf{w}_i^c - \overline{\mathbf{w}_c})(\mathbf{w}_i^c - \overline{\mathbf{w}_c})^t \tag{2.39}$$

where $\overline{\mathbf{w}_c} = \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{w}_i^c$ is the mean of total factors vectors of each speaker, C is the number of speakers and n_c is the number of utterances for each speaker c. Then a mapping function φ_{wccn} is defined as

$$\varphi_{wccn}(\mathbf{w}) = \mathbf{B}^t\mathbf{w} \tag{2.40}$$

where $\mathbf{B}$ is obtained through Cholesky decomposition of matrix $\mathbf{W}^{-1} = \mathbf{B}\mathbf{B}^t$.

Linear Discriminant Analysis

LDA [132] is a technique for dimensionality reduction that is widely used in the field of pattern recognition. It attempts to seek new orthogonal axes that give better discrimination between pairs of classes by maximising between-class variance and minimising intra-class variance. In general, the purpose of LDA is to maximise the Rayleigh coefficients as shown in equation (2.41) for space direction v. The Rayleigh coefficients represent the amount of information ratio of the between-class variance S_b and within-class variance S_w as shown in equation (2.42) and (2.43) respectively.

$$J(v) = \frac{v^t S_b v}{v^t S_w v} \tag{2.41}$$

$$S_b = \sum_{s=1}^{S} (y_s - \bar{y})\,(y_s - \bar{y})^t \tag{2.42}$$

$$S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (y_i^s - \bar{y_s})\,(y_i^s - \bar{y_s})^t \tag{2.43}$$

This maximisation is used to define a projection matrix **A** composing of the *k* best eigenvectors (for dimension reduction: only the top *k* eigenvectors with highest eigenvalues will be retained) of the eigenvalue equation

$$S_b v = \lambda S_w v \tag{2.44}$$

where λ is the diagonal matrix of eigenvalues. Similar to WCCN, the total factors are then submitted to the projection matrix **A** obtained from LDA as follows

$$\varphi_{LDA}(\mathbf{w}) = \mathbf{A}^t \mathbf{w} \tag{2.45}$$

Despite the findings by Dehak as mentioned above, JFA is chosen over *i-vectors* in the experiments reported in this thesis since the comparison of the recognition performance in [136] indicated that JFA consistently outperforms *i-vectors*.

2.5.3 Score-based Normalisation

Score normalisation attempts to remove the effect of noise, channel variability and session variability by modifying the score distribution. The basic normalisation uses the world model which is based on Bayes' Theorem, or a cohort which is based on a set of speakers closest to the target speaker [137]. A score normalisation usually takes on the form

$$s' = \frac{s - \mu_I}{\sigma_I} \tag{2.46}$$

where s' and s are the normalised and original score respectively, and μ_I and σ_I are the estimated mean and standard deviation of the impostor score distribution respectively. In Zero-normalisation (Z-norm) [138], the impostor statistics μ_I and σ_I are estimated from the scores of a set of impostor speaker utterances tested against the target speaker model. On the other hand, in Test-normalisation (T-norm) [137] the parameters, μ_I and σ_I, are estimated from scores of each test segment against a set of impostor speaker models at test time. Therefore, Z-norm has an advantage in that the estimation of the normalisation parameters can be performed offline in the speaker enrolment phase. However T-norm avoids the test-to-normalisation mismatches which are possible in Z-norm since the mean and variance normalisation parameters are estimated from the test utterance.

Z-norm is mostly utilised to scale various output scores caused by different speaker models, and T-norm is to transform output scores caused by various test utterances. In order to enhance the robustness of decision threshold and normalise the uncertainty of score variability between trials entirely, two kinds of combination mode, Test-dependent zero-score normalisation (TZ-norm) and Zero-dependent test-score normalisation (ZT-norm) were proposed in [139]. In this thesis, unless otherwise mentioned, ZT-norm is used for score normalisation.

2.6 Fusion

The vast range of feature extraction and modelling processes described thus far brought about opportunities to exploit the complementary information that exists between these different techniques to obtain more robustly trained speaker models, and subsequently, improved verification scores [23, 140, 141]. In current speaker recognition systems, the combination of information from multiple sources of evidence, referred to as fusion, is widely applied. In the literature, fusion is divided into two main approaches: *feature-level fusion* wherein two or more features are concatenated before passing through the classifier and *score-level fusion* which refers to the soft combination of classifier outputs [142].

For feature-level fusion, it is necessary that the individual feature vectors be available at the same frame rate (i.e., the feature extraction must be synchronous). In many approaches, some features (e.g. pitch, only defined in voiced speech frames) might not be generated for a complete utterance and are therefore not available synchronously to the spectral frame-based features (e.g. MFCC). Furthermore, feature concatenation introduces long feature length which may leads to curse of dimensionality since the number of training vectors needed for robust density estimation increases exponentially with the dimensionality [141]. Thus, score-level fusion is typically utilised in current speaker recognition systems. In score fusion, each individual data source is modelled separately, and the outputs of the individual classifier scores are combined as a weighted sum to give the overall match score. That is, given the sub-scores s_k, where k is the index of the classifier, the fused score is $s = \sum_{n=1}^{N_c} w_n s_n$. Here N_c is the number of classifiers and w_n is the fusion weight which determines the relative contribution of the n^{th} classification system. The fusion weights w_n are determined based on logistic regression [143] using the Fusion and Calibration (FoCal) toolkit[5] in this thesis. In this thesis, unless otherwise mentioned, score-level fusion is used over feature-level fusion.

In general, performance improvement can be attained through score fusion of systems with different front-ends or back-ends [144]. Examples of this kind of variation in front-ends include the use of different voice activity detectors (VADs) and/or feature extraction techniques across systems [23, 141]; and in back-ends, these include the use of different

[5] https://sites.google.com/site/nikobrummer/focal

classifiers and/or compensation techniques [20, 21, 23] across systems. All are typical of NIST SRE consortium submissions.

2.7 Performance Measures

In speaker verification systems, misses and false alarms are the two types of errors that can happen. A miss occurs when a valid identity is rejected and false alarm occurs when an invalid identity is accepted. The probability of miss is estimated as the ratio of the number of falsely rejected speaker tests to the total number of correct speaker trials. Similarly, the probability of false alarm is estimated as the ratio of the number of falsely accepted speaker tests to the total number of impostor trials. By varying the decision threshold, the miss and false alarm probabilities can be changed in opposing directions. For higher decision thresholds, false alarm will be fewer but misses will be more common. On the other hand, for lower decision thresholds, false alarm will be more common but misses will be fewer.

Given a fixed decision threshold, the Detection Cost Function (DCF) [145] can be used as a performance measure of a speaker verification system. It is defined as the weighted sum of the miss probability ($P_{Miss|Target}$) and false alarm probability ($P_{FalseAlarm|NonTarget}$) as follows

$$\begin{aligned} DCF = C_{Miss} \times P_{Target} \times P_{Miss|Target} + C_{FalseAlarm} \\ \times \left(1 - P_{Target}\right) \times P_{FalseAlarm|NonTarget} \end{aligned} \tag{2.47}$$

where C_{Miss} and $C_{FalseAlarm}$ are the relative costs of detection errors and P_{Target} is the a priori probability of the target. The typical values for the cost parameters and a priori probabilities are $C_{Miss} = 10$, $C_{FalseAlarm} = 1$ and $P_{Target} = 0.01$. In the NIST 2010 speaker recognition evaluation[6], the interest point was shifted to lower false alarms by setting $C_{Miss} = 1$ and $P_{Target} = 0.001$. The value of DCF depends on the value of the decision threshold. The MinDCF is the minimum value of the DCF obtained when the decision threshold is changed. This last value was used as the principal metric in the most recent NIST 2010 speaker recognition evaluation campaign.

The Equal Error Rate (EER) is another criterion used to compare the performance of speaker verification systems. It represents the operating point where the false alarm probability is equal to the miss probability as shown in Figure 2.13.

In addition to the scalar measures of EER and MinDCF, the detection error trade-off (DET) curve [146] is also used as a performance measure. The DET curve is the curve of miss probability plotted against false alarm probaility across different decision thresholds, as shown in Figure 2.13 where System 1 outperforms System 2 in terms of EER and minDCF.

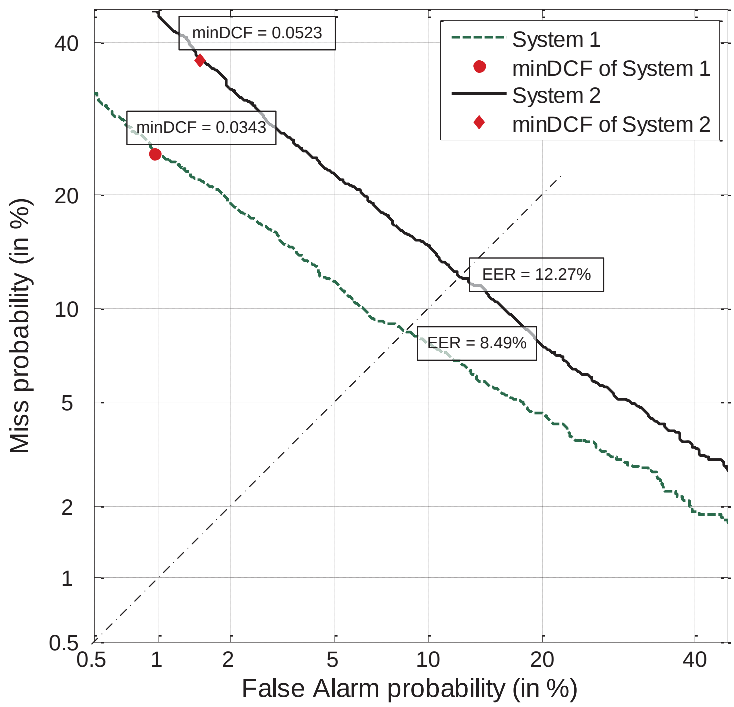


Figure 2.13 Plot of a DET curve for a speaker recognition task

[6] http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf

2.8 Databases

This section details a number of corpora available for evaluating the performance of a speaker verification system. Generally, the data involved in building an automatic speaker verification system is split into three sets. Namely, the background set, development set and the evaluation set. The background set is a huge dataset containing speech from many speakers from the expected target population. It is used to train the UBM, subspace models and/or to draw impostor models for score normalization. The development set is an evaluation database that is used for tuning system parameters to ensure classification performance is maximized for the expected conditions of audio acquisition. The evaluation set is a second independent evaluation database that is used to evaluate the final system (that was optimized on the development set).

2.8.1 Switchboard Series of Corpora

The Switchboard series of corpora was collected by the Linguistic Data Consortium (LDC) as part of the Effective, Affordable, Reusable Speech-to-text (EARS) project, sponsored by the Defense Advanced Research Projects Agency (DARPA) [147]. The Switchboard series comprises several releases recorded from early 90's until 2004.

Switchboard-2 Phase I consists of 3,638 five-minute telephone conversations from 657 participants which were mainly recruited from US universities [148]. Each participant was asked to take part in at least ten five-minute phone calls. Phase II consists of 4,472 conversations involving 679 participants [149]. Phase III consists of 5,456 sides (2,728 calls) from 640 participants under varied environmental conditions (i.e. indoors, outdoors or moving vehicle) [150].

The Switchboard Cellular Part 1 focused primarily on GSM cellular phone technology with 2,618 sides (1,309 calls) from 254 pariticipants under varied environmental conditions [151]. Part 2 focused on cellular phone technology from a variety of network types (i.e. CDMA, GSM or TDMA), with CDMA technology being the dominant type [152]. A total of 4,040 sides (2,950 cellular calls) from 419 participants were collected.

2.8.2 Mixer Corpora

The Mixer corpora [153] was collected in three phases with approximately 600 subjects completing 10 calls and 550 completing at least 20 calls of at least 6 minutes duration using unique handsets and multichannel recording devices for a subset of calls. In contrast to the Switchboard corpora, the Mixer corpora collected speech spoken in a number of languages from both native and non-native English speakers. The languages in the Mixer corpora include Arabic, Mandarin, Russian, Spanish and English. Bilingual speakers complete at least four calls in languages other than English as well as additional calls in English. Notably, since the Switchboard and Mixer corpora contain many recordings from the same speakers over diverse session characteristics, they are used in this thesis to estimate characteristics of intersession variability (as discussed in section 2.5.2).

2.8.3 NIST Speaker Recognition Evaluation Corpora

The NIST Speaker Recognition Evaluation (SRE) is part of an ongoing series of evaluations conducted by NIST since 1996 [154]. The overarching objective of the evaluations has always been to drive the technology forward, to measure the state-of-the-art, and to find the most promising algorithmic approaches for text independent speaker recognition systems [154].

The NIST SRE 2001 core conditions marked the first trial of cellular data evaluation with data sourced from the Switchboard Cellular corpus. A number of challenging conditions was introduced into the SREs, this includes factors such as voice compression and rapidly changing environment conditions due to the nature of mobile communication.

In contrast to previous years' evaluations whereby data were sourced from the Switchboard corpus, the Mixer data were introduced into the NIST SRE 2004 – 2006 evaluations. The NIST SRE 2004 corpus consists of 10,743 telephone call segments recorded from 480 participants (181 Male, 299 Female) over landline and cellular phones. The NIST SRE 2005 corpus consists of 16,537 telephone call segments recorded from 528 participants (220 Male, 308 Female). The NIST SRE 2006 corpus consists of 24,637 telephone call segments recorded from 1088 participants (462 Male, 626 Female). Furthermore, telephone calls were recorded over auxiliary microphones (*mic*) of eight different kinds and many segments have different lengths (from 10 seconds to five minutes). Apart from the native English spoken data, the corpora consist of non-native English speech and speech corresponding to other languages.

In the 2008 and 2010 evaluation, NIST broadened the scope of the evaluation by introducing interview speech (*int*) that was recorded over several microphones. However, this was a somewhat difficault task in the NIST SRE 2008 evalaution due to the lack of microphone-recorded development data available at the time of evaluation. Therefore in the 2010, an additional set of interview data termed the NIST SRE 2008 *follow-up* corpus was released for system development leading to the NIST SRE 2010 evaluation.

2.9 Summary

This chapter provided an overview of the speech production organs in humans that carry speaker-specific properties. The vocal tract is one of the most important organs, in addition to the vocal cord, to characterise speakers and its effect on the spectrum of speech is discussed. The effect of vocal tract on the frequency spectrum is mainly in higher frequency areas (i.e. above 2.5 kHz), demonstrating the non-uniform distribution of speaker-specific information in the speech signal.

In the front-end of the automatic speaker recognition systems, the most successful features currently being used are cepstral features, which are computed purely from the magnitude spectrum of each frame of data, ignoring the phase spectrum. However, recent research has shown success fusing phase-based features with magnitude-based features (MFCCs) for speaker recognition. Hence one of the aims in this work is to explore new phase-based features in particularly the group delay-based feature and frequency modulation-based feature for the purpose of developing an enhanced automatic speaker recognition system. Nevertheless, although complementary features have been widely researched in speaker recognition system, why one feature (in particularly MFCCs) usually outperforms another feature and what represent a good pair of features for the purpose of fusing complementary speaker recognition systems have yet to be explored in detail.

Following feature extraction, some of the back-ends which are used in current state-of-the art systems, namely Gaussian mixture model-universal background model (GMM-UBM) and support vector machine (SVM) were discussed in some detail. For speaker recognition, GMM-UBM has been one of the predominant modelling approaches, particularly in text-independent tasks, because of the ability of the UBM component densities to represent the broad phonetic class distribution with the MAP adaptation providing speaker discrimination. However the relative importance of the acoustic and speaker modelling in the speaker recognition problem has yet to be analysed in detail.

Furthermore, a brief background on sparse representation classification (SRC) was given. Its ability to achieve comparable performance to SVM in various other applications (i.e. face recognition, speaker identification, etc), without the need of a training phase and time-consuming parameter tuning (kernel selection, the SVM cost parameter and kernel parameters) as in SVM, has been shown in literature. Although SRC has been applied in speaker identification systems, the relatively small TIMIT database used for testing, characterises an ideal speech acquisition environment and does not include e.g. reverberant noise and session variability. Moreover, the application of SRC in speaker verification systems has yet to be explored.

Chapter 3

Proposed Phase and Frequency Based Features

Due to the increasing use of fusion in speaker recognition systems, features that are complementary to MFCCs offer opportunities to advance the state of the art. Two promising classes of features are phase-based features (i.e. group delay) and frequency-based features (i.e. frequency modulation and subband spectral centroid), as reviewed briefly in the previous chapter.

This chapter will first provide a detailed discussion of the group delay (GD) feature used in speaker recognition systems. It will then explore the use of least squares regularisation for reducing the large variability in GD features caused by zeros of z-transform polynomial of speech signal. Then in Section 3.2, a highly computationally efficient method to extract frequency based features, Spectral Centroid Frequency (SCF) will be investigated. Furthermore, an analytically complementary feature to SCF, the Spectral Centroid Magnitude (SCM) will be introduced either as a more accurate representation of the subband energy compared with MFCC or for use as complementary features to MFCCs.

3.1 Proposed Group Delay Features

As mentioned in section 2.3.2.1, the conventional group delay estimate is not a suitable feature for speaker recognition because it suffers from extreme estimates creating undesirable variability in the feature distribution, resulting in difficulties for use in speech processing applications. Hence several techniques have been proposed to suppress the peaks as discussed in section 2.3.2.1.

The various types of group delay namely conventional group delay, cepstral smoothed group delay, modified group delay and log compressed group delay (expressed in equation (2.4) – (2.7) respectively) are shown in Figure 3.1(b) – (e) respectively for comparison. The magnitude response of a 10^{th} order LPC and FFT magnitude spectrum is given in Figure 3.1(a), showing the location of the estimated formant frequencies. As shown in Figure 3.1(b), it can be observed that the peaks are highly pronounced in the conventional approach to extracting the group delay (equation 2.4), which masks the group delay information corresponding to the vocal tract system (peaks at formant frequencies are not visible). In contrast, the formant peaks are visible in the cepstrally smoothed GD, modified GD (MODGD) and log compressed GD (LogGD) as shown in Figure 3.1(c) – (e) respectively. Although the formant peaks are visible in the cepstrally smoothed GD (Figure 3.1(c)), the dynamic range of the GD variation still remains relatively large. On the whole, the MODGD and LogGD seem to give the best representation through the suppression of the peaks in Figure 3.1(c). However, despite the fact that the MODGD has shown success in the task of speaker identification, it requires an extensive search of empirical parameters, γ and β, which are data dependent [24].

In an attempt to circumvent the above problems, alternative group delay features regularised using least squares approach is proposed in this section.

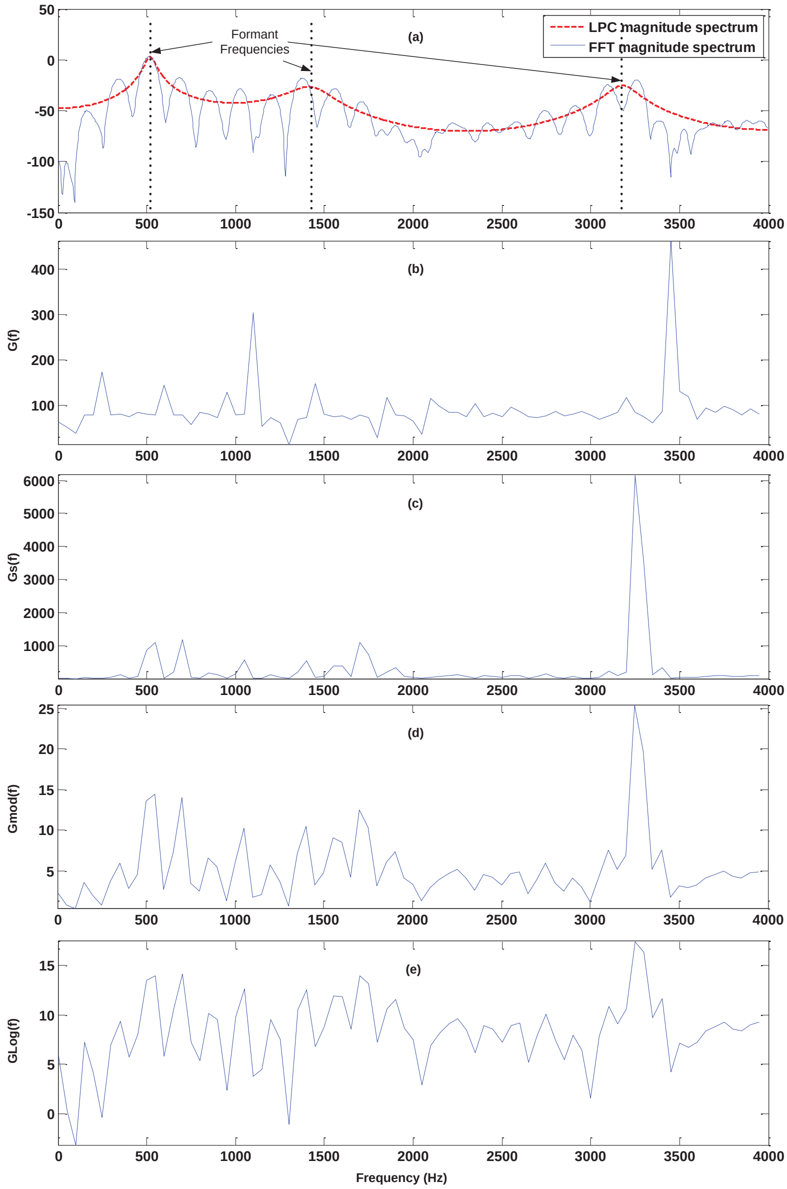


Figure 3.1 Comparisons of existing group delay spectra for a 20ms voiced frame of speech (a) Magnitude Spectrum (b) Conventional group delay, $G(f)$ (c) Cepstral smoothed group delay, $G_s(f)$ (d) Modified group delay, $G_{mod}(f)$ with γ=0.9 and β=0.4 (e) Log compressed group delay, $G_{Log}(f)$.

3.1.1 Proposed Least Squares Regularisation of Group Delay Features

As indicated in equation (2.5), smaller values of $|\mathbb{S}(f)|^2$ (when zeros of the excitation component are located close to the unit circle) lead to large-amplitude peaks in the group delay function [65] which can commonly create large variability in cepstrally smoothed group delay, $G_s(f)$ as shown in Figure 3.2 (b). Here, we achieve a smoother estimate which we term the least square group delay, $G_{LS}(f)$, using least squares (LS) regularisation. This is achieved by rewriting equation (2.5) in a matrix-vector notation over a window of length L

$$\mathbf{n} \approx G_{LS}(f)\mathbf{d} \tag{3.1}$$

where $\mathbf{n}$ and $\mathbf{d}$ are the numerator and denominator of equation (2.5) calculated at consecutive points along a frequency domain window of length L as in equations (3.2)–(3.3) and (3.4)–(3.5) respectively, where the subscript k denotes the frequency index.

$$\mathbf{n} = \begin{bmatrix} n(f_k) \\ \vdots \\ n(f_{k-L}) \end{bmatrix} \tag{3.2}$$

$$n(f_k) = S_R(f_k)F_R\{ts(t)\} + S_I(f_k)F_I\{ts(t)\} \tag{3.3}$$

$$\mathbf{d} = \begin{bmatrix} |\mathbb{S}(f_k)|^2 \\ \vdots \\ |\mathbb{S}(f_{k-L})|^2 \end{bmatrix} \tag{3.4}$$

$$d(f_k) = |\mathbb{S}(f_k)|^2 \tag{3.5}$$

Then

$$G_{LS}(f) = [\mathbf{d}^T\mathbf{d}]^{-1}\mathbf{d}^T\mathbf{n} \tag{3.6}$$

As shown in Figure 3.2(d), following least squares regularisation, the dynamic range of variation is reduced when compared with cepstrally smoothed GD (Figure 3.2(b)). Significant additional compression can be attained by incorporating log compression, without reducing the significance of the peaks relative to the remainder of the GD spectrum resulting in the log compressed least square group delay, G_{LogLS}, as shown in Figure 3.2(e). Compared with log compression (Figure 3.2(c)), G_{LogLS} preserves the relative significance of the GD spectral peaks more effectively.

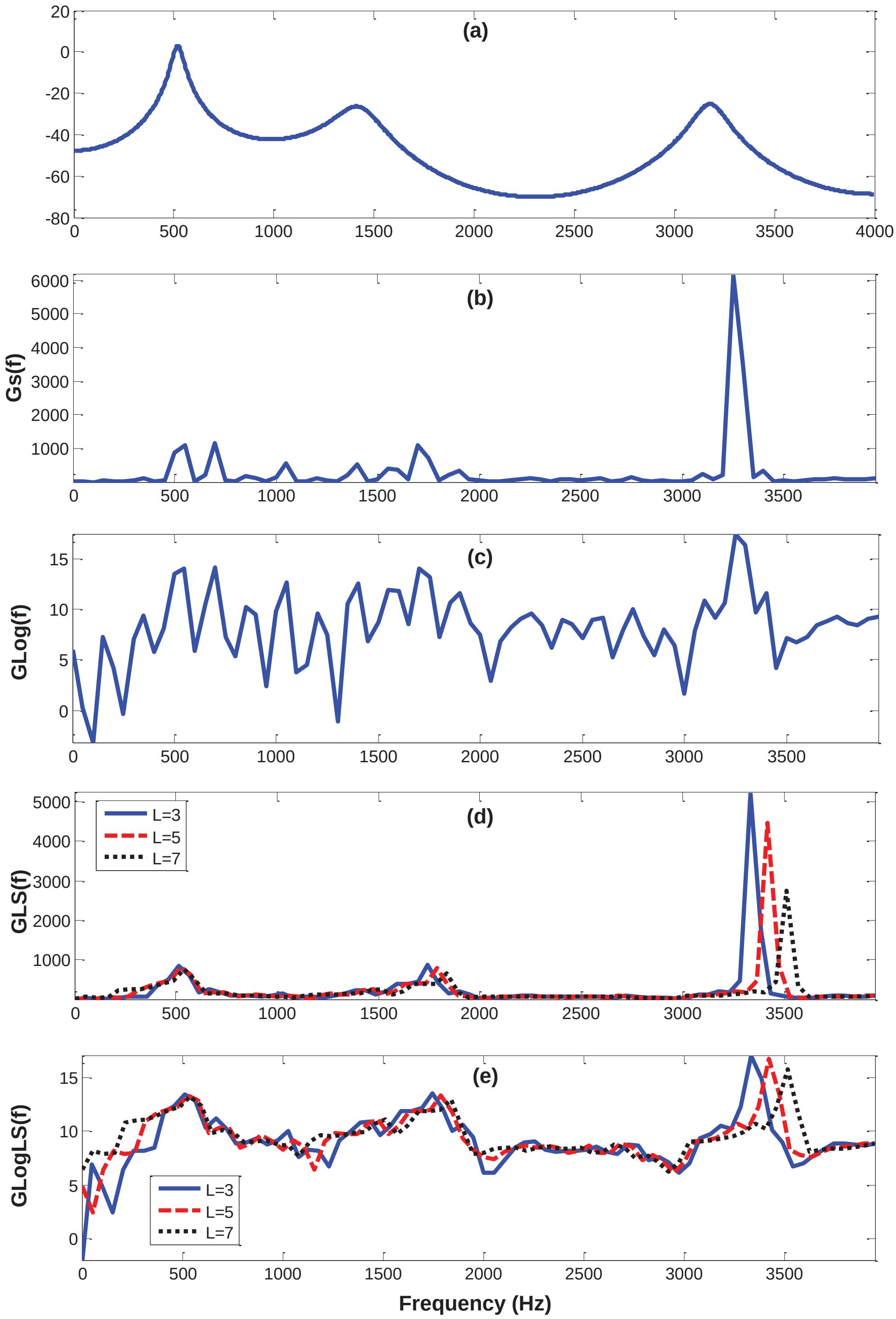


Figure 3.2 Comparison of existing and proposed group delay spectra for a 20 ms voiced frame of speech. (a) Magnitude spectrum (b) Cepstral smoothed group delay , $G_s(f)$ (c) Log compressed group delay, $G_{Log}(f)$ (d) Least squares regularised group delay, $G_{LS}(f)$ (e) Log compressed least square regularised group delay, $G_{LogLS}(f)$. As expected, longer regularisation windows L produce smoother group delay spectra

3.1.2 Group Delay Feature Extraction

The group delay features are extracted from speech using frames of length 20 ms, overlapping by 10 ms. In order to decorrelate the resulting GD features, a Discrete Cosine Transform (DCT) is applied to the GD function and the first 14 DCT coefficients are taken as a feature vector [64]. If log compression is desired, the absolute value is taken before the compression. For the purpose of comparison, the modified group delay (MODGD) [24] and log compressed group delay (LogGD) [25] are used as the baselines. As it is very time consuming to determine the optimal parameters values of γ and β for MODGD, a limited number of experiments were performed and the best parameters are used in the ensuing comparisons.

3.1.3 Evaluation

Speaker recognition experiments were conducted using the NIST 2001 SRE database and core condition of the NIST 2006 SRE database (1conv4w-1conv4w). The back-end of the recognition system for the NIST 2001 database was based on GMM-UBM with 512 mixtures. For UBM creation, the development set of NIST 2001 SRE database was used. For this and following experiments, the baseline system is based on 16 dimensional MFCC with appended delta coefficients (unless specified). The MFCC features were extracted every 10 ms, using Hamming analysis window of 20 ms and a filterbank of 26 triangular mel-spaced filters. The features were also normalised using feature warping.

Results for speaker recognition experiments based on the NIST 2001 SRE database for MFCC, MODGD, LogGD, LSGD and LogLSGD are given in Table 3.1 and Figure 3.3. The LogLSGD feature with L = 3 gave the best performance among the different approaches. This is mainly because by employing a least squares approach prior to log compression, a better numerical stability could be achieved as compared with log compression alone, while preserving the small variation as shown in Figure 3.2(e). However there is a trade-off between the window length and spectral blurring where the performance degrades slightly for longer windows. Furthermore, it has been noted that LSGD gave the worst performance which is mainly due to the large dynamic range in the GD estimates as shown in Figure 3.2(d); supporting the claim that having smooth estimates in the extracted features is essential. This was further validated with the invariant cluster separation index in the feature space study to determine the degree to which speakers can be separated across different group delay features (results are shown in Appendix A).

Table 3.1: Comparison of GD feature extraction techniques for speaker recognition on the NIST 2001 SRE database

Features	Window Size (*L*)	EER (%)	Fused EER (%) with MFCCs
MFCC	-	8.49	-
MODGD	-	13.35	8.48
LogGD	-	11.73	8.09
LSGD	3	17.86	8.04
LogLSGD	3	**10.01**	**7.82**
LogLSGD	5	10.16	7.94
LogLSGD	7	10.26	7.89

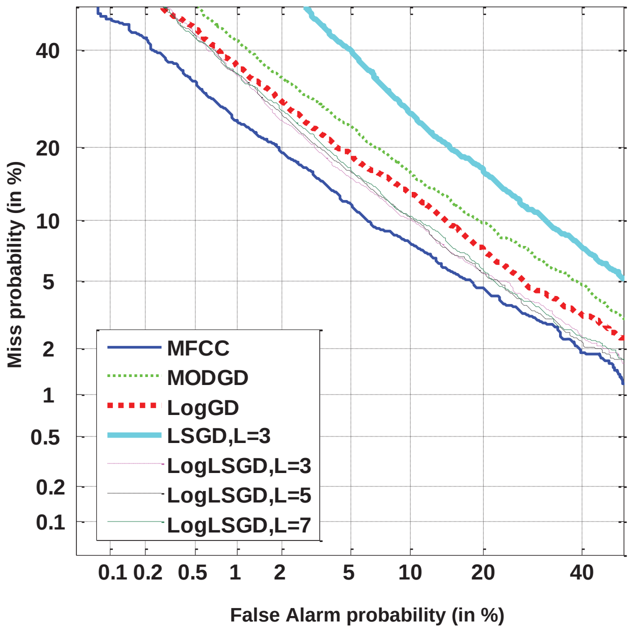


Figure 3.3 DET curves for various MFCC and group delay based speaker recognition systems, tested on NIST 2001.

Finally, the combination of MFCC and LogLSGD (*L*=3) was evaluated on the larger and more contemporary NIST 2006 SRE database, in order to see the database independency of the results. The back-end was based on the GMM-SVM technique. The background data consists of 3079 speech utterances from the NIST 2004 SRE, which cover a number of speakers (female and male). The Nuisance Attribute Projection (NAP) training data includes approximately 10000 speech utterances from the NIST 2004 and 2005 SRE corpus. The training data in the NIST 2004 SRE corpus and NIST 2005 SRE corpus are used for training cohort models in Z-norm and T-norm score normalization respectively.

The results and the DET curves are shown in the Table 3.2 and Figure 3.4 respectively. The improvements discussed on NIST 2001 SRE database were also found for the more contemporary NIST2006 database, where LogLSGD improved on a 5.09% EER MFCC baseline to 4.54% after fusion as shown in Figure 3.4.

Table 3.2: Speaker recognition results for MFCC, LogLSGD and fused system on the NIST 2006 SRE database with speaker detection cost model parameters of $C_{Miss} = 10$, $C_{FalseAlarm} = 1$, $P_{Target} = 0.01$

Features	EER (%)	minDCF
MFCC	5.09	0.0236
LogLSGD (*L*=3)	5.84	0.0274
MFCC + LogLSGD	4.54	0.0213

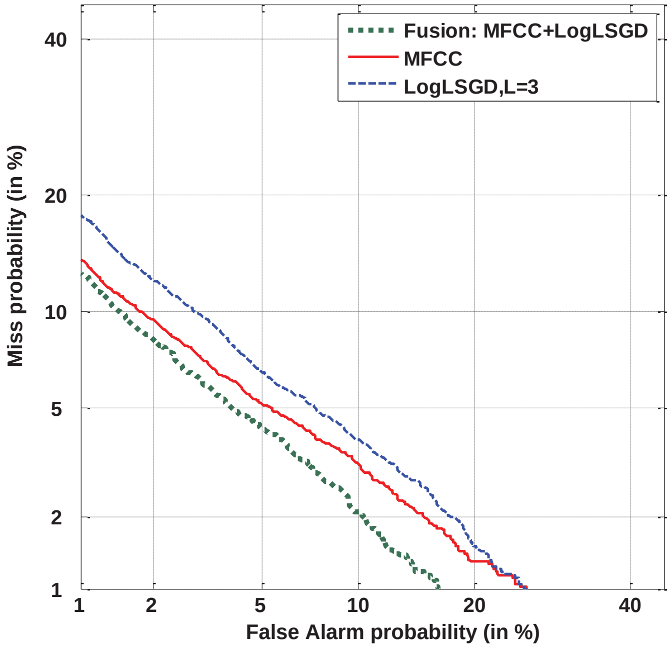


Figure 3.4 DET curves showing the MFCC, LogLSGD (L=3) and fused speaker recognition system, tested on NIST 2006 SRE core condition

3.2 Proposed Spectral Centroid Features

In the recent NIST 2008 SRE evaluation, the effectiveness of the frame-averaged FM components extracted using second order all pole method (discussed in section 2.3.2.2) on speaker recognition and its complementary nature to magnitude based information was demonstrated [22]. One problem with using the all-pole FM extraction in practical implementations is computational complexity [36], due to the need to model each subband FM component as a second order all-pole resonator [36]. However, a comparison between the feature values from all-pole FM extraction and the deviation of subband spectral centroid (SSC) features [35] from the center frequency of the subband, discussed in [76] and replicated in Figure 3.5, reveals that both SSC and all-pole FM features carry similar information. Notably, the calculation of SSC is more efficient (by a factor of 1.5) than the estimation of frame-averaged FM components.

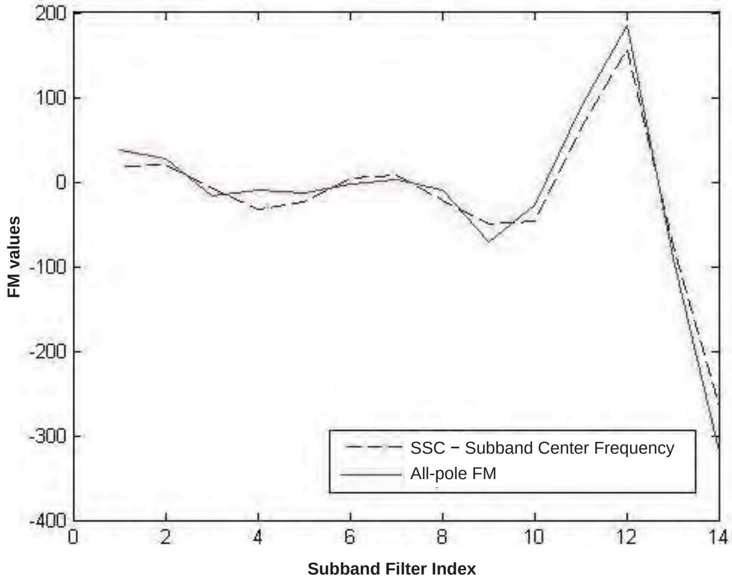


Figure 3.5 Frame-averaged Frequency Modulation, based on the all-pole method [26], compared with deviation of subband spectral centroid [35] from the center frequency of the subband for a frame of voiced speech signal

The subband spectral centroid (SSC) was originally proposed as a feature for speech recognition systems [35], and has been described as a formant feature, as it provides the approximate location of formant frequencies within the subbands [35]. However, this feature can be estimated easily and reliably, and in fixed dimension, unlike formant features [35]. Recently, SSC was also used for speaker recognition [155, 156] to complement cepstral based features with very slight success in contrast to FM features. Considering the similarity with frame-averaged FM, seen in Figure 3.5, however, the slight improvements over MFCC in speaker recognition applications seems something of an anomaly. In this section, the effectiveness of SSC is re-evaluated and an improved implementation of SSC demonstrated.

3.2.1 Spectral Centroid Feature Extraction

3.2.1.1 Spectral Centroid Frequency

As mentioned in section 2.3.2.2, spectral centroid frequency (SCF)[7] is the weighted average frequency for a given subband, where the weights are the normalised energy of each frequency component in that subband. The m^{th} subband spectral centroid frequency F_m is defined as follows [35]:

$$F_m = \frac{\sum_{k=l_m}^{u_m} k|S[k]w_m[k]|}{\sum_{k=l_m}^{u_m} |S[k]w_m[k]|} \tag{3.7}$$

where $S[k]$ represents the spectrum of a frame of speech, k denotes the discrete frequency value and $w_m[k]$ is the frequency-sampled frequency response of the m^{th} subband filter that is defined by a lower frequency edge (l_m) and an upper frequency edge (u_m). The final SCF vector corresponding to each frame is obtained by concatenating all the F_m values extracted from that frame. Figure 3.6 outlines the process of extracting the SCF features from a speech signal.

In the preliminary speaker recognition experiments, subband spectral centroid features based on mel-scaled triangular filters as proposed in [35] did not outperform second order all pole FM [26], achieving an EER around 2% poorer than FM. Since SCF is a frequency-based feature, we experimented with extracting SCF using a Bark scale Gabor filterbank which is motivated by the extraction of frequency modulation components in [26, 70]. In addition, we increased the number of FFT points by an order of magnitude (from 160 to 2048 for f_s = 8 kHz by zero-padding) to better approximate the speech power spectrum and filterbank frequency response, which was found to have an *absolute* improvement of 4% in terms of EER on the SCF performance.

[7] Spectral centroid frequency is commonly known as subband spectral centroid, however, we use the term spectral centroid frequency in order to avoid the ambiguity with spectral centroid magnitude, proposed herein.

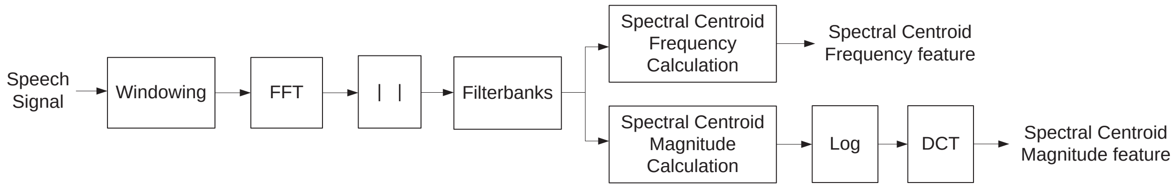


Figure 3.6 Proposed spectral centroid features extraction scheme

3.2.1.2 Proposed Spectral Centroid Magnitude

Looking at the expression of SCF in equation (3.7), an analytically complementary feature termed Spectral Centroid Magnitude (SCM) is proposed. The spectral centroid magnitude can be viewed as the weighted average magnitude for a given subband, where the weights are the frequency of each magnitude component in that subband as shown in equation (3.8). A feature vector is obtained by concatenating all M_m in that frame, then a logarithm is applied to reduce the dynamic range of the feature vector. The discrete cosine transform (DCT) is then applied to obtain the final SCM feature vector as shown in Figure 3.6. The use of the DCT is intended to decorrelate the feature vector, as it does when conventionally used in computing MFCCs.

$$M_m = \frac{\sum_{k=l_m}^{u_m} k|S[k]w_m[k]|}{\sum_{k=l_m}^{u_m} k} \qquad (3.8)$$

The SCM captures, to a first order approximation, the distribution of energy in a subband, as shown in Figure 3.7, for two arbitrary signals with the same average energy. Due to the weighting function, the two signals are each represented by different SCF and SCM values. The different steepness of the weighting function with respect to the subband bandwidth may also be noted; this results in different feature element variances (prior to feature warping). Average energy could be computed using equation (3.8) by simply setting k = 1 (i.e. MFCCs). As the spectral centroid magnitude is the magnitude at the position of the spectral centroid frequency, it is assumed to carry formant-related information which is useful for speaker recognition.

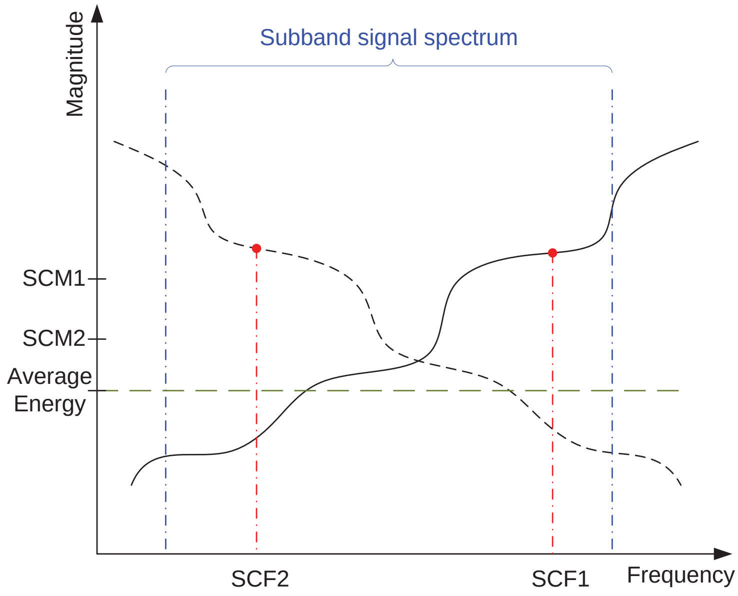


Figure 3.7 SCF and SCM extraction for two different example subband signals (solid (1) and dashed (2)) with equal average energy. Due to the SCM frequency weighting, $SCM_1 > SCM_2$.

However, in equation (3.8), the denominator is not speaker-dependent, unlike for the SCF. In order to increase the speaker dependency of the SCM, an alternative formulation, using only the P most significant frequency components within each subband can also be proposed, as follows:

$$M_{sc,m} = \frac{\sum_{k' \in I_m} k' |S[k'] w_m[k']|}{\sum_{k' \in I_m} k'} \qquad (3.9)$$

where I_m is a set of frequencies corresponding to the P largest values of $|S[k]w_m[k]|$. This alternative method of SCM will be referred to as the SCM based on significant components (SCM_SC) in this thesis. As shown in Figure 3.8, when SCM is plotted against SCF, it provides a better approximation to the LPC spectrum compared with average energy plotted against the center frequency of each subband. To confirm this result, the average MSEs of average energy, SCM and SCM_SC of 100 speakers (50 male and 50 female from NIST2001 SRE) were computed against the LPC spectrum. The resulting MSEs were 2.6 for the average energy and 2.48 for SCM. Hence suggesting that the combination of SCM and SCF carry more information than just average energy (i.e. MFCCs). The MSEs of SCM_SC for *P*= 3, 5 and 7 were 3.67, 3.69, 3.7 respectively.

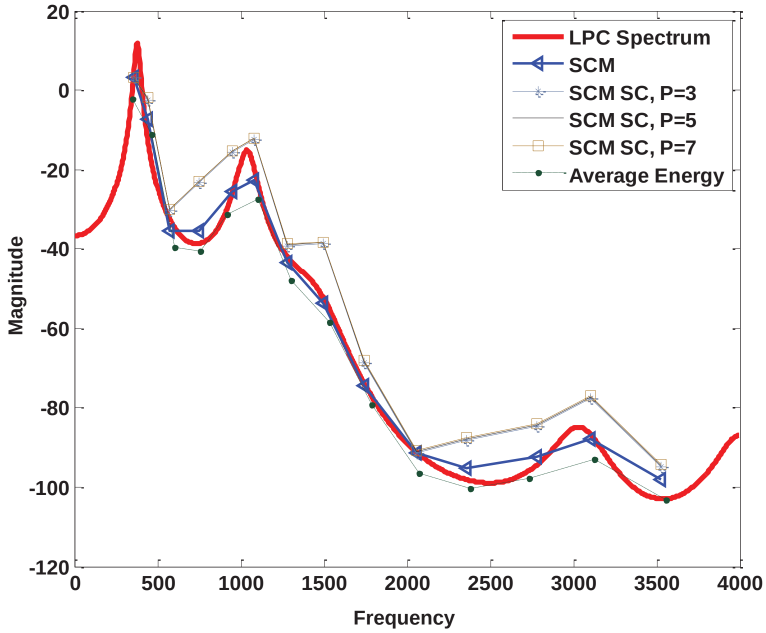


Figure 3.8 LPC spectrum, SCM vs SCF and Average energy vs subband center frequency, frame size = 20ms

3.2.2 Evaluation

In this section, due to its extensive nature, the NIST 2001 SRE database is employed for the initial part. Later the NIST 2006 SRE database was used to evaluate a selection of feature combinations (selected based on the initial investigations). The same baseline system based on MFCC and system configurations outlined in Section 3.1.3 are used for the following experiments.

3.2.2.1 Comparison of Normalization

Feature warping and cepstral mean subtraction (CMS) are commonly used feature normalisation techniques for magnitude based features [4]. As SCF is a frequency based feature we empirically studied its behaviour with both normalisation techniques. In these experiments 14 uniformly spaced Gabor filter banks across the bandwidth of 0.3 to 3.4 kHz were chosen to analyse the cellular telephone speech data in NIST 2001 SRE database. Table 3.3 shows the EERs from speaker recognition experiments on the NIST 2001 SRE database with both normalisation techniques.

These experiments verify that the feature warping is the better normalisation technique for SCF. Hereafter in all subsequent experiments, feature warping was used as the feature normalisation for both SCF and SCM.

Table 3.3: The speaker recognition results for spectral centroid frequency with various normalisation approaches on the NIST 2001 SRE database

Normalisation techniques	EER (%)
No normalisation	12.17
Cepstral Mean Subtraction (CMS)	10.11
Feature warping	9.47

3.2.2.2 Comparison of SCF and FM

After comparing the normalisation techniques in the previous section, we investigate the effects of different frequency scales and filterbanks on SCF. In these comparative experiments three different frequency scales: Bark, uniform and mel scales, and two different filter shapes: Gabor and triangular were chosen. Uniform-scaled and Bark-scaled Gabor were chosen for comparative studies between SCF and FM, since SCF and FM carry similar information as discussed above, and also because of the significant improvement of the uniform scale over Bark scale observed for FM features in [157]. In all these experiments, the number of filters was fixed at 14. Results for speaker recognition experiments based on the NIST 2001 SRE database are given in Table 3.4.

Table 3.4: The speaker recognition results for spectral centroid *frequency* and all-pole FM with various frequency scales and filterbanks on the NIST 2001 SRE database

Filterbank	SCF EER (%)	All-pole FM EER **[157]** (%)
Mel Scale Triangular	11.19 [35]	-
Mel Scale Gabor	**8.83**	11.04
Bark Scale Gabor	9.42	12.71
Uniform Scale Gabor	12.17	10.45

The results indicate that the Gabor filterbank with a mel-scale produces the best results for SCF, outperforming the mel-scale triangular filterbank SCF implementation proposed in [35]. One reason might be that SCF is a frequency based feature and previously for another frequency based feature, FM feature, the Gabor filter bank was chosen for its optimum time, frequency sensitivity and the absence of large side lobes [70]. In addition, Bark scale filters performed slightly better than uniform scale filters for SCF in contrast to the results in [157] for FM features on NIST 2001 database.

Furthermore, it was also reported in [157] that the auditory motivated scales are not an optimal scale for designing a speaker verification scale based on FM, as it does not take into account the non-uniform distribution of speaker discriminative information across the spectrum as discussed in section 2.1.1. Motivated by Thiruvaran et al. [157], a similar approach based on Kullback-Leibler distance is taken to design an optimal filter for SCF, and a relative improvement of 4% in EER (on NIST 2001 database) was observed. The details of the design process along with recognition performance are reported in Appendix B.

3.2.2.3 Comparison of Filterbanks for SCM

The same filterbank configuration mentioned in Section 3.2.2.2 was used for SCM extraction. In all the experiments in this thesis, the number of SCM DCT coefficients were fixed at 14. Results for speaker recognition experiments based on the NIST 2001 SRE database for SCM are given in Table 3.5. For SCM, mel-scale triangular filters performed best among our comparisons. This result is perhaps expected since MFCCs also employ triangular mel-scale filters, and SCM is equivalent to a frequency-weighted MFCC feature.

Table 3.5: The speaker recognition results for spectral centroid *magnitude* with various frequency scales and filterbanks on the NIST 2001 SRE database

Filterbank	SCM EER (%)
Mel Scale Triangular	**8.88**
Mel Scale Gabor	9.12
Bark Scale Gabor	9.53
Uniform Scale Gabor	9.62

3.2.2.4 Combination of SCM and SCF

In this section, we investigate the effectiveness of the combination of SCF and SCM features for speaker recognition. First, SCM and SCF were combined using score level fusion with results as given in Table 3.6. Linear fusion was used, with weights calculated using the same NIST 2001 database. The fusion can thus be considered optimum. When the filter banks were fixed to the same shape and scale, the best fused results were obtained with mel scale triangular filters. Keeping the same filterbanks for both SCF and SCM is preferred as it reduces the computational complexity significantly. It could be observed from equations (3.7) and (3.8) that only the denominator of equation (3.7) and (3.8) differs when using the same filter banks for both SCM and SCF. This is one advantage of using the combination of the (FFT-based) SCM and SCF over the alternative feature combination of MFCC and FM, where FM extraction occurs in the time domain and is very computationally demanding. System performance was further improved by fusing the best SCF and best SCM features, as shown on Table 3.6.

Though score level fusion is usually used to combine different subsystems, in our case as both features are extracted using the same filterbanks, feature level concatenation is a reasonable alternative. The advantage of feature level concatenation over score level fusion is that a development database is not required, while score level fusion is biased by the choice of development data for computing the fusion weights. Although the performance of concatenation is slightly less than that of fusion, it should be noted that the fusion is the optimal fusion trained using the same evaluation database.

It can be observed that both fused and concatenated SCF + SCM systems perform better than the baseline MFCC system (EER = 8.49%), with an increment in feature dimension from 32 (16 MFCCs + 16 Δs) to 56 (14 SCFs + 14 Δs + 14 SCMs + 14 Δs).

Table 3.6: Score level fusion and feature concatenation of SCM and SCF speaker recognition performance on the NIST 2001 SRE database

Features	EER (%)
MFCC	8.49
SCF (Mel Scale Gabor) + SCM (Mel Scale Gabor)	8.05
SCF (Bark Scale Gabor) + SCM (Bark Scale Gabor)	8.43
SCF (Mel Scale Triangular) + SCM (Mel Scale Triangular)	7.99
SCF (Mel Scale Gabor) + SCM (Mel Scale Triangular)	7.90
Feature concatenation of SCF (Mel Scale Gabor) and SCM (Mel Scale Gabor)	8.19

3.2.2.5 SCM based on significant components

In this section, the alternative expression of equation (3.8) to calculate SCM based on significant components is briefly explored. As shown on Table 3.7, the EER for SCM based on significant components did not outperform SCM for which all frequency components are taken into consideration and hence SCM_SC was not used in subsequent experiments. On the other hand, the performance of SCM_SC is close to that of SCM even when we use just a few frequency components.

Table 3.7: The speaker recognition performance for SCM based on significant components (SCM_SC) on the NIST 2001 SRE database

Features	Number of significant components (P)	EER (%)
SCM	-	8.88
SCM_SC	3	9.57
SCM_SC	5	9.51
SCM_SC	7	9.08

3.2.2.6 SCF and SCM performance for NIST2006 SRE (1conv4w-1conv4w)

Finally, the combination of SCF and SCM was evaluated using the larger and more contemporary NIST 2006 database, in order to see the database independency of the results. Based on the results in Section 3.2.2.4, the fusion of mel-scale Gabor SCF and mel-scale triangular SCM gave the best result. However the mel-scale triangular SCF and mel-scale triangular SCM can be implemented more efficiently since they share the same filterbank, at the cost of a slightly higher EER. Hence SCF and SCM extracted with mel-scale triangular filterbank along with SCF extracted with mel-scale Gabor filters were all selected for comparisons with MFCC when tested on the NIST 2006 database. The performance of SCM and SCF when used alone is given in Table 3.8, together with the MFCC baseline, and the fusion results are given in Table 3.9.

It can be observed that SCF extracted using Gabor filters performed significantly better than SCF extracted using triangular filters as proposed in [35] or all-pole based FM [157]. In addition, triangular filter extracted SCF performs worse than all-pole based FM as found in earlier results.

Table 3.8: Speaker recognition results for spectral centroid features on the NIST 2006 SRE database with speaker detection cost model parameters of $C_{Miss} = 10$, $C_{FalseAlarm} = 1$, $P_{Target} = 0.01$

Features	EER (%)	minDCF
Mel Scale Triangular SCM	5.40	0.0238
Mel Scale Triangular SCF	9.23	0.0380
Mel Scale Gabor SCF	6.45	0.0291
MFCC	5.09	0.0236
FM	7.01	0.0302

Table 3.9: Fused speaker recognition results for spectral centroid features on the NIST 2006 SRE database with speaker detection cost model parameters of C_{Miss} = 10, $C_{FalseAlarm}$ = 1, P_{Target} = 0.01

Features	EER (%)	minDCF
Mel Scale Triangular SCM + Mel Scale Triangular SCF	4.82	0.0235
Mel Scale Triangular SCM + Mel Scale Gabor SCF	4.40	0.0217
MFCC + Mel Scale Triangular SCM	4.26	0.0212
MFCC + Mel Scale Gabor SCF	4.31	0.0216
MFCC + FM	4.57	0.0213
MFCC + Mel Scale Triangular SCM + Mel Scale Triangular SCF	4.13	0.0213
MFCC + Mel Scale Triangular SCM + Mel Scale Gabor SCF	**3.73**	**0.0200**

Interestingly the fusion of MFCC with SCM extracted using mel scale triangular filters gave substantial improvement over the individual subsystems, which was not expected. This might be attributed partly to the different number of filters used for MFCC (26 filters) and SCM (14 filters) and partly to the different extraction methods that is, MFCC is based on average energy while SCM is based on weighted average energy.

Results from this experiment showed that the improvements discussed in Section 3.2.2.4 (fusion of SCM + SCF outperforms MFCC) were also found for the more contemporary NIST2006 database, where SCM and SCF improved on a 5.09% EER MFCC baseline to 4.4% after fusion as shown in Table 3.8 and Table 3.9. When the best performing SCM and SCF (extracted using mel-scale triangular and mel-scale Gabor filterbanks respectively) were further fused with MFCC, the EER dropped to 3.73% as shown in Figure 3.9. These results provide strong encouragement that SCM and SCF carry complementary information to MFCCs.

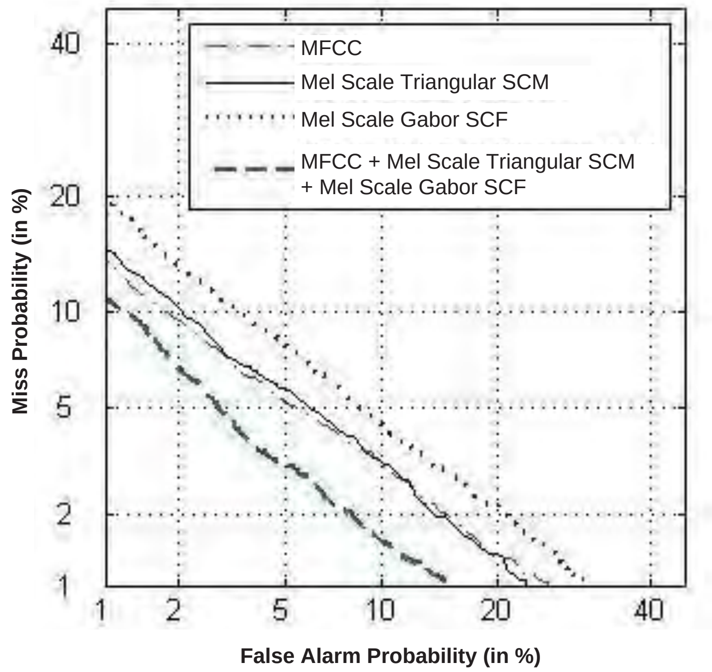


Figure 3.9 DET curves showing the speaker recognition results of MFCC and spectral centroid features on the NIST 2006 SRE database

3.3 Summary

This chapter firstly presented a technique based on least squares regularisation as an alternative complementary feature extraction method to reduce the variability of group delay (GD) features derived from the speech spectrum. The proposed method involved rewriting the GD extraction expression in a matrix-vector notation over a window of pre-defined length to achieve a smoothing effect. Interestingly, the proposed log compressed least squares group delay (LogLSGD) feature successfully reduces the dynamic range while retaining the fine structure of the group delay. Experimental results (Table 3.1) indicate that the proposed LogLSGD features alleviate the ill-conditioning of the MODGDF calculation due to strong excitation components and the need to determine any data-dependent empirical parameters in the GD feature extraction algorithm.

An alternative centroid feature extraction method for subband magnitude-based and frequency-based features, termed spectral centroid magnitude (SCM) and spectral centroid frequency (SCF) respectively from the speech spectrum were proposed. Evaluation on the NIST 2006 database using a fusion of SCM-based and SCF-based subsystems, demonstrated relative improvements of 13% over the performance of an MFCC-only system (shown in Table 3.9). This result demonstrated that the combination of SCM and SCF carries more information than MFCC alone. SCF was also shown to perform significantly better than the previously proposed subband spectral centroid and frame-averaged FM features for speaker recognition.

Chapter 4

Investigation of Front-end Diversity in Speaker Recognition Systems

The previous chapter has shown that the fusion of systems based on new features to complement MFCC-based systems can advance the performance of the baseline (MFCC-based) system. Furthermore, many speaker recognition researchers, have been motivated to investigate features derived from different sources of information in speech (e.g. frequency, phase, modulation energy), with the assumption that systems built on these features will model different aspects of the speaker voices (resulting in different speaker modelling[8]) and eventually will fuse well with MFCCs [24-26, 76, 81, 158-160]. The fusion of diverse speaker recognition systems with different front-end features is commonly referred to as classifier ensembles [37] in general terms (alternatively known as feature set diversity [141] or front-end diversity in this thesis).

The SCM, which is based on the same information as MFCC (as discussed in section 3.3.2.6), was found to fuse well with MFCCs. Considering the similarity between MFCC and SCM, the significant improvements over MFCC only systems seems something of an anomaly. This leads us to the hypothesis that front-end diversity is instead achieved through different 'partitioning' of the acoustic space (acoustic modelling rather than speaker modelling).

[8] Acoustic and speaker modelling refers to the GMM-UBM training and MAP adaptation in this chapter respectively. The UBM covers the space of speaker-independent, broad acoustic classes of speech sounds, while adaptation is the speaker-dependent tuning of those acoustic classes based on features observed in the speaker's training speech [8].

This chapter looks at two sets of experiments that attempt to test our hypothesis to some extent. In section 4.2, motivated by SCM, a range of MFCC-variant features (features that carries similar information to conventional MFCC) are introduced and fused with the baseline MFCC system to determine whether MFCC-variant features do *generally* carry complementary properties to MFCCs. These experiments may give some insight into the common assumption that only features derived from different sources of information in speech will fuse well with MFCCs [24, 159-162]. Then in section 4.3, we introduce a novel way to separately investigate the acoustic and speaker modelling 'stages' of the GMM-UBM based systems, towards determining the contributions of each stage to the speaker recognition performance across different features.

4.1 Feature-based Approaches to Front-end Diversity

In this section, some possible variations to the extraction of MFCCs that produce diversity with respect to fused subsystems based on different MFCC-variant features are investigated. The variations have been chosen to produce minimal/no additional speaker-related information with respect to MFCCs. The assumption is thus that any improvement observed in the speaker recognition performance may be related to the differences in acoustic modelling. Figure 4.1 shows a modular representation of the computation of MFCCs, with some possible variations to achieve diversity. The weighting block, which does not appear in the conventional MFCCs computation, is incorporated here to permit the computation of SCM discussed in section 3.2.1.2. In the mel scale filterbank block, the number and type of filterbanks are investigated, as are the dependencies between frequency components and the individual characteristics for speaker recognition through band selection. In the Discrete Cosine Transform (DCT) block, a drop-one-out cepstral experiment is conducted.

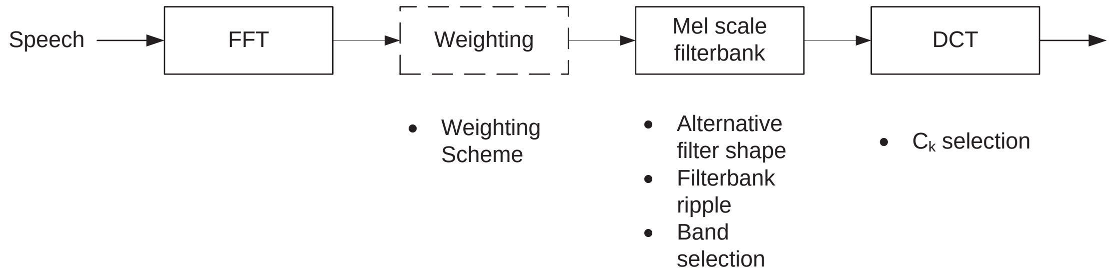


Figure 4.1 Block diagram of the MFCC extraction process, with an additional weighting stage and approaches to MFCC front-end diversity, for investigating if MFCC-variant features do *generally* carry complementary properties to MFCCs in this section listed below the relevant blocks

4.1.1 Subband Energy Weighting

The weighting scheme of SCM, $\mathbb{w}_k = f$, was generalised to study the effect of incorporating a weighting function on the subband energy with respect to the MFCCs and to allow alternative weightings as shown in equations (4.1) and (4.2). The Anti-SCM (ASCM), whose weights are in the reverse order of those for the SCM (where higher frequencies within the subband are weighted more than lower frequencies) was proposed to evaluate the effects of different weighting schemes on the computation of MFCCs. The various weighting functions are illustrated in Figure 4.2.

$$M_m = \frac{\sum_{k=l_m}^{u_m} \mathbb{w}_m |S[k] w_m[f]|}{\sum_{k=l_k}^{u_k} \mathbb{w}_m} \tag{4.1}$$

$$\mathbb{w}_m = \begin{cases} k & SCM \\ 1 & MFCC \\ -k + (l_m + u_m) & ASCM \end{cases} \tag{4.2}$$

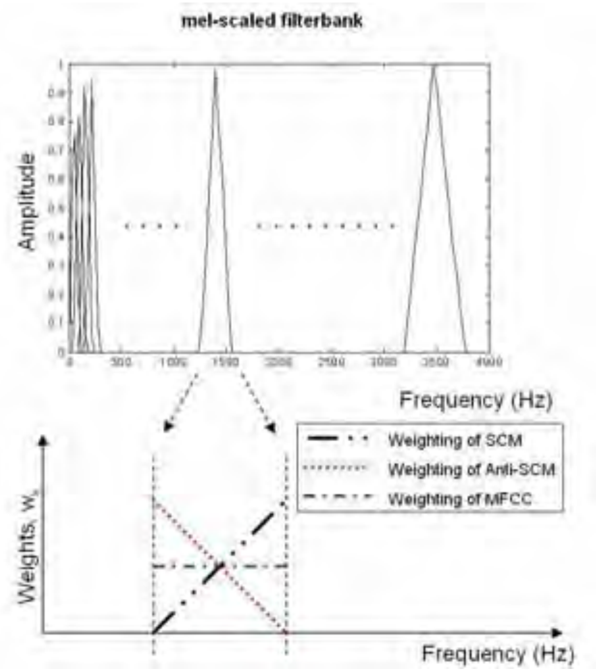


Figure 4.2 Differences in weighting schemes, w_k, between MFCC, SCM and ASCM

The speaker recognition results based on the NIST 2001 SRE are given in Table 4.1. For this and following experiments, the systems are based on 14 dimensional DCT coefficients with appended delta coefficients. The features were extracted every 10 ms, using Hamming analysis window of 20 ms and a filterbank of 26 triangular mel-spaced filters (unless otherwise stated). The features were also normalised using feature warping. It can be observed that by altering the weighting function in each band, the individual system performance does not improve upon that of MFCCs. However the fused results show a slight improvement, which could be due to the difference in steepness of the weighting function with respect to the subband bandwidth causing different feature element distributions.

Table 4.1: The speaker recognition results for MFCCs with different weightings on the NIST 2001 SRE database

Features	EER (%)	Fused EER (%) with MFCCs
MFCC	8.78	-
SCM	9.86	8.68
Anti-SCM	9.72	8.57

4.1.2 Filterbank

The effects of different filterbanks on the computation of cepstral coefficients were investigated, while keeping the frequency scale fixed to the mel scale. In these comparative experiments, three different filter shapes: triangular, Gabor and gammatone and two different numbers of filters: 26 and 14 were chosen. Gabor-shaped filters and 14 filters were chosen for comparative studies between MFCC and SCM (discussed in section 3.2.2.3) and gammatone was chosen here as an alternative auditory motivated filterbank (26 triangular filters is used for the standard MFCC computation). Results for speaker recognition experiments based on the NIST 2001 SRE database are given in Table 4.2.

Table 4.2: The speaker recognition results for cepstral coefficients with different filterbanks on the NIST 2001 SRE database

Filterbank	EER (%) for number of filters	
	14	26
Triangular	9.13	8.78
Gabor	9.08	9.27
Gammatone	9.76	8.82

According to the results, the standard 26 mel-scaled triangular filterbank produced the best results. This result is perhaps expected since the total magnitude response across all triangular filters with 50% overlap is an exactly uniform magnitude spectrum as shown in Figure 4.3, so the filterbank energies from each band are not attenuated. As a slight digression, in order to investigate the effect of filterbank ripple in feature extraction, a set of experiments was devised in which the number of filters was fixed at 26 and the total filterbank maximum passband ripple was adjusted by tuning the overlap ratio between neighbouring bands. As shown in Figure 4.4, the larger the ripple in the passband, the higher the EER. This result is significant for two reasons: (i) it helps to explain the general success of MFCCs as features; and (ii) alternative features often tend to employ non-triangular filter banks, where minimising ripple may not always be given the priority it deserves. Nonetheless, the purpose of examining filterbank configurations was to see whether slight modifications of the MFCC extraction process could produce systems that fused well with the baseline.

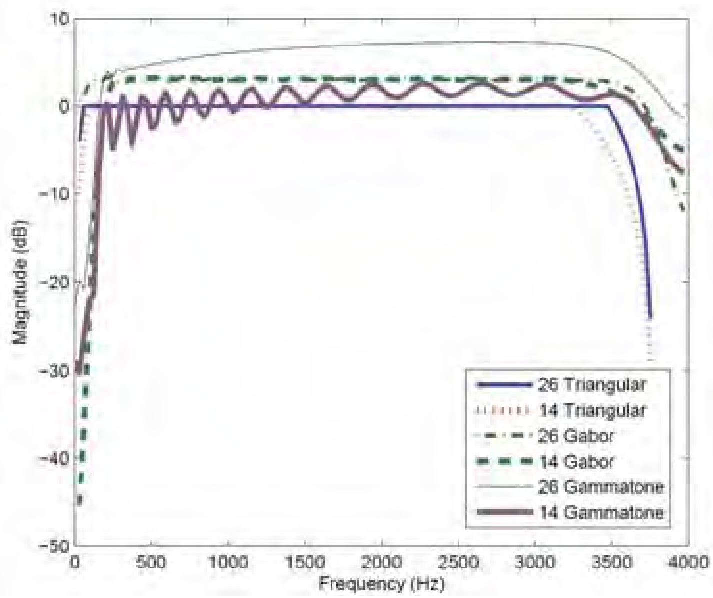


Figure 4.3 Total magnitude response of various filterbanks

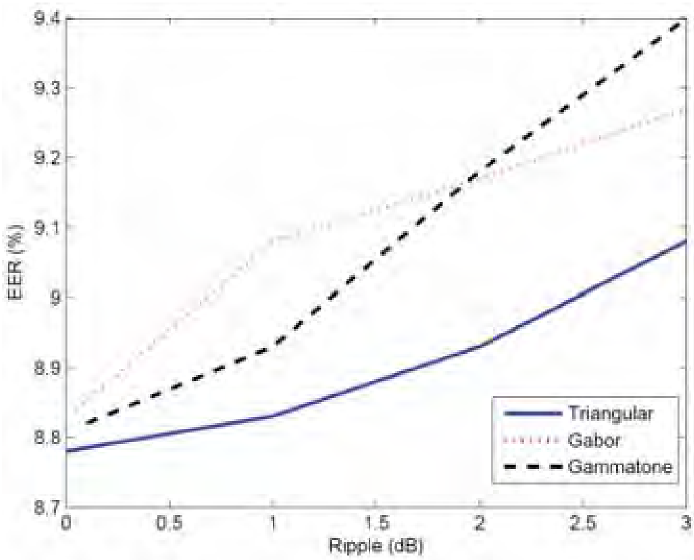


Figure 4.4 EER vs passband ripple

Next, the combination of systems employing cepstral coefficients computed using different filterbanks was investigated. It can be observed from the results in Table 4.3 that fusion of cepstral coefficients extracted using different filterbanks improves on the individual subsystems. This might be attributed to the different number and type of filters used which 'partition' the acoustic space in slightly different ways.

Table 4.3: Fused EER of different filterbanks for speaker recognition on the NIST 2001 SRE database

Filterbank	EER (%)	Fused EER (%)				
		14 Triangular	**26 Gabor**	**14 Gabor**	**26 Gammatone**	**14 Gammatone**
26 Triangular	8.78	8.41	8.68	**8.14**	**8.19**	8.56
14 Triangular	9.13	-	8.73	8.53	8.24	8.83
26 Gabor	9.27	-	-	8.53	8.36	8.68
14 Gabor	9.08	-	-	-	8.61	8.93
26 Gammatone	8.82	-	-	-	-	8.82
14 Gammatone	9.76	-	-	-	-	-

4.1.3 Band Selection

A series of speaker recognition experiments were conducted by leaving out one band of the filterbank energies at a time (in each experiment only 25 filters out of 26 are used), which will be termed as the drop-one-band system, before applying the DCT. These results, generated for NIST2001, indicate how speaker-specific information (discussed in section 2.1.1) is distributed along different bands (Figure 4.5). These results are generally similar to the observations by Lu et al [85] based on the analyses using the Fisher F-ratio to determine the dependencies between frequency components and individual speaker characteristics on filterbank energies.

Observations from a comparison of the distribution of speaker-specific information on telephone channel (telephone bandwidth of 300 Hz to 3700 Hz) obtained for subband energy (i.e MFCCs) in [85] with Figure 4.5 are: (1) both reveal that the contribution to speaker recognition is high in the area around 500 Hz and drops beyond this to a minimum around 1100 Hz; (2) the contribution increases significantly between 2500 Hz to 2800 Hz and drops after 2800 Hz. In addition, comparing Figure 4.5 and Figure B-1 (in Appendix B), it can be observed that the trend is almost similar across two different features, MFCC versus SCF, except for frequencies approximately below 500Hz. This could be due to frequencies being around the edge of the telephone bandwidth [157].

As in 4.1.1 and 4.1.2, the conventional MFCC system was fused with the drop-one-band system. Results showed slight improvements over the MFCC baseline, with a similar trend to the EER curve in Figure 4.5. Interestingly 23 out of the 26 fused systems showed an improvement of 0.03% - 0.35% in EER despite the fused systems carrying essentially the same information as shown in Figure 4.6.

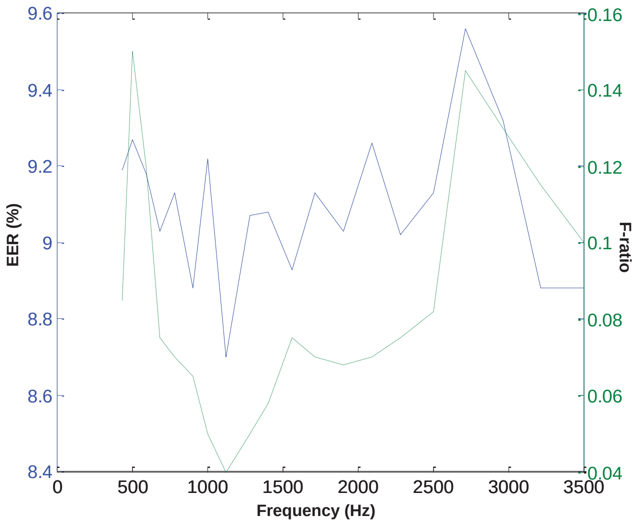


Figure 4.5 EER (left y-axis, solid line) in a series of 'leave-one-out experiments' and F-ratio (right y-axis, dash-dot line) (after [85]) using MFCCs, for the NIST2001. Higher EER indicates that valuable speaker-specific information is contained in the respective dropped frequency band.

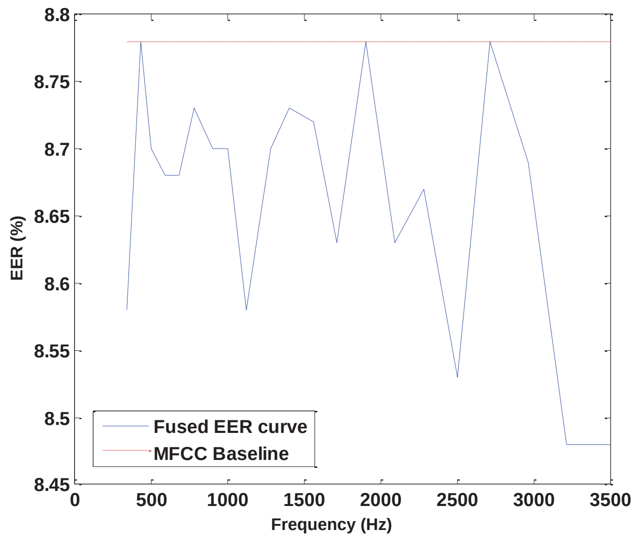


Figure 4.6 Fused EER of drop-one-band system (for bands centred at frequencies as shown on x-axis) with MFCC baseline system for speaker recognition on the NIST 2001 SRE database

4.1.4 Cepstral Coefficient Selection

Motivated by ensemble classifier methods, drop-one-out cepstral coefficient experiments were conducted and the resulting EERs can be seen in Table 4.4. Each result does not outperform the conventional MFCC EER of 8.78% since the representation of the filterbank energies has been altered, and less information is available to the classifier. The fused results of conventional MFCCs and the drop-one-out system are shown in Table 4.4 and all the possible combinations of two different drop-one-out cepstral systems as shown in Figure 4.7 suggest that system performance can be further improved by fusing systems of similar features whereby no additional information of the speech is included, consistent with the observation in section 4.1.3. The fusion of a system based on all 14 cepstral coefficients with a system based on the last 13 coefficients improved the MFCC baseline to 8.34% and when further fused with all the 14 different drop-one-out MFCC elements system, the EER dropped to 8.19%.

Table 4.4: Speaker recognition results for drop-one-out MFCC elements on the NIST 2001 SRE database

Dropped Cepstral Coefficient	EER (%)	EER fused with MFCC (%)	Dropped Cepstral Coefficient	EER (%)	EER fused with MFCC (%)
1	8.93	**8.34**	8	9.22	8.63
2	9.37	8.78	9	9.14	8.62
3	8.78	8.46	10	8.97	8.68
4	9.27	8.68	11	8.88	8.67
5	9.52	8.78	12	9.13	8.63
6	9.14	8.68	13	9.08	8.69
7	9.52	8.68	14	8.87	8.68

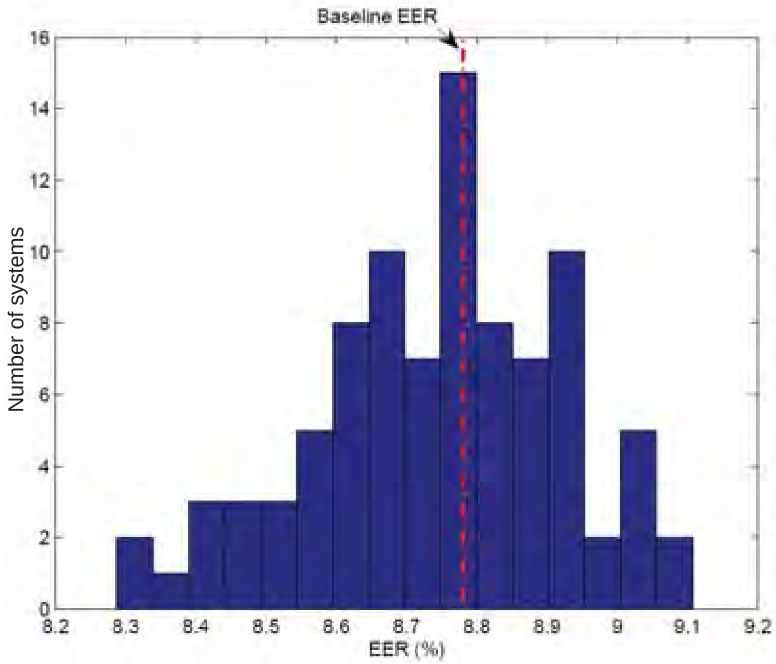


Figure 4.7 EER histogram of all possible combination of two drop-one-out cepstral systems

4.1.5 Speaker Recognition Performance on NIST 2006 SRE

Finally, a selected set of feature combinations from the previous sub-sections was evaluated on the core condition (1con4w-1con4w) of NIST 2006 SRE database, in order to confirm that the foregoing improvements hold for a more typical, contemporary speaker recognition system applied to a larger database. The back-end was based on a GMM-SVM classifier with configurations as discussed in section 3.1.3.

The performance of the selected features when used alone (leftmost column) and all the possible combinations (remaining columns) of fused systems is given in Table 4.5. It can be observed that the individual performance is consistent with the experimental results in section 4.1 in that the modified MFCCs alone do not outperform the MFCC baseline. However all fused systems do improve the performance of the individual subsystems. Interestingly, the fusion of MFCCs extracted using 26 Triangular and 26 Gammatone filterbanks attained an EER of 4.21% (Figure 4.8), which is similar to the fused results of MFCC and 14 mel-scaled Gabor filterbank-extracted spectral centroid frequency (SCF) with an EER of 4.31% (same database and back-end configuration) in section 3.2.2.6.

Table 4.5: Speaker recognition results for MFCC variant features on the NIST 2006 SRE database

Features	EER (%)	Fused EER (%)				
		SCM	26 Tri MFCC	14 Gab MFCC	26 Gam MFCC	DC 1
ASCM	5.87	4.75	4.35	5.09	5.14	5.87
SCM	6.25	-	4.48	4.48	4.78	4.90
26 Tri MFCC	5.09**	-	-	4.25	**4.21**	4.51
14 Gab MFCC	5.76	-	-	-	5.25	5.32
26 Gam MFCC	5.62	-	-	-	-	5.30
DC 1	6.20	-	-	-	-	-

* Tri: Triangular, Gab: Gabor, Gam: Gammatone, DC: Dropped Cepstral
** MFCC Baseline

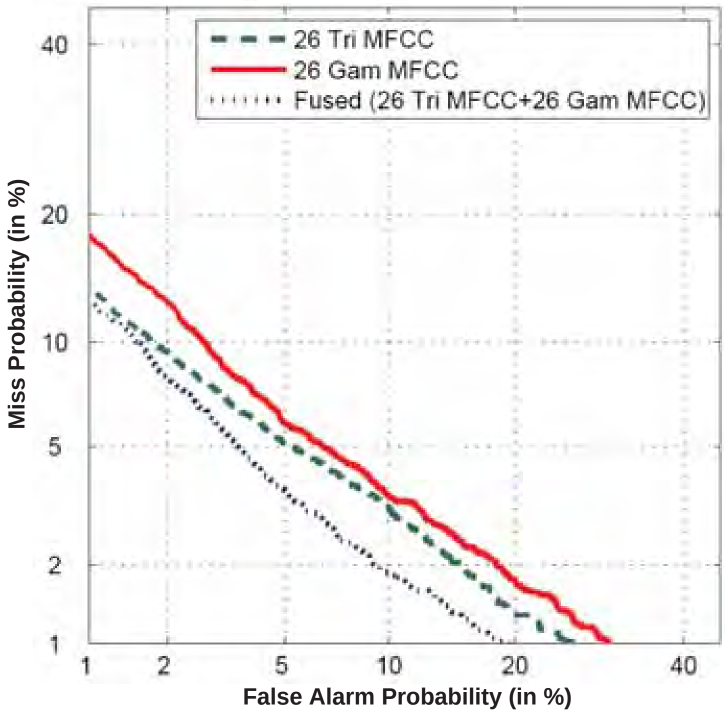


Figure 4.8 DET curves showing the speaker recognition results on the NIST 2006 SRE database

Studies in the literature [76] have shown the advantage of frequency-based features as being complementary to MFCCs for the purpose of fusion in speaker recognition systems. However here we were surprised to find that the fused modified MFCC variants, which carry essentially the same information as MFCCs, were able to achieve comparable performance to the fusion of MFCC and frequency-based features. In addition, we experimented with the fusion of possible combinations of three different subsystems and the EER was further improved to 3.98% (26 Triangular MFCC + 14 Gabor MFCC + SCM), representing a 22% relative reduction in error rate over the MFCC baseline. A comparison of the three subsystems fused performance with those given in Table 3.9, again indicates that the fusion of MFCC variant features are able to achieve comparable performance to the fusion of magnitude- and frequency-based features.

4.2 Clustering-based Approaches to Front-end Diversity

In the previous section, it can be observed from the results that the fusion of MFCC and MFCC-variant features (suboptimal systems based on features comprising essentially the same information as MFCCs) consistently outperforms an individual MFCC based system, and that for particular design choices, the improvement can be significant. This shows that it is not essential for the features to be extracted from different sources of information in speech to carry complementary information. These results support the hypothesis that diversity is achieved mainly through different ’partitioning’ of the acoustic space.

In addition, it can be observed that features proposed as complementary to MFCCs usually are unable to outperform MFCCs in stand-alone systems, and one hypothesis we propose is that their ability to describe the acoustic space of a speaker is poorer than that of MFCC. Based on our hypothesis and motivated by Loquendo/Politecnico di Torino’s Phonetic GMM [94] (discussed in section 2.4.1), we attempted to separate acoustic modelling (UBM) and speaker modelling (MAP adaptation) by training a UBM based on

MFCC feature and then adapting the speaker model using alternative features (i.e. group delay, SCF, etc…) in order to utilise the 'well-trained' UBM model for MAP adaptation. However experiments conducted have so far been unsuccessful[9] and suggest that phonetic and speaker modelling must be based on the same feature in the current GMM paradigm. Next, motivated by the fact that GMM UBMs are outcomes of clusterings (e.g. k-means clusterings), we utilise the clustering comparison measures to 'separately' investigate the acoustic and speaker modelling.

4.2.1 Clustering Comparison Measures

Over the past few decades, along with the proposal of new clustering algorithms, there has been concerted interest in the development of effective measures for clusterings comparison. There are currently three main types of clustering comparison methods: pair-counting based measures [163], set-matching based measures [164] and information theoretic based measures [165].

In this section, an information-theoretic based measure for clustering comparison, the Normalised Information Distance (NID) [166] is employed. Compared with other measures, the NID is normalised, with a range of [0, 1], has a strong mathematical foundation and has the unique advantage of being a metric [166]. These properties facilitate comparison across different features or different data subsets.

4.2.2 Normalised Information Distance

NID is a measure of disagreement between alternative data (hard) partitions that can be used even when considering partitions with different numbers of clusters. Let **S** be a set of N data points $\{s_1, s_2, \dots, s_N\}$. Given two clusterings, $\mathbf{U} = \{U_1, U_2, \dots, U_R\}$ with R clusters,

[9] It has been observed that any attempt at hard acoustic partitioning caused a significant EER degradation. Furthermore, partitioning based on phonetic transcription is difficult because databases like TIMIT [122] are unsuitable for serious speaker recognition investigation.

and $\mathbf{V} = \{V_1, V_2, \dots, V_C\}$ with C clusters $\left(\bigcap_{i=1}^{R} U_i = \bigcap_{j=1}^{C} V_j = \emptyset \text{ and } \bigcup_{i=1}^{R} U_i = \bigcup_{j=1}^{C} V_j = S\right)$, the cluster label for each data point in **S** is computed and are represented by a string of symbols. For example:

$$\begin{aligned} \mathbf{S}_{\mathrm{U}} &= \{U_1, U_7, U_2, U_R, U_1, \dots\} \\ \mathbf{S}_{\mathrm{V}} &= \{V_5, V_3, V_C, V_1, V_2, \dots\} \end{aligned} \tag{4.3}$$

which means that the first data point s_1 belongs to the cluster labelled "U_1" in the clustering **U** whereas in the clustering "**V**" it belongs to the cluster labelled "V_5" and so on. In brief, $\mathbf{S_U}$ and $\mathbf{S_V}$ will contain the cluster labels of the corresponding data points in **S** with respect to clustering **U** and **V** respectively. Then the information on cluster overlap between $\mathbf{S_U}$ and $\mathbf{S_V}$ can be summarised in the form of an $R\times C$ contingency table $M = \left[n_{ij}\right]_{j=1\dots C}^{i=1\dots R}$ where n_{ij} denotes the number of data points that are common to clusters U_i and V_j as illustrated in Table 4.6. The normalised information distance is calculated as [166, 167]

$$NID(U,V) = 1 - \frac{MI(U,V)}{\max\{H(U), H(V)\}} \tag{4.4}$$

where mutual information (MI) and entropy (H) are defined as

$$MI(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{R}\sum_{j=1}^{C} P(i,j)\log\frac{P(i,j)}{P(i)P'(j)} \tag{4.5}$$

$$H(\mathbf{U}) = -\sum_{i=1}^{R} P(i) log P(i) \tag{4.6}$$

$$H(\mathbf{V}) = -\sum_{j=1}^{C} P'(j) log P'(j) \tag{4.7}$$

The $P(i,j)$ denotes the probability that a point belongs to cluster U_i in **U** and cluster V_j in **V**:

$$P(i,j) = \frac{|U_i \cap V_j|}{N} = \frac{n_{ij}}{N} \tag{4.8}$$

$$P(i) = \frac{|U_i|}{N} = \frac{a_i}{N} \qquad\qquad P'(j) = \frac{|V_j|}{N} = \frac{b_j}{N} \tag{4.9}$$

where $|X|$ denotes the number of data points in X. For NID, a larger value indicates discordance between the two clusters.

Table 4.6: The contingency table, $\boldsymbol{n_{ij}} = |\boldsymbol{U_i} \cap \boldsymbol{V_j}|$ represents the number of data points that are common to clusters U_i and V_j .

U\V	V_1	V_2	...	V_C	*Sums*
U_1	n_{11}	n_{12}	...	n_{1C}	a_1
U_2	n_{21}	n_{22}	...	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	...	n_{RC}	a_R
Sums	b_1	b_2	...	b_C	$\Sigma_{ij}\, n_{ij} = N$

NID for Gaussian Mixture Model

As discussed in section 2.4.1, clustering using Gaussian mixture models (the clustering method used in this thesis) is considered a soft clustering/assignment method [168], since the posterior probabilities for each point indicate that every data point has some probability of belonging to each Gaussian mixture component (equation (2.15)). In the clustering comparison literature, although clustering comparison methods for soft clustering algorithms are widely available, the final clustering results are often converted to hard clustering for ease of interpretation and comparisons [169] or alternatively the data are first partitioned into disjoint groups (hard partitioning) before the clusters are compared. In order to be in line with the clustering comparison literature, and to use the desirable properties of the NID, the latter option of converting all soft partitioning to hard partitioning is employed herein. Consequently, after the Gaussian mixture model parameters have been estimated, the assignment of data points to clusters[10] is achieved by assigning each data point to the mixture component which it has the highest posterior probability of belonging (hard assignment). This step will result in a list of mixture component indices as the cluster labels as shown in equation (4.3). Thereafter, the NID for GMM can be computed using equation (4.4).

4.2.3 Investigation of Fused Acoustic Features

Although it has been shown in section 4.1 that diversity is achieved mainly through different 'partitioning' of the acoustic space through the fusion of similar features, the effect on the speaker recognition performance with respect to the amount of difference between the acoustics clustering of two different features (e.g. fusion of magnitude-based and phase-based features) hasn't been quantified. In this section, we investigate if the amount of difference in terms of acoustic clustering between two different features is correlated with the fused speaker recognition performance.

The investigation was conducted by measuring the NID between the acoustic (UBM) clustering of MFCC and that for an alternative feature, for all the utterances used in UBM training. The alternative features chosen for comparison include Linear Prediction Cepstrum Coefficient (LPCC), perceptual linear prediction coefficients (PLP), spectral centroid frequency (SCF) [35] (discussed in section 3.2.1.1), log compressed least squares group delay (LogLSGD) (discussed in section 3.1.1) and dropping the first cepstral coefficient (C_1) of MFCC (MFCCDropCep1) (discussed in section 4.1.4). The first two

[10] In this thesis, the terms 'clustering' and 'clusters' refer to Gaussian mixture model clustering and Gaussian mixture components respectively.

features were chosen as they are the typical features of NIST SRE consortium submissions besides MFCC, and the latter are examples of phase-based, frequency-based and MFCC variant features. In this experiment, LPCC, PLP and MFCC (14 dimensions each) were computed from the pre-emphasised speech (with a pre-emphasis factor of 0.97), every 10 ms using an analysis window of 20 ms. Feature warping was then performed, and delta coefficients are appended to achieve the final feature vector. However any features that represent diversity in the characterisation of speech would suffice in an experiment of this kind. To calculate the NID between the UBM clusterings of two features (e.g. feature *A* and feature *B*), the UBM for each feature is first trained and will be referred to as the clustering **U** for feature *A* and clustering **V** for feature *B* in this experiment for simplicity (as discussed in section 4.2.2). Then, for the utterances that have been used to trained the UBM for each feature, the most probable individual Gaussian mixture component on a frame level were computed with respect to its corresponding clustering (feature *A* with respect to clustering **U** and feature *B* with respect to clustering **V**). Finally, with the two sets of cluster labels (equation (4.3), one set for feature *A* with respect to clustering **U** and one set for feature *B* with respect to clustering **V**), the cluster overlap across all the utterances (frame-by-frame basis) is summarised using the contingency table (as shown in Table 4.6) and NID between the UBM clustering are computed using equation (4.4).

The results for the female subset of the core condition of the NIST 2006 SRE database with gender-dependent universal background models trained using NIST 2004 SRE database are summarised in Table 4.7. It can be observed from the result that the greater the distance (more difference) between the pairs of UBMs (higher NID), consistently the greater the reduction in terms of EER for the fused system. This is most probably due to features being derived from different sources of information in speech (e.g. frequency, phase) which models different aspects of the acoustic space. Thus, the amount of complementary information is observed to be proportional to the amount of difference between the partitions. This further supports the hypothesis that diversity can be achieved purely through different partitioning of the acoustic space as speculated in this chapter. Hence for researchers working on the development of new/complementary features (and perhaps VADs and front-ends in general), the NID between the UBMs could be utilised as an initial indicator of the amount of complementary information and optimal settings of the proposed feature before extracting features for all the databases, which could be time-consuming. Also, the fact that the UBM alone (with no speaker modelling) could be somewhat predictive of EER seems to be an interesting new perspective on speaker recognition research using the GMM-UBM approach.

Table 4.7 NID between UBM of fused systems on NIST 2004 SRE female dataset and system EER on NIST 2006 SRE core condition (512-mixtures UBM)

Features	EER (%)	Feature-MFCC NID (UBM)	EER Fused with MFCC (%)
MFCC	6.66	-	-
MFCCDropCep1	6.90	0.47	6.31
LogLSGD	8.12	0.65	6.04
SCF	7.82	0.73	5.87
PLP	7.45	0.71	5.87
LPCC	6.74	0.78	5.72

4.2.4 Investigation of Acoustic Modelling

In addition to shedding some light on the fusion of systems employing different features, clustering comparison can be used to understand single-feature systems better. The aim of this experiment was to assess the “stability” of the clusters (GMMs) with respect to different features, that is, the robustness of the putative clusters to sampling variability. As mentioned previously, one hypothesis for the poorer performance of alternative features is that their ability to describe the acoustic space of a speaker is less stable/reliable than that of MFCC. This notion is motivated by research into clustering stability by Smith et al. [170]. The experiment was conducted through the use of resampling, proposed in [171, 172], to assess the stability of the clustering results with respect to sampling variability by simulating permutations of the original data set. The basic assumption of this method is intuitively simple: if the data represent a sample of items drawn from distinct sub-populations, and if different samples are drawn from the same sub-populations, the induced cluster compositions should not be radically different from each other. Therefore, the more the attained clusters are robust to sampling variability, the more we can be confident that these clusters represent the real underlying structure.

Following studies in [171, 172], a similar experiment was conducted on the NIST 2004 female database (1849 speakers, one utterance per speaker), with the NID as a measure of agreement between alternative sets of data. In order to determine if a feature provide stable clustering in the acoustic space, each utterance is first resampled randomly into k disjoint subsets (on the frame level). Using the k subsets of utterances, a UBM is trained for each subset. Next, the procedure of finding the highest probable Gaussian mixture components (discussed in section 4.2.2) is repeated across the j^{th} subset of utterances with respect to the j^{th} UBM resulting in k sets of cluster labels. This procedure is termed the feature stability assessment procedure and is summarised in Figure 4.9. Finally, a NID is computed for all possible pairwise set of cluster labels. Shown in Table 4.8 are the average NID and EER for various features with k=10. It can be observed from the results that better feature stability, or smaller average distance between the putative clusters (lower average NID) corresponds with a lower EER, demonstrating the importance of stability in acoustic modelling, and explaining in clustering terms why MFCC usually outperforms alternative features. As with section 4.2.3, it can be observed that the modelling of the acoustic space by a particular feature has a strong relationship with the EER - in both sections the NID is computed only on the UBM and speaker modelling is not a part of these NID calculations. The fact that the smallest average NID between UBMs trained on subsampled UBM occurs for the MFCC (which has dominated speaker recognition feature extraction for many years now) is probably no surprise, however it does provide a novel perspective on the success of MFCC as a feature.

1. Given a set of speakers/ utterances $\mathbf{S} = \{s_1, s_2, \dots, s_N\}$ for UBM training
2. Resample/generate k (at least partly) disjoint permuted subsets of size $|S|/k$ for all utterances in S
3. Generate k UBMs with the k permuted subsets of S using the EM algorithm
4. For each training utterance, determine the closest (highest probable) UBM mixture in each UBM

Figure 4.9 Feature stability assessment procedure (after [171])

Table 4.8 NID between UBMs trained on subsampled UBM and EER on NIST2004 female subset

Features	Average NID	EER (%)
MFCC	0.44	6.66
LPCC	0.45	6.74
MFCCDropCep1	0.47	6.90
PLP	0.47	7.45
SCF	0.56	7.82
LogLSGD	0.57	8.12

4.2.5 Investigation of Speaker Modelling

In Section 4.2.3, we have shown that different features result in different partitioning of the acoustic space. In this section, we investigate the extent to which different features give different clusterings after MAP adaptation, with respect to the UBM. This methodology is necessarily different because MAP adaptation of the UBM is done on a per-utterance basis, and it is not possible to associate an EER with each NID between the

UBM and utterance-adapted UBM. In order to compute the average NID between clustering of UBM and speaker models for each feature, the training utterance of each speaker is first used to create a speaker model through MAP adaptation from the UBM. Then with respect to the clustering of the UBM and speaker models, the closest Gaussian mixture component for the training utterance (at frame level) is computed. Finally the average NID for each feature is computed through the averaging of all NID values between the clustering of UBM and speaker models.

For illustration, we look at the adaptation of the UBM for a single speaker (total 1849 speakers) from the NIST 2004 female database for two very differently clustered features (MFCC and SCF), expecting to see that adapted models exhibit more deviation in one feature domain than the other if we employ the assumption that different features carry different speaker-specific information. Figure 4.10 shows the NID comparison between MFCC and SCF with an average NID of 0.5002 for MFCC and an average NID of 0.5361 for SCF. Surprisingly, we observed that all feature sets have almost equal amounts of adaptation in both feature domains, with respect to the UBM as shown in shown Table 4.9. This might suggest that different features do not differ very much in how much they model individual speakers, or that speaker modelling may not be as important as acoustic modelling.

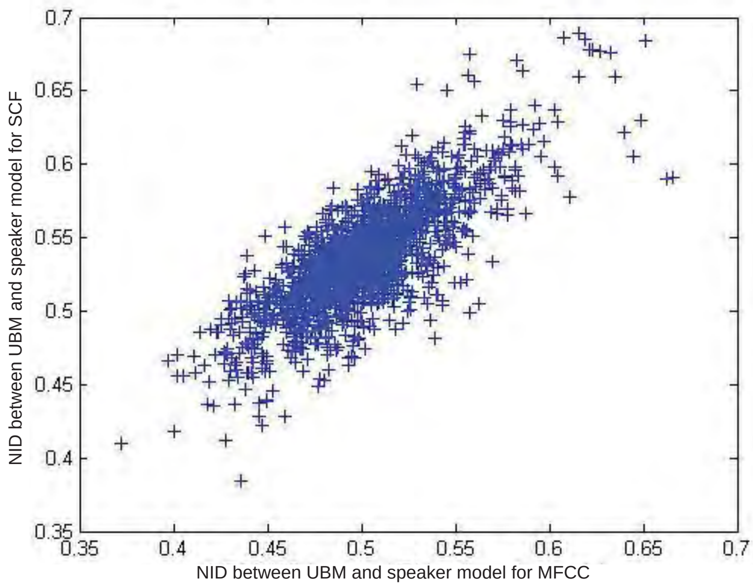


Figure 4.10 NID between clustering of UBM and speaker model for MFCC and SCF where each '+' sign correspond to one speaker (1849 speakers)

Table 4.9 Average NID between clustering of UBM and speaker model

Features	Average NID
MFCC	0.5002
LPCC	0.5000
MFCCDropCep1	0.5050
PLP	0.5211
SCF	0.5361
LogLSGD	0.5312

4.2.6 UBM Data Selection Using Clustering Comparison

Based on the results of the foregoing sections, it was observed that front-end diversity can be achieved by fusing systems with different 'partitioning' in the acoustic space (UBM). - Traditional clustering algorithms (i.e. k-means) have shown good performance through the creation of a single good clustering solution. Data, however, often bear multiple equally reasonable clusterings and this has led to the recent emergence of the field of alternative clustering [173]. Alternative clustering aims to create different clustering solutions that are distinctive from each other. In an attempt to investigate if the fusion of systems with different UBM clustering created through the use of alternative clustering algorithm improves on the speaker recognition performance, we used one of the alternative clustering algorithm, minCEntropy [173], to generate alternative clustering for the MFCC-based system with the same number of clusters. The fusion of the k-means clustered MFCC systems (conventional system used in this thesis generated using HTK[11]) and one of the minCEntropy clustered MFCC systems achieved an EER of 6.35% (baseline EER = 6.66%) and 5.2% (baseline EER = 5.4%) on the NIST 2006 SRE female and male subsets respectively. This finding was consistent across two genders, suggesting that the fusion of alternative clustering systems carry complementary information.

In addition, in regards to acoustic modelling, a common assumption in UBM training for speaker recognition is that the more utterances used, the better the system performance. However according to [8], Reynolds et al. mentioned that a small amount of data is sufficient for a reasonable system. Recently Hasan et al. proposed a novel feature subsampling method for selecting UBM feature vector frames [174], in which they also demonstrate the significance of the data to be selected for effective UBM training in terms of entire speaker recognition system performance.

Since it seems that acoustic modelling is a good area to focus investigation in speaker recognition, and further that feature stability has an important role to play, the consensus clustering based approach in Figure 4.9 suggests itself for application to utterance selection for UBM training. First, the frame log-likelihood of each training utterance (of the j^{th} UBM) was computed against its j^{th} UBM. Then in order to determine which utterance constantly contributes more to the clustering of the UBMs, the frame-based log-likelihoods were averaged (across the same frame) across the k UBMs. Finally, the top

[11] http://htk.eng.cam.ac.uk/

x% utterances with the highest averaged log-likelihood were chosen for training a new UBM (procedure summarised in Figure 4.11).

Figure 4.12 shows the EER versus the percentage of data used for UBM training data (*x*%) when the utterances are selected according to the proposed technique and randomly on NIST SRE 2006 database. It can be observed in both cases that using fewer utterances for training UBM results in better system performance. This could be due to the fact that introducing more utterances to a "stable" UBM may simply increase the variability within the UBM which might not be desirable in terms of acoustic space modelling. Using the proposed UBM data selection algorithm, a 4% and 11% relative reduction in EER using only 20% and 30% of the usual female and male UBM training data sets respectively (~370 utterances for each gender) was achieved; better performance improvement as compared with random utterance selection. In addition, this corresponds to a reduction in UBM development time of up to 3 times (once the data have been selected) as compared with using all utterances.

1. Repeat step 1 – 3 of feature stability assessment procedure (Figure 4.9)
2. For each training utterance, determine frame log-likelihood against its corresponding UBM.
3. Compute averaged log-likelihood (averaged across only those partitions that overlap for that utterance)
4. The top *x*% utterances with the highest averaged loglikelihood are chosen for training a new UBM

Figure 4.11 Proposed UBM data selection procedure

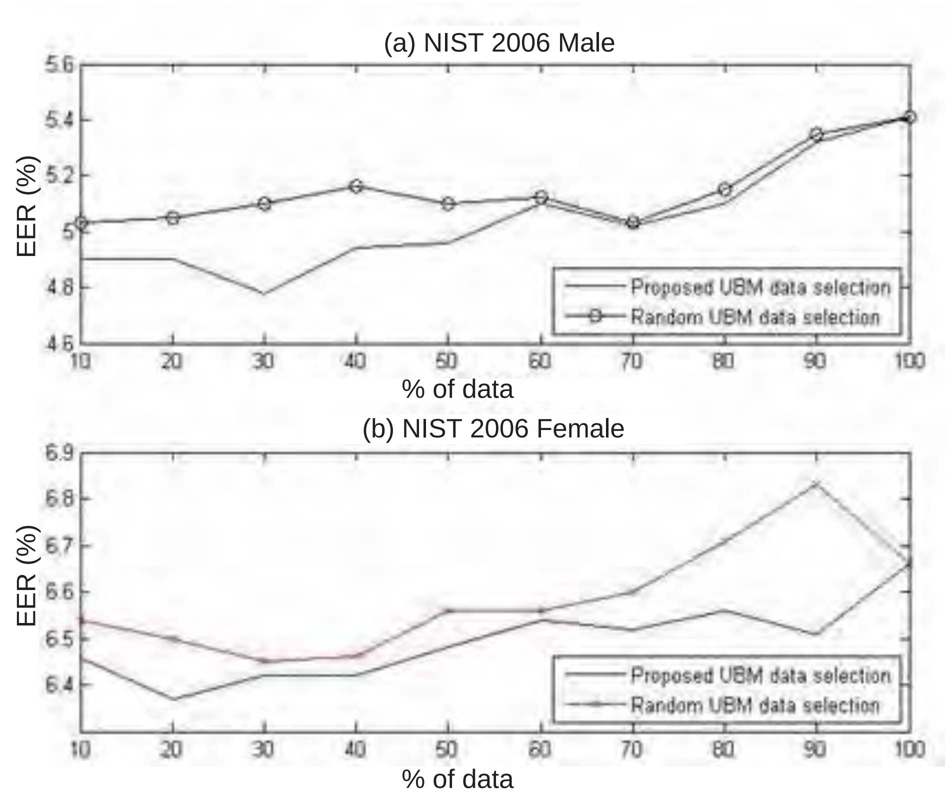


Figure 4.12 EER vs. percent of selected data for gender dependent UBM training. The EER performances for each random UBM data selection have been averaged across 10 individual speaker recognition experiments.

4.3 Summary

In this chapter, the assumption that performance improvement in terms of front-end diversity (feature-level) can be best attained through fusion of systems based on acoustic features that are from different origins of the features (e.g. magnitude, phase, modulation information) was explored. We proposed an ensemble of different variants of MFCCs and showed that the fusion of suboptimal systems based on features comprising essentially the same information as MFCCs consistently outperformed an individual MFCC based system. Furthermore, for particular design choices of MFCC-variant features, the improvement can be significant. In particular, the use of different filter shapes was found to provide modified MFCCs that perform promisingly in fused systems, providing the filterbank ripple is minimised. Evaluations on the NIST 2006 SRE database show a relative improvement of 17% in EER when one modified MFCC subsystem is fused with a conventional MFCC-based system, and an improvement of 22% when two modified MFCC subsystems are fused. This perhaps prompts a re-evaluation of what types of features might be considered complementary, and we believe that this supports the hypothesis that diversity can be achieved purely through different ’partitioning’ of the acoustic space.

Next, in an attempt to *separately* investigate the acoustic and speaker modelling ‘stages’ of the GMM-UBM based systems (which is new in the context of speaker recognition), towards determining the contributions of each stage to the speaker recognition performance across different features, we proposed the use of clustering comparison measures, in particular the Normalised Information Distance. It was observed from the experimental results that the amount of difference in terms of acoustic clustering between pairs of different features is correlated with the fused speaker recognition performance, again demonstrating that front-end diversity can be achieved purely through different ‘partitioning’ of the acoustic space. Further, features that exhibit good ‘stability’ with respect to repeated clustering are shown to also give good EER performance in speaker recognition. This has implications for feature choice, fusion of systems employing different features, and for UBM data selection, to be discussed further in Chapter 6. A novel utterance selection algorithm for the latter problem on training a “stable” UBM was presented and evaluated on the NIST 2006 database. Results show that using NID-based resampling to select utterances during UBM training can improve speaker recognition performance up to an 11% relative reduction in EER despite employing a smaller set of training data (20-30% of the usual UBM training data set).

Chapter 5

Sparse Representation Classification for Speaker Recognition

In the previous chapters, the focus is on achieving diversity through front-end processing in which the ASV systems described are based on a Gaussian mixture model (GMM) back-end (either GMM-UBM or GMM-SVM), which is useful for comparing different front-ends. However, the fact that systems employing different classifiers fused in a complementary way, reducing error rates substantially (as shown in [20]), brings our attention to alternative classification methods for ASV. Recently, one machine learning technique that has received significant focus in pattern recognition literature is sparse representation classification (SRC). The discriminative nature of SRC has been successfully demonstrated in pattern recognition tasks such as face recognition [28], signal classification [31] and speaker identification [29] as discussed in section 2.4.3.

In the speaker recognition area, Naseem et al. [29] were the first to introduce a supervector-based sparse representation classifier (GMM-SRC) for speaker identification. Although their experiments showed good performance compared with GMM-SVM/GMM-UBM, the investigations were conducted on the relatively small TIMIT database that characterises an ideal speech acquisition environment and does not include reverberant noise and session variability. In this chapter, we will first extend their work to a speaker verification task on the contemporary NIST SRE databases, followed by an investigation on the inclusion of inter-session variability compensation methods for GMM-SRC. Then we will investigate sparse representation classification of low-dimension speaker factors[12] as an alternative to the large-dimensional supervectors, producing an approach we term Joint Factor Analysis – Sparse Representation Classification (JFA-SRC) for speaker verification. Since as discussed earlier (in section 2.4.2), ever since SVMs were introduced to the field of speaker recognition by Campbell et al. [20], various investigations have been conducted in each individual component of SVM (e.g type of kernel, SVM cost parameter, kernel parameters and background dataset) with the intent of improving the system performance and/or increasing the computational efficiency of SVM training. Hence, in this chapter, we extend our analysis to different types of sparseness methods, dictionary composition and ways to improve the robustness of SRC against corruption to determine the best configuration for speaker recognition using SRC.

5.1 Classification based on Sparse Representation

As discussed in section 2.4.3, for classification problems, a test sample (**S**) can be written as a linear combination of the training samples from L classes in the overcomplete dictionary, **D** as follows

$$\begin{aligned}\mathbf{S} &\approx \mathbf{D}\boldsymbol{\gamma} \\ &\approx \alpha_{1,1}\mathbf{v}_{\mathbf{1},\mathbf{1}} + \cdots + \alpha_{1,l_1}\mathbf{v}_{\mathbf{1},\boldsymbol{l}_\mathbf{1}} + \cdots + \alpha_{i,\boldsymbol{l}_\mathbf{1}}\mathbf{v}_{\boldsymbol{i},\boldsymbol{l}_\mathbf{1}} + \cdots + \alpha_{i,l_i}\mathbf{v}_{\boldsymbol{i},\boldsymbol{l}_{\boldsymbol{i}}} + \cdots + \alpha_{L,l_L}\mathbf{v}_{\boldsymbol{L},\boldsymbol{l}_{\boldsymbol{L}}}\end{aligned} \tag{5.1}$$

where the coefficient vector $\boldsymbol{\gamma} = \left[0, \dots, 0, \alpha_{i,1}, \alpha_{i,2}, \dots \alpha_{i,l_i}, 0, \dots, 0\right]^T$, termed the sparse coefficients [28], has entries that are mostly zero except those associated with the *i*th class after solving the linear system of equations $\mathbf{S} = \mathbf{D}\boldsymbol{\gamma}$ using ℓ_1-norm minimization (in equation (2.29)). In this case, the indices of the sparse coefficients encode the identity of

[12] Herein speaker factors is chosen over i-vectors as features for the SRC due to their excellent discriminative capabilities as compared with i-vectors, as reported in [136].

the test sample **S**, and these form the non-zero entries of what we term the 'sparse coefficient vector', $\boldsymbol{\psi}$.

In order to demonstrate the basic concept of sparse representation classification using ℓ_1-norm minimization (Equation (2.29)), an example matrix $\mathbf{D} = [[8.33\ 8\ 7.43]^T, [8.14\ 7.24\ 8.34]^T, [6.19\ 8.11\ 4.11]^T, [5.3\ 8.03\ 4.87]^T, [10.66\ 5.23\ 6.04]^T, [10.72\ 4.19\ 6.81]^T, [7.21\ 3.03\ 1.28]^T, [8.11\ 2\ 2.53]^T, [4.69\ 1.83\ 3.97]^T, [4.01\ 0.54\ 2.2]^T, [5.01\ 4.17\ 8.52]^T, [4.16\ 4.32\ 8.03]^T]$ was created using a small number of synthetic 3-dimensional data (K = 3), where the columns of **D** represent 6 different classes with 2 samples for each class (L = 6 , N = 12). A test vector $\mathbf{S} = [8\ 7\ 7]^T$ was chosen near to class 1, as shown in Figure 5.1. Solving Equation (2.29)[13] produces the vector $\boldsymbol{\gamma} \approx [0.59, 0.22, 0, 0, 0.10, 0, 0, 0, 0, 0, 0, 0]^T$, where the largest value (0.59) corresponds to the correct class (1), but $\boldsymbol{\psi}$ also has entries in other training samples of classes 1 and 3. Although ideally $\boldsymbol{\psi}$ would only be associated with the columns of **D** from a single class i, we can still easily assign the test sample **S** to that class. However, noise may lead to small nonzero entries associated with multiple classes.

[13] This example is solved using the MATLAB implementation of Gradient Projection for Sparse Reconstruction (GPSR) which is available online on http://www.lx.it.pt/~mtf/GPSR/.

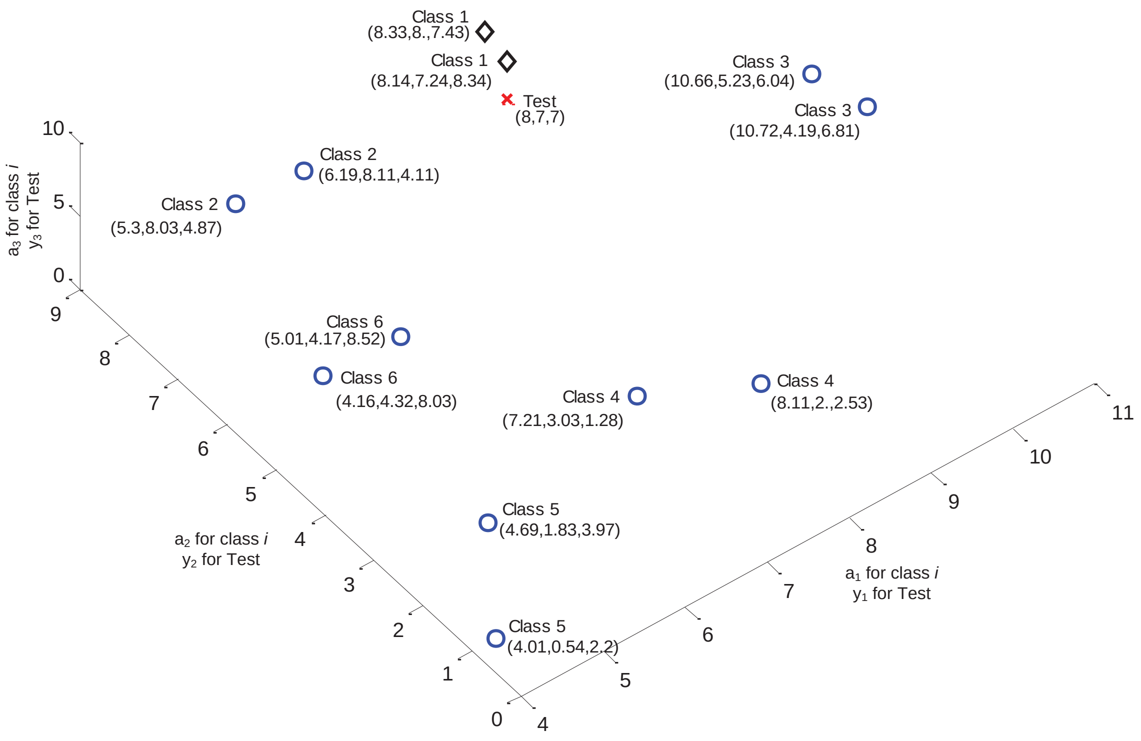


Figure 5.1 Example of sparse representation classification on synthetic 3-dimensional data (K = 3) comprising 6 classes with two training samples each (L = 6, N = 12) where × and ♢correspond to the test and training data from class 1 (correct class) respectively and circles correspond to the training data from classes 2 to 6.

For more realistic classification problems, **S** can be classified based on how well the coefficients associated with all training samples of each class reproduce **S,** instead of simply assigning **S** to the object class with the single largest entry in $\boldsymbol{\gamma}$ [28]. For each class i, let $\delta_i: \mathbb{R}^N \longrightarrow \mathbb{R}^N$ be the characteristic function that selects the coefficients associated with the ith class as shown in equation (5.2).

$$\delta_i(\boldsymbol{\gamma}) = \begin{bmatrix} \sigma_{1,1} \\ \sigma_{1,2} \\ \vdots \\ \sigma_{1,l_1} \\ \vdots \\ \sigma_{L,1} \\ \sigma_{L,2} \\ \vdots \\ \sigma_{L,l_L} \end{bmatrix} where\ \sigma_{j,k} = \begin{cases} 0\,, j \notin class\ i\ \forall\ k \\ \alpha_{i,k}\,, j \in class\ i\ \forall\ k \end{cases} \tag{5.2}$$

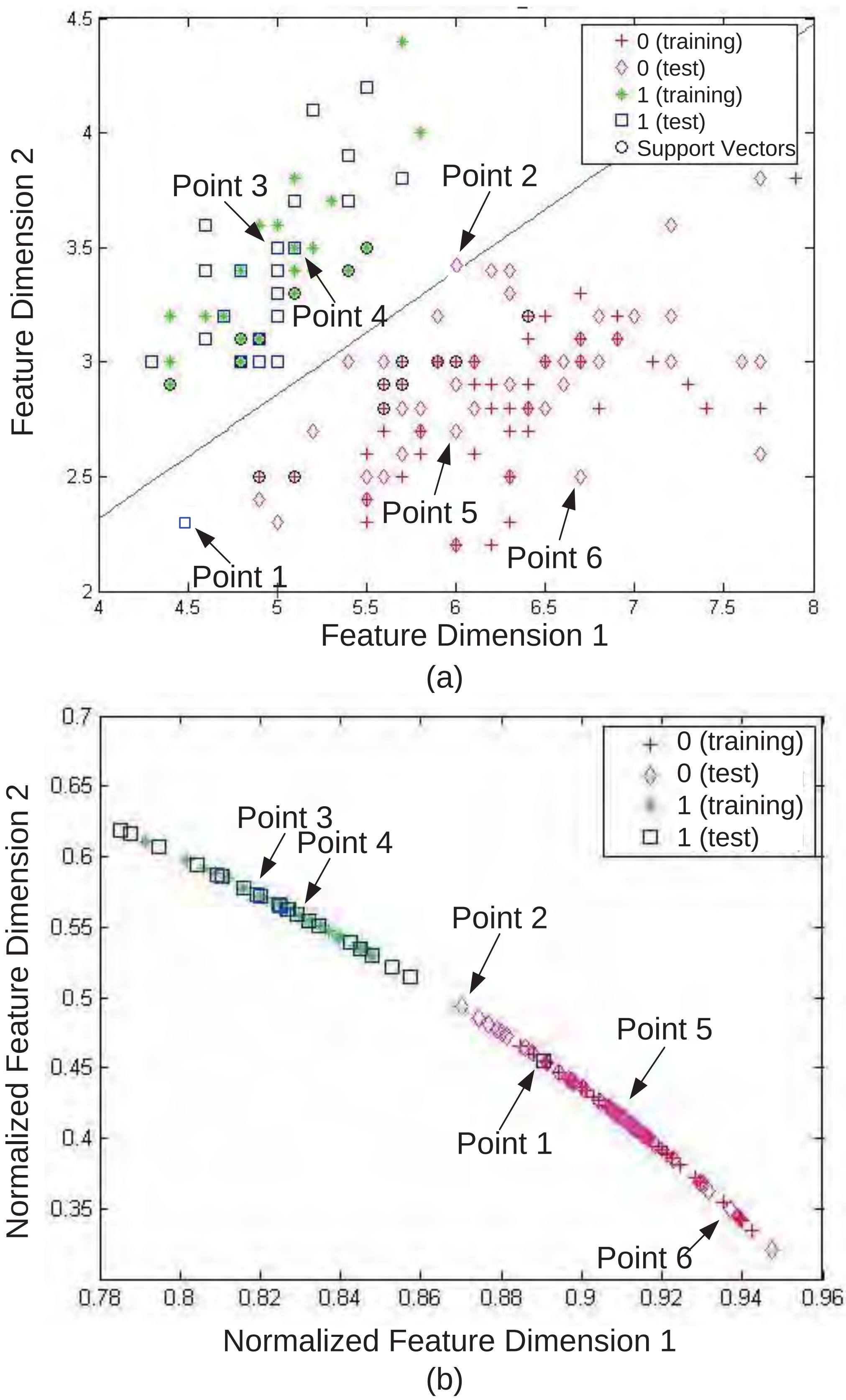
Hence for the example shown in Figure 5.1, the characteristic function for class 1 would be $\delta_1(\boldsymbol{\gamma}) = [0.59, 0.22, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]^T$. Using only the coefficients associated with the *i*th class, one can approximate the given test sample $\mathbf{S}$ as $\hat{\mathbf{S}}_i = \mathbf{D}\delta_i(\boldsymbol{\gamma})$. $\mathbf{S}$ is then assigned to the object class, $\mathbb{C}_{\mathbf{S}}$, that minimises the residual between $\mathbf{S}$ and $\hat{\mathbf{S}}_i$:

$$\mathbb{C}_{\boldsymbol{S}} = \arg\min_i r_i(\mathbf{S}) \quad where \quad r_i(\mathbf{S}) \approx \left\| \mathbf{S} - \hat{\mathbf{S}}_i \right\|_2 \tag{5.3}$$

5.2 Comparison of SVM and SRC classification

In this section, a comparison of SVM and SRC in terms of recognition performance was conducted with the aim of understanding the similarities and differences between the classifiers, since in the literature, various authors [28, 29, 33, 175] have shown experimentally that SRC is able to achieve comparable performance to SVM and at times outperform it. Furthermore, it has been shown through mathematical derivation (in [176]) that to some extent, SVM and SRC are equivalent (when the data are noiseless). We considered simple 2-dimensional data for easy visualisation, as shown in Figure 5.2. For sparse representation-based classification, all the samples are normalised to have unit ℓ_2-norm (as shown in Figure 5.2 (b)), which matches the length normalisation in the SVM kernel. This experiment is conducted on the Fisher iris data [177] using the sepal length and width for classifying data into two groups: Setosa and non-Setosa shown as “Class 1” and “Class 0” respectively on Figure 5.2. The experiment was repeated 20 times, with the training and testing sets selected randomly.

Notably, the performance of SRC matches that of the SVM in 19 out of the 20 trials. Similarly to SVM, the sparse representation approach also finds it difficult to classify the same test point indicated as “point 1” in Figure 5.2 (a) for SVM and (b) for SRC, since it is in the subspace of class 0 for both classifiers. However “point 2” (shown in Figure 5.2) is correctly classified as class 0 for SRC and misclassified as class 1 by SVM. This could be because SVM does not adapt the number and type of supports to each test example. It selects a sparse subset of relevant training data, known as support vectors (shown as circle in Figure 5.2 (a)) which correspond to the data points from the training set lying on the boundaries of the trained hyperplane, and uses these supports to characterise "all" data in the test set. Although visually "point 2" is closer to the training subset of class 0, it is misclassified since it is on the left hand side of the hyperplane, corresponding to class 1. SRC allows a more adaptive classification with respect to the test sample by changing the number and type of support training samples for each test sample [178] as shown in the sparse coefficients of four test samples (Figure 5.2 (c) – (f)) chosen from Figure 5.2 (b), indicated as "point 3" to "point 6" respectively, whereas the SVM classifies with the same support vector weights, α (refer to equation 2.25), as shown in Figure 5.2 (c) – (f), across all test data in the test set. In addition, Figure 5.2 supports the concept that test samples can be represented as a linear combination of the training samples from the same class since it can be observed from Figure 5.2 (c) – (d) that for test samples from Class 1 (indicated as Point 3 and 4 on Figure 5.2 (b)), the sparse coefficients have larger values for the dictionary indices belonging to class 1 and the same applies to Point 5 and 6 for Class 0 (shown in Figure 5.2 (e) – (f)).



0 (training)
0 (test)
1 (training)
1 (test)
Support Vectors
Point 1
Point 2
Point 3
Point 4
Point 5
Point 6
Feature Dimension 1
Feature Dimension 2
(a)
0 (training)
0 (test)
1 (training)
1 (test)
Point 1
Point 2
Point 3
Point 4
Point 5
Point 6
Normalized Feature Dimension 1
Normalized Feature Dimension 2
(b)

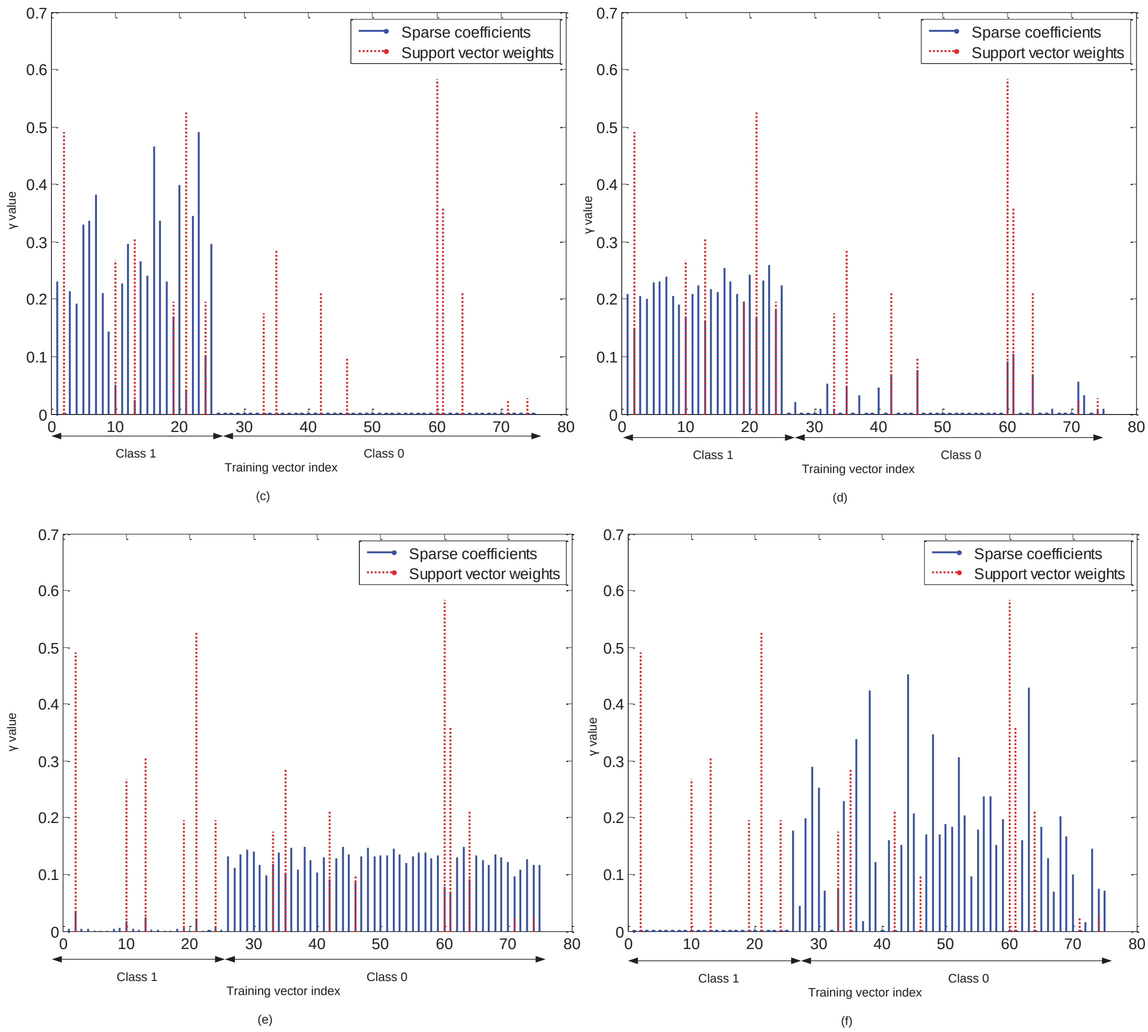


Figure 5.2 Comparison between (a) SVM and (b) SRC for a two-class problem (class 0 and class 1) where '+' and '*' correspond to the training set instances for class 0 and class 1 respectively. ◊ and □ correspond to the test points for class 0 and class 1 respectively. ○ are the support vectors chosen from the training data sets of each class for SVM. (c) – (f) The values of the sparse coefficients and weights of the support vectors, α (shown in Figure 5.2 (a)) for test points 3 – 6 respectively

5.3 Speaker Recognition based on SRC

5.3.1 Supervector-based SRC

In [29], Naseem et al. proposed the use of the GMM mean supervector, $\mathbb{M}$, to develop an over-complete dictionary using all the training utterances of speakers in a database for speaker identification where they achieved better recognition accuracy as compared with current state-of-the-art systems (GMM-SVM and GMM-UBM). Likewise, we begin by employing a similar approach, termed GMM-Sparse Representation Classification

(GMM-SRC), in the context of speaker verification, whereby the over-complete dictionary (**D**) is composed of the normalised supervectors (with unit ℓ_2 norm) of training utterances from the target speaker ($\mathbf{D}_{tar}$) and the background speakers ($\mathbf{D}_{bg}$) as shown in equations (5.4). The normalisation process is analogous to the length normalisation in the SVM kernel and in this thesis the dictionary data composition is the same as the kernel training data for SVM unless otherwise specified. In the context of speaker verification, usually $l_{bg} \gg l_{tar}$ with l_{tar} = 1, where l_{bg} and l_{tar} represent the number of utterances from the background and target speakers respectively.

$$\mathbf{D} = [\mathbf{D}_{tar}\mathbf{D}_{bg}] \tag{5.4a}$$

$$\mathbf{D}_{tar} = [\mathbb{M}_{tar,1}, \dots, \mathbb{M}_{tar,l_{tar}}] \tag{5.4b}$$

$$\mathbf{D}_{bg} = [\mathbb{M}_{bg,1}, \dots, \mathbb{M}_{bg,l_{bg}}] \tag{5.4c}$$

Following this, the GMM mean supervector of a test utterance (**S**) from an unknown speaker is represented as a linear combination of this over-complete dictionary, a process referred to as sparse representation classification for speaker recognition, as follows

$$\mathbf{S} = \mathbf{D}\boldsymbol{\gamma} \tag{5.5}$$

Throughout the testing process, the background samples $\mathbf{D}_{bg}$ are fixed and only the target samples $\mathbf{D}_{tar}$ are replaced with respect to the claimed target identity in the test trial. Due to the high dimensionality of the supervectors, equation (5.5) usually represents an overdetermined system of equations, but it has been shown that sparse approximate solutions $\boldsymbol{\gamma}$ can still be obtained by solving the ε-relaxed ℓ_1-minimization [28, 179]. Although a least squares approach is usually used to find an approximate solution to overdetermined systems, it requires large computational cost to solve large systems of equations and the least-square solution can exhibit severe bias if the system is not properly regularised [180]. In addition, the data are of very high dimension, with only a few samples per subject. The small sample size exacerbates “the curse of dimensionality” that plagues high-dimensional statistics [181]. This problem is exhibited by nearest neighbour classifiers in this context, resulting in poor performance [28, 182].

In the context of speaker verification, $\boldsymbol{\gamma}$ is sparse since the test utterance corresponds to only a very small fraction of the dictionary. As a result, $\boldsymbol{\gamma}$, obtained efficiently via ℓ_1-minimisation, will have large $\boldsymbol{\psi}$ corresponding to the correct target speaker of the test utterance as shown in Figure 5.3(a), where the dictionary index n=1 corresponds to the true target speaker. On the other hand, if the test utterance is from an impostor, the coefficients will be sparsely distributed across multiple speakers in the dictionary [33, 183], as shown in Figure 5.3(b). Here, the membership of the sparse representation in the over-complete dictionary itself captures the discriminative information since it adaptively selects the relevant vectors from the dictionary with the fundamental assumption that test samples from a class lie in the linear span of the dictionary entries corresponding to the class of the test samples [28, 34].

Therefore, given sufficient training samples of an object class, any new sample **S** from the same class can be expressed as a linear combination of the corresponding training samples. This assumption is valid in the context of speaker recognition since it has been shown by Ariki et al. that each individual speaker has their own subspace [184, 185]. In addition, even though the number of background examples significantly outweighs that of target speaker examples, the SRC framework is not affected by the unbalanced training set, in contrast to an SVM system which requires tuning of the SVM cost values. This is because for SVM, a hyperplane trained by an unbalanced training set will be biased toward the class with more training samples [186, 187], but this is not the case for SRC. On the other hand, SRC utilises the highly unbalanced nature of the training examples to form a sparse representation problem [188].

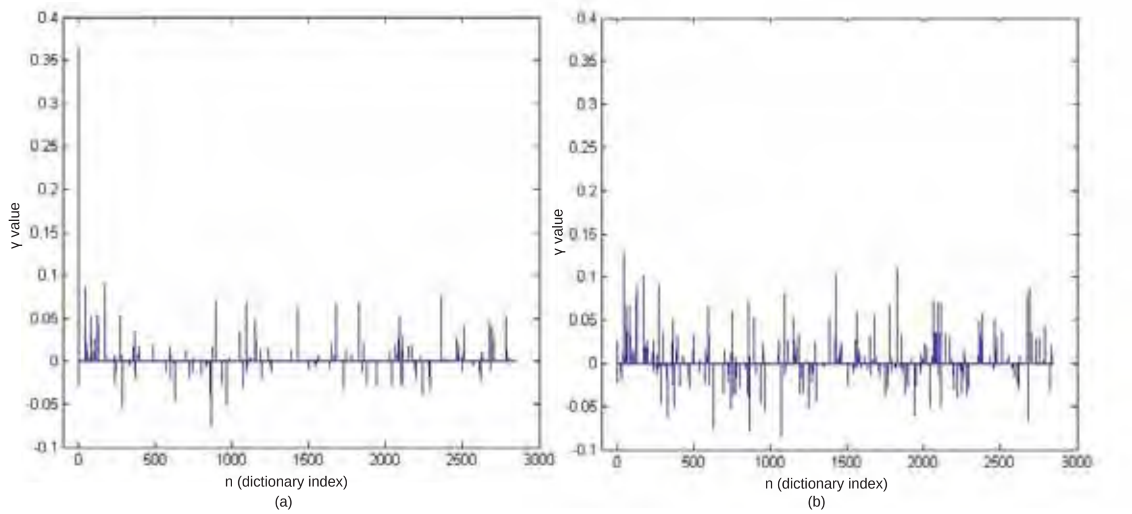


Figure 5.3 The sparse solution $\boldsymbol{\gamma}$ of two example speaker verification trials where $\boldsymbol{\gamma}$ is a function of the dictionary index n, (a) True target (n = 1) (b) Impostor

The target ℓ_1-norm, $\wp$ shown in equation (5.6) is used as the decision criterion for verification, where the operator δ_{target} as shown in equation (5.2) selects only the coefficients associated with the target class. The example shown in Figure 5.3 has ℓ_1-norm of 0.354 and 0.035 for the true target (Figure 5.3 (a)) and impostor (Figure 5.3(b)) respectively.

$$\wp = \left\| \delta_{target}(\boldsymbol{\gamma}) \right\|_1 \tag{5.6}$$

Finally, the detailed architecture of the proposed GMM-SRC system based on GMM supervectors for speaker verification is shown in Figure 5.4, where $K = MD$ corresponds to the supervector dimension, M is the total number of mixtures, D is the dimension of the feature vector and N is the total number of utterances from target and background speakers.

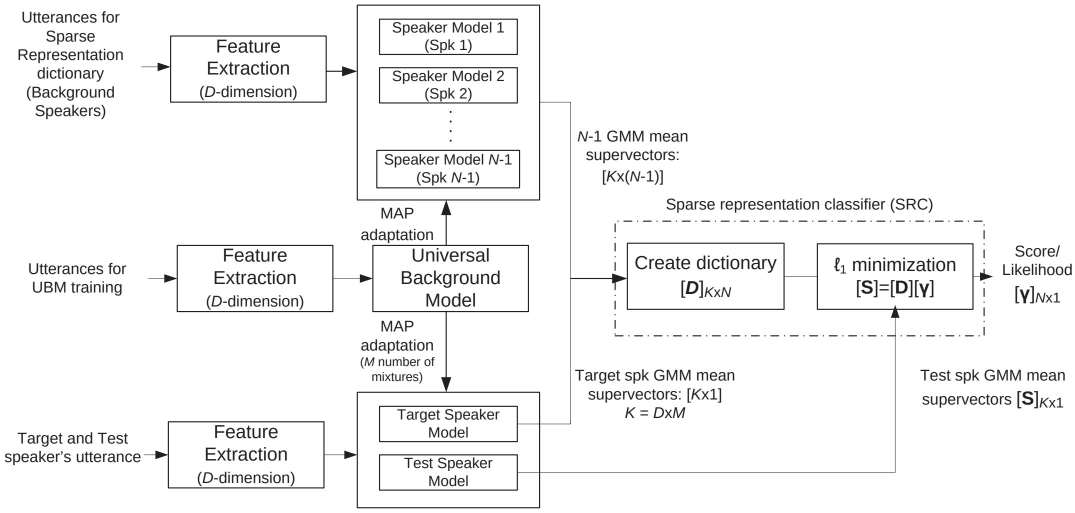


Figure 5.4 Architecture of the GMM-SRC system based on GMM supervectors.

5.3.2 Proposed Speaker Factor-based SRC

Motivated by [189], where the authors proposed the use of speaker (**y**) and common (**z**) factors components defined by the JFA model (in equation (2.35)) as features for the SVM, we adopt the speaker factors, which correspond to speaker coordinates in the speaker space defined by **V**, as feature vectors for the SRC. In addition, the JFA model seems to dominate in recent years of speaker recognition evaluations (SRE) [4] and was pursued further in the Johns Hopkins University (JHU) workshop by Burget et al. [18] and Najim et al. [190] in 2008 and 2011 respectively. Independent evaluations by different research groups have clearly indicated the potential of JFA.

The underlying structure of the speaker factor-based SRC, which we term Joint Factor Analysis Sparse Representation Classification (JFA-SRC), is similar to GMM-SRC except that the speaker factors are used to develop the over-complete dictionary and for testing as opposed to supervectors in the previous section, as shown in equations (5.7). A detailed architecture of the JFA-SRC system is shown in Figure 5.5, where *F* corresponds to the speaker factor dimension.

$$\mathbf{D} = \left[\mathbf{D}_{tar}\mathbf{D}_{bg}\right] \tag{5.7a}$$

$$\mathbf{D}_{tar} = \left[\mathbf{y}_{tar,1}, \dots, \mathbf{y}_{tar,l_{tar}}\right] \tag{5.7b}$$

$$\mathbf{D}_{bg} = \left[\mathbf{y}_{bg,1}, \dots, \mathbf{y}_{bg,l_{bg}}\right] \tag{5.7c}$$

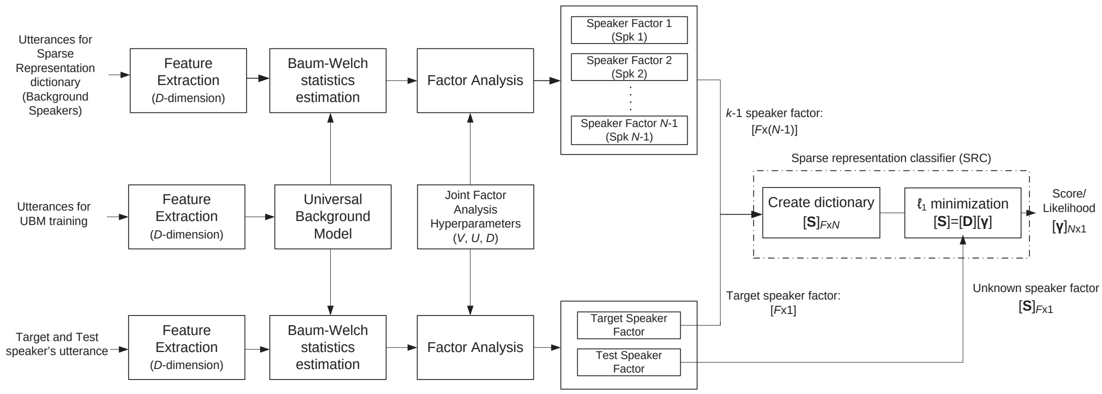


Figure 5.5 Architecture of the proposed JFA-SRC system based on speaker factors.

5.4 System Development using SRC

5.4.1 Experimental Setup

All experiments reported in this section were based on MFCC (refer to section 3.1.3 for configuration) and carried out on the core condition female trials of the NIST 2006 SRE dataset. Four current state of the art systems, namely GMM-SVM [12], JFA-SVM in speaker factor space [189], joint factor analysis-cosine distance scoring (JFA-CDS) in speaker factor space [191] and JFA [192] were implemented as baseline systems. In our SVM system, we took 2843 female SVM background impostor models from NIST 2004 to train the SVM. In addition, for the GMM-SVM system, NAP (rank 40) trained using NIST 2004 and 2005 SRE corpora was incorporated to remove unwanted channel or intersession variability [12]. On the other hand for JFA-SVM and JFA-CDS, LDA (trained using Switchboard II, NIST 2004 and 2005 SRE) without any dimensionality reduction (dim = 300) followed by WCCN (trained using NIST 2004 and 2005 SRE) were used for session compensation [132]. For JFA-SVM, JFA-CDS and JFA, the (gender dependent) factor analysis models were trained using LDC releases of Switchboard II, Phases 2 and 3; switchboard Cellular, Parts 1 and 2 and NIST 2004-2005 SRE. The JFA configuration was composed of 300 speaker factors and 100 channel factors with diagonal matrix in order to have speaker and common factors. Finally, the decision scores obtained for GMM-SVM and JFA-SVM were normalised using T-norm, and for JFA-CDS and JFA, they were normalised using ZT-norm (Z-norm followed by T-norm) and TZ-norm[14] (T-norm followed by Z-norm) respectively. We used 367 female T-norm models and 274 female Z-norm utterances from NIST 2004 and 2005 SRE respectively. Note that any utterances from speakers in NIST 2005 that appear in NIST 2006 have been excluded from the training set. The speaker verification results for all the baseline systems are shown in Table 5.1.

In the following subsections, results for various SRC systems will be presented and unless specified all optimisation was performed by the Gradient Projection for Sparse Reconstruction (GPSR) [193] MATLAB toolbox[15]. Alternatively, other freely available MATLAB toolbox including ℓ_1-magic [194], SparseLab [195] and l1_ls [196] can be used. During initial investigations, all toolboxes gave similar performance so GPRS was chosen as it is significantly faster, especially in large-scale settings [193]. Score normalisation (i.e T-norm) has been excluded from the SRC system because the conventional method of score normalisation (individual scoring against each T-norm model) slows down the verification process significantly (by a factor of three to six depending on the number of T-norm model and dictionary size) as compared with other systems (i.e SVM, CDS). Although a novel SRC-based T-norm has been proposed in [188] through the replacement of the T-norm data as the background samples in the over-

[14] Although the scores of JFA systems are usually normalised using ZT-norm (5.04%), we achieved slightly better performance with ZT-norm (4.96%).
[15] Gradient Projection for Sparse Reconstruction (GPSR) MATLAB toolbox is available online on http://www.lx.it.pt/~mtf/GPSR/

complete dictionary, no performance improvement was observed in the proposed method over the conventional T-norm as reported in [188]. In addition, the direct replacement of the background samples in the over-complete dictionary using T-norm data seems somewhat heuristic.

Table 5.1: Baseline speaker verification results on the NIST 2006 Female Subset database

Systems	EER (%)	minDCF
GMM-SVM	14.79	0.0760
GMM-SVM + NAP + T-norm	5.78	0.0285
JFA-SVM + LDA + WCCN + T-norm	5.39	0.0275
JFA-CDS + LDA + WCCN + ZT-norm	5.40	0.0270
JFA + TZ-norm	4.96	0.0251

5.4.2 Supervector-based SRC

The experiments carried out in this section compare the results of a GMM-SRC system with and without NAP with those of the GMM-SVM system. The dictionary $\mathbf{D}_{bg}$ matrix of SRC was composed of 2843 female utterances from the NIST 2004 SRE database, which was the same as the background training speaker database for SVM. The results based on supervectors with SVM and SRC are given in Table 5.1 and Table 5.2 respectively. We observed that by incorporating NAP compensation [12], the EER is improved significantly for both systems and SRC outperforms SVM if NAP is not incorporated. Comparing Tables 5.1 and 5.2, the GMM-SRC-NAP based classifier was able to achieve comparable results to the GMM-SVM-NAP system.

Table 5.2: Speaker verification results for supervector-based SRC on the NIST 2006 SRE Female Subset database

Systems	EER (%)	minDCF
GMM-SRC	11.21	0.0561
GMM-SRC + NAP	5.90	0.0334

On the other hand, when we compared SRC with supervector-based nearest neighbour (GMM-NN) (since SRC is considered a generalisation of NN [28]), a GMM-NN system with NAP incorporated (EER of 14.05%) was unable to achieve comparable performance to SVM/SRC. This supports the claim in [182, 197] that for high dimensional spaces, the concept of proximity, distance or nearest neighbour may not be qualitatively meaningful.

5.4.3 Speaker Factor-based SRC

As shown in the previous section, supervector-based SRC is able to achieve comparable performance to GMM-SVM. However the sparse representation of large dimension supervectors requires a large amount of memory due to the over-complete dictionary, which can limit the training sample numbers and could slow down the recognition process. In this section, we evaluate the JFA-SRC system (detailed in Section 5.3.2) in comparison with JFA-SVM. Furthermore, we tried various channel compensation steps in the speaker factor space that are reported in [132] and the best performance for JFA-SRC was found to be based on LDA (JFA-SRC-LDA, dim = 300) with an EER of 7%. However, the result indicates that the initial performance of the JFA-SRC is significantly poorer than that of JFA-SVM and JFA-CDS. In the following sub-sections, we investigate some techniques presented in [33, 132, 188, 198] with a view to improving the system performance.

5.4.3.1 Robustness to Corruption

In many practical recognition scenarios, the test sample **S** can be partially corrupted due to large session variability. Thus it has been suggested in [28, 33, 188] to introduce an error vector $\boldsymbol{e}$ into the linear model in equation (5.8) as follows

$$\mathbf{S} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{e} = [\mathbf{D}\ \mathbf{I}]\begin{bmatrix}\boldsymbol{\gamma}\\ \mathbf{e}\end{bmatrix} \doteq \mathbf{B}\boldsymbol{w} \tag{5.8}$$

Here, $\mathbf{B} = [\mathbf{D}, \mathbf{I}] \in \mathbb{R}^{K\times(N+K)}$ where the system is always underdetermined. As before, the sparsest solution $\boldsymbol{w}$ by solving the following extended ℓ_1-minimization problem

$$\begin{gathered} \widehat{\boldsymbol{w}} = \min\|\boldsymbol{w}\|_1 \ \ subject\ to\ \mathbf{S} = \mathbf{B}\boldsymbol{w} \\ \widehat{\boldsymbol{w}} = [\hat{\boldsymbol{\gamma}}^t \ \ \hat{\mathbf{e}}^t]^T \in \mathbb{R}^{N+K} \end{gathered} \tag{5.9}$$

If the error vector $\mathbf{e}$ is sparse and has no more than $\frac{K+l_{tar}}{2}$ nonzero entries, the new sparse solution $\widehat{\boldsymbol{w}}$ is the true generator [28]. Finally, the decision criterion in equation (5.6) is used for verification.

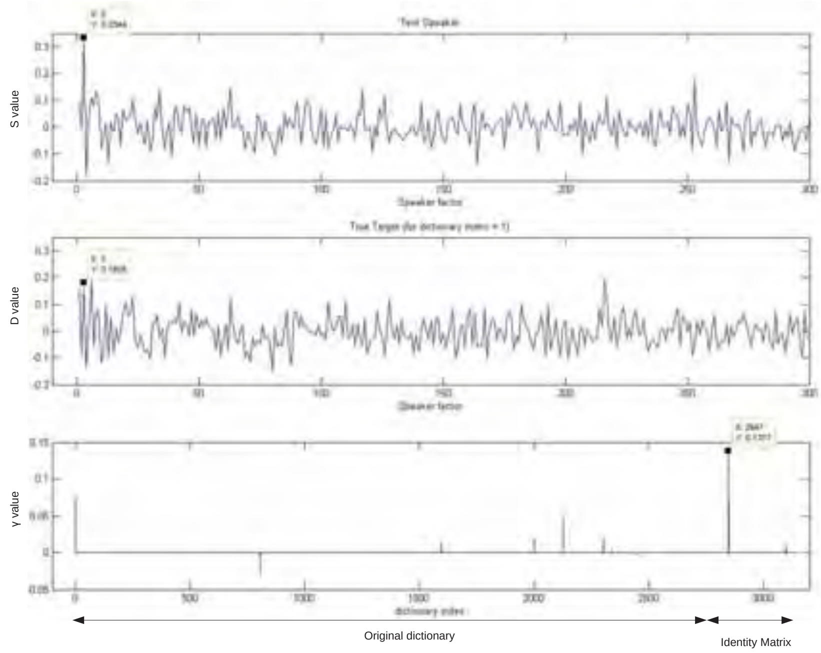


Figure 5.6 Illustration of inclusion of identity matrix (a) Test speaker's speaker factor (b) Target speaker's speaker factor (c) Sparse solution $\boldsymbol{w}$ with identity matrix included

Here we briefly illustrate the effect of including the identity matrix in the overcomplete dictionary and show the incremental improvement in accuracy. An example speaker from the NIST 2006 database was chosen, such that the test speaker's speaker factor had a large outlier in the third dimension relative to its training speaker factor, as shown in Figure 5.6(a) and (b) respectively. It has been reported in [28, 199] that the identity matrix will capture any redundancy between the test sample and dictionary, hence the outlier is captured by the identity matrix at the location corresponding to the third dimension in this example, for an original dictionary size of $N = 2844$ as shown in Figure 5.6(c). The inclusion of the identity matrix in the dictionary improves the recognition performance from 7% to 6.4% EER. The improvement supports the claim in [28, 33, 188] that by adding a redundant identity matrix at the end of the original over-complete dictionary, the sparse representation is more robust to variability. Therefore, in subsequent experiments, the identity matrix is included as a part of the overcomplete dictionary.

5.4.3.2 Sparseness Constraint

The use of SRC for speech classification and recognition tasks [34, 198, 200] has become increasingly popular in recent years. However, little analysis has been done on the appropriateness of different types of sparsity regularisation constraints in speech processing applications. One such study was conducted by Kanevsky et al., wherein a comparative study across different sparseness methods in terms of classification performances for speech recognition were conducted [198]. The sparseness methods investigated include the LASSO [201] and Bayesian Compressive Sensing (BCS) [202] that use an ℓ_1 sparseness constraint (known as a Laplacian prior), Elastic Net [203] and Cyclic Subgradient Projections (CSP) [204] which use a combination of an ℓ_1 and ℓ_2 (Gaussian prior) constraint and Approximate Bayesian Compressive Sensing (ABCS) [34] that uses an ℓ_1^2 constraint, which is known as a Semi-Gaussian prior. It was found that the methods based on a combination of an ℓ_1 and ℓ_2 constraint (i.e. Elastic Net, CSP and ABSC) gave the best classification accuracies. Given the encouraging performance reported in [198], the investigation of the appropriateness of different types of sparsity regularisation constraints for speaker recognition (which is new in the context of speaker recognition) will be conducted in this section. Since ℓ_1 sparsity constraint coupled with an ℓ_2 norm showed almost similar results, Elastic Net (which gave the best performance reported in [198]) was selected for comparison in this section. It can be formulated as follows:

$$\min_{\boldsymbol{w}} \|\mathbf{S} - \mathbf{D}\boldsymbol{w}\|_2 + \lambda \|\boldsymbol{w}\|_1 + (1-\lambda)\|\boldsymbol{w}\|_2^2 \,, \;\; where\ \lambda \in [0,1) \tag{5.10}$$

where $\lambda\|\boldsymbol{w}\|_1 + (1-\lambda)\|\boldsymbol{w}\|_2^2$ is termed the elastic net penalty, which is a convex combination of the LASSO and ridge regression [205]. Ridge regression is an exemplar-based technique that uses information about all training examples in the dictionary to make a classification decision about the test example, in contrast to sparse representation techniques that constrain $\boldsymbol{w}$ be sparse. When $\lambda = 0$, the naïve elastic net penalty becomes simple ridge regression and when $\lambda = 1$, it becomes LASSO. In this section, Elastic Net is implemented using the Glmnet MATLAB package[16] [206] with $\lambda = 0.8$ since it gave the best EER as shown in Figure 5.7.

[16] MATLAB implementation of Glmnet is available online on http://www-stat.stanford.edu/~tibs/glmnet-matlab/.

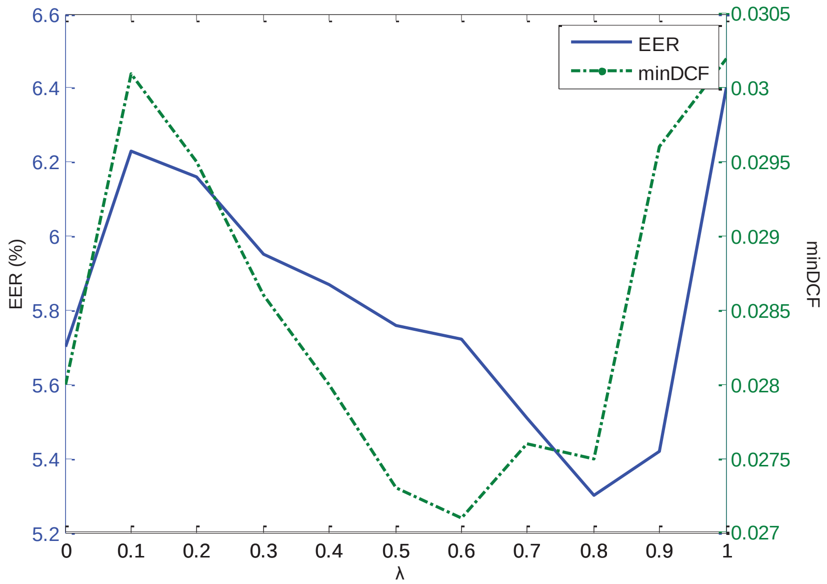


Figure 5.7 Speaker recognition performance (EER: left y-axis, solid line and minDCF: right y-axis, dash-dot line) on NIST 2006 as the elastic net penalty, λ, is refined.

As shown in Figure 5.7 and Table 5.3, the method using only ℓ_1 norm or ℓ_2 norm has slightly lower accuracy, showing the decrease in accuracy when a high or low degree of sparseness is enforced respectively (similar results are observed in [198]). Thus, it appears that using a combination of a sparsity constraint on $\boldsymbol{\gamma}$, coupled with an ℓ_2 norm, does not force unnecessary sparseness and offers the best performance, comparable with JFA-SVM and JFA-CDS.

Table 5.3: Speaker verification results for different types of sparsity regularisation constraints on the NIST 2006 SRE Female Subset database

Systems	EER (%)	minDCF
JFA-SRC-LDA with ℓ_1-constraint	6.40	0.0302
JFA-SRC-LDA with ℓ_2-constraint	5.71	0.0280
JFA-SRC-LDA with ℓ_1 and ℓ_2 -constraint	5.30	0.0275

5.4.3.3 Proposed Dictionary Design

In recent years, apart from the study of different pursuit algorithms for sparse representation, the design of dictionaries to better fit a set of given signals has attracted growing attention [207-210]. We briefly tried the K-SVD[17] algorithm for training an overcomplete dictionary that best suits a set of given signals proposed in [207]. However no improvement/degradation was observed since the aforementioned method aims at better representing the signals with respect to a tuned dictionary rather than classifying them.

As mentioned previously (in section 2.4.2), McLaren et al. [38] proposed SVM background speaker selection algorithms for speaker verification. In this section, a similar but novel idea, which we termed column vector frequency, is considered for choosing the dictionary of SRC based on the total number of times each individual column of the background dictionary $(\mathbf{D}_{bg})$ is chosen, as shown in (5.11)

$$\mathbf{D}_{bg} = \left[\mathbf{y}_{bg,1}\ \mathbf{y}_{bg,2}\ \cdots\ \mathbf{y}_{bg,l_{bg}}\right]$$
$$\mathfrak{V}(\mathbf{y}_{bg,t}) = \sum_{c=1}^{P} \mathfrak{M}(\alpha_{bg,t}^{c}), \qquad \text{where } \mathfrak{M}(x) = \begin{cases} 1, x \neq 0 \\ 0, x = 0 \end{cases} \tag{5.11}$$

where t is the column index of the background dictionary with values from 1 to l_{bg}, P is the number of test trials and $\mathfrak{V}$ is the frequency counter for the corresponding t^{th} column.

First, the results for using a number of different dictionary dataset configurations without any background speaker selection (with ℓ_1+ℓ_2 constraint, $\lambda = 0.8$) are detailed in Table 5.4. It can be observed that using the NIST 2004 dataset alone gave the best performance, which is the same as the results reported for SVM in [110]. Combining the NIST 2004 dataset with NIST 2005 resulted in the degradation of EER performance despite the significant increase in the number of background speaker examples.

[17] The K-SVD algorithm is implemented with Matlab toolbox available online on http://www.cs.technion.ac.il/~elad/software/

Table 5.4: Speaker verification results for different dictionary datasets on the NIST 2006 SRE Female Subset

Dictionary	EER	minDCF
NIST 2004	5.30%	0.0275
NIST 2005	5.64%	0.0291
NIST 2004 + NIST 2005	5.63%	0.0268

As an initial indicator of whether the column vector frequency is an adequate metric to represent the suitability of a background speaker, the 500 highest ranked and 500 lowest ranked background speakers from the NIST 2004 (2843 speakers) and NIST 2005 (673 speakers) datasets based on column vector frequency were selected on gender-dependent basis and the evaluation results are detailed in Table 5.5. The performance demonstrates that the dictionary basis chosen based on column vector frequency is an appropriate measure of the impostor example. Furthermore, to determine an optimal size for the dictionary, the experiment was repeated using only the highest R column vector frequencies with R varying from 300 to 3516 in steps of 200. The resulting EER was 5.3% for all values of R and minDCF of 0.0275 or 0.0276 for $R \geq 500$ (R=300 has a minDCF of 0.0285, R=250 has an EER of 5.37% and minDCF of 0.0289) as shown in Figure 5.8(a), indicating that a smaller size dictionary can be used. In addition, a 79% relative reduction in computation time is achieved using the refined dictionary over the full dictionary (as shown in Figure 5.8(b)), allowing a faster verification process. The refined dictionary with R=500 will be used for all subsequent experiments and will be shown to generalise well to the NIST 2010 dataset in Section 5.5. On the other hand, despite the significant improvement in time, the SRC is still much slower than the JFA-SVM (1800s) and JFA-CDS scoring (244s).

Table 5.5: Speaker verification results on NIST 2006 Female Subset trials when using SRC background datasets refined by impostor column vector frequency.

Dictionary	EER	minDCF
Full Dataset	5.63%	0.0268
500 highest frequency	5.30%	0.0276
500 lowest frequency	6.50%	0.05

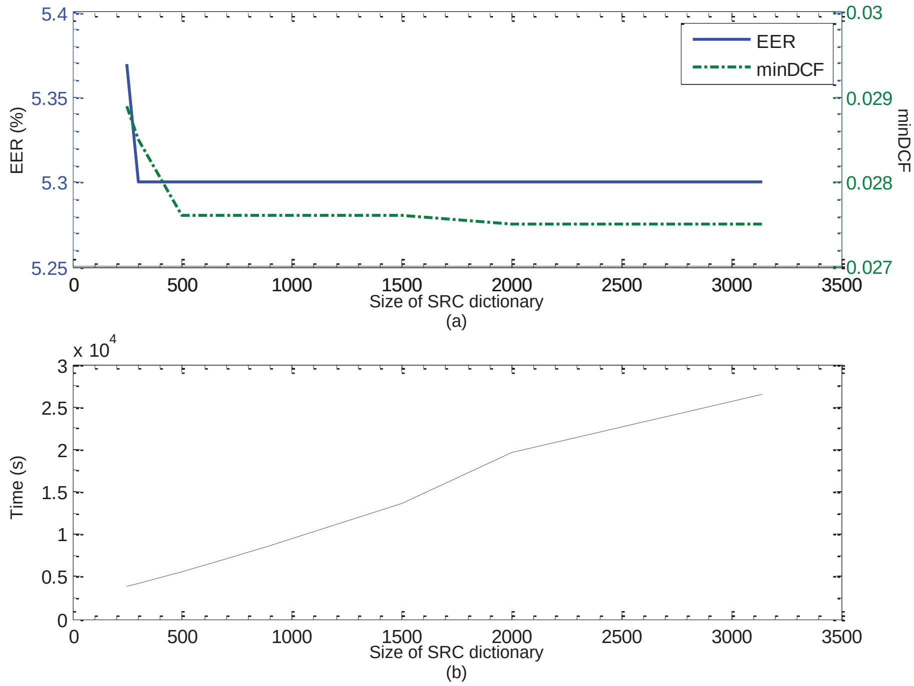


Figure 5.8 Speaker recognition performance on NIST 2006 as the SRC dictionary is refined. (a) EER (left y-axis, solid line) and minDCF (right y-axis, dash-dot line) (b) Total time taken (in seconds) for computing the ℓ_1-norm score across all test utterances.

5.4.3.4 Related Work

Recently[18] similar work has been conducted by Li et al. [188] using i-vectors as features for SRC with LDA and WCCN incorporated for channel variability compensation, in which the focus is on enhancing the robustness and performance of speaker verification through the concatenation of a redundant identity matrix at the end of the original over-complete dictionary. They also propose new scoring measures: ℓ_1-norm ratio, ℓ_2-residual ratio and background normalised (Bnorm) ℓ_2-residual (as shown in equations (5.12) to (5.14) respectively) and a simplified T-norm procedure for SRC system by replacing the

[18] Indeed, in parallel with this thesis work.

dictionary with T-norm i-vectors. Although three different decision criteria are proposed in [188], our experiments showed that simply using the target ℓ_1-norm (as shown in equation (5.6)) gave the best minDCF and comparable/better EER to the ℓ_1-norm ratio, ℓ_2-residual ratio and Bnorm ℓ_2-residual as shown in Table 5.6.

$$\ell_1 - norm\ ratio = \|\delta_{tar}(\boldsymbol{\gamma})\|_1 / \|\boldsymbol{\gamma}\|_1 \tag{5.12}$$

$$\ell_2 - residual\ ratio = \frac{\left\|\mathbf{S} - \mathbf{D}\left(\sum \delta_{bg}(\boldsymbol{\gamma})\right)\right\|_2}{\|\mathbf{S} - \mathbf{D}\delta_{tar}(\boldsymbol{\gamma})\|_2} \tag{5.13}$$

$$\begin{gathered} Bnorm\ \ell_2 - residual = \frac{-\|\mathbf{S} - \mathbf{D}\delta_{tar}(\boldsymbol{\gamma})\|_2 - mean(\phi)}{std(\phi)} \\ \phi_j = -\left\|\mathbf{S} - \mathbf{D}\delta_j(\boldsymbol{\gamma})\right\|_2;\ j = bg\ index \end{gathered} \tag{5.14}$$

Table 5.6: Speaker verification performance for different scoring measures (with respect to configurations used for result in Table 5.5) on the NIST 2006 SRE database (female subset).

Scoring measure	EER (%)	minDCF
ℓ_1-norm ratio **[188]**	5.34	0.0285
ℓ_2-residual ratio **[188]**	5.73	0.0328
Bnorm ℓ_2-residual **[188]**	5.68	0.0328
ℓ_1-norm	5.30	0.0276

Next, we compare the results reported in this thesis with the best baseline system configuration reported in [188] which is based on ℓ_1–minimisation with ℓ_1-constraint[19], inclusion of identity matrix, Bnorm-(ℓ_2-residual) scoring and T-norm (conventional). Using these configurations on NIST 2006 SRE database (female subset), an EER of 6.11% and minDCF of 0.0302 was achieved. It could be observed that similarly to other classifiers, incorporating T-norm does improve the EER performance (from 6.4%). Furthermore, comparing the above mentioned results (proposed system by Li et al. [188]) with the proposed system reported in this thesis (Table 5.3 and Table 5.5), we observed that sparse representation based on a combination of ℓ_1 and ℓ_2 constraint on $\boldsymbol{\gamma}$

[19] The ℓ_1-constraint refers to the constraint on $\boldsymbol{\gamma}$ (as discussed in section 5.4.3.2) and not the quadratic constraints on the error tolerance as indicated in [188].

outperformed the former system significantly, with a relative EER reduction of 13.25%. This improvement seems to be mainly attributable to the degree of sparseness constraint on $\boldsymbol{\gamma}$. In addition, a faster verification process with no deterioration in performance can be achieved with a smaller dictionary refined based on column vector frequency, as opposed to the direct heuristic replacement of the dictionary with T-norm samples in [188].

5.4.4 Fused Speaker Verification Results

In this section, we will explore whether SRC provides complementary information to the conventional baseline introduced in Section 5.4.1, since the study of systems which fuse well has held sustained interest in the speaker recognition community in recent times [96]. The best performing configuration for each individual system was chosen. Since the base classifier scores may have different interpretations (e.g. log-likelihood ratios, SVM inner products, cosine distance and regression coefficients) and their scales may vary a lot, it is important to equalise their global range to avoid the large-variance base classifier dominating the fused score [211]. Shown in Figure 5.9 (a) – (d) are the score distributions of GMM-SVM, JFA, JFA-SVM and JFA-CDS, from which Gaussian distributions may be observed, whereas JFA-SRC (shown in Figure 5.9 (e)) has an approximately exponential distribution with the majority of scores concentrated near zero. The same observation was reported as future work in [188]. In this thesis, we propose the use of the S-calibration (Scal) [143], to discriminatively train a mapping to convert arbitrary scores to well-calibrated log-likelihood ratios (LLRs) so that all of the scores maps to the same distribution and global range.

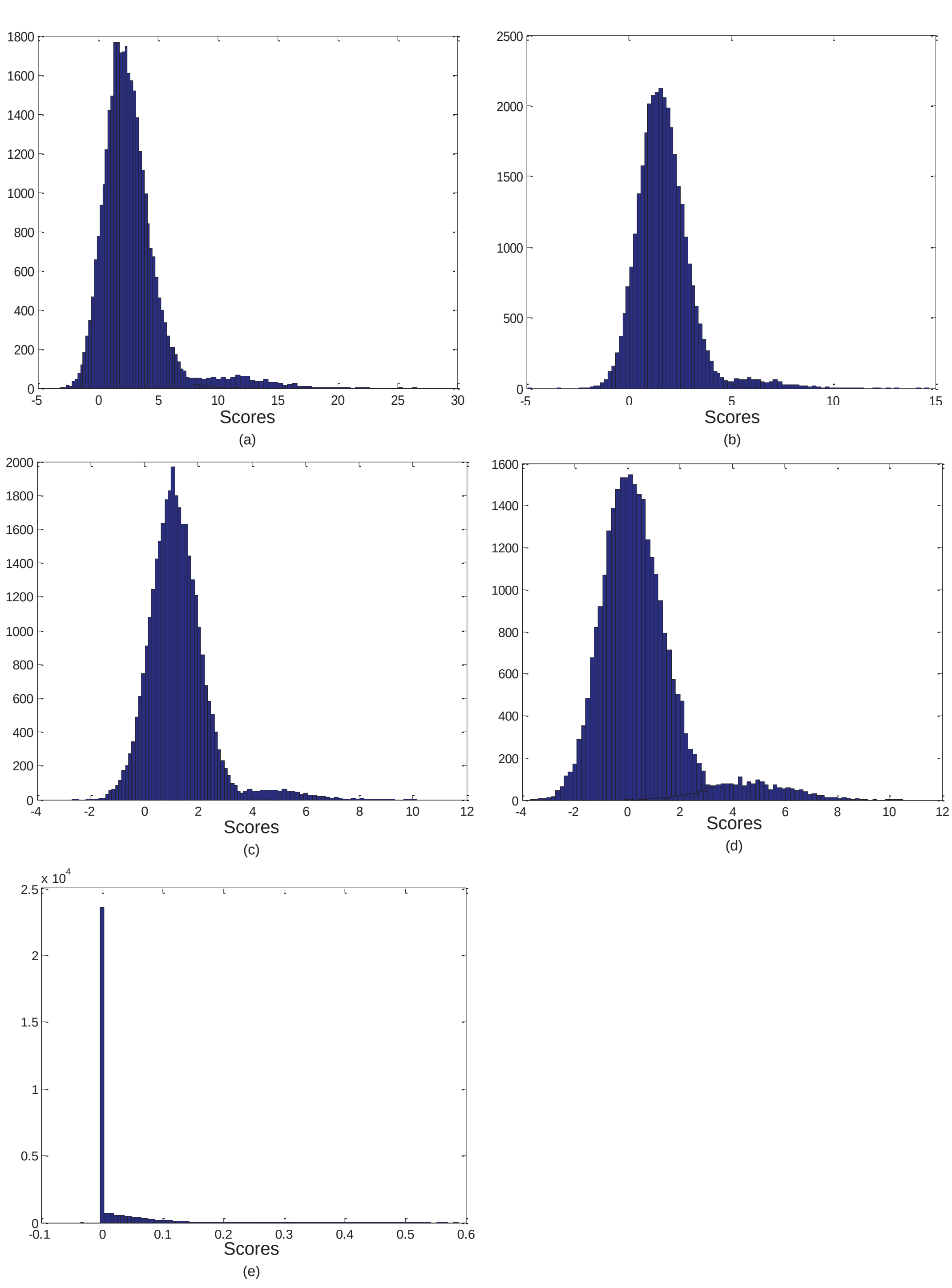


Figure 5.9 Scores distributions for (a) GMM-SVM (b) JFA (c) JFA-SVM (d) JFA-CDS (e) JFA-SRC

In Table 5.7, we see a significant improvement from fusing the proposed sparse representation systems (GMM-SRC and JFA-SRC) with the current state-of-the-art JFA baseline in terms of EER value. Both proposed SRC systems are able to achieve comparable performance to the fusion of SVM-based systems with JFA. The fusion of

JFA and JFA-SRC achieved the best result with 4.50% EER, i.e. a 9.27% relative EER reduction (c.f. JFA alone at 4.96% EER). Moreover, JFA-SRC also fused well with GMM-SVM bringing the EER down to 4.58% from 5.3% (relative reduction of 13.58%). This could be attributed to different representation of speaker information variation (i.e. speaker factor and supervectors) and classifiers.

Table 5.7: Fused speaker verification performance on the NIST 2006 SRE database (female subset) with speaker detection cost model parameters of $C_{Miss} = 10$, $C_{FalseAlarm} = 1$, $P_{Target} = 0.01$ (EERx100)

System	Individual Performance		Fused Performace									
			GMM-SRC		**JFA-SVM**		**JFA-SRC**		**JFA-CDS**		**JFA**	
	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
GMM-SVM	5.78	0.0285	5.69	0.0283	4.73	0.0244	4.58	0.0242	4.76	0.0246	4.54	0.0238
GMM-SRC	5.90	0.0334	-	-	4.88	0.0256	5.14	0.0261	4.89	0.0246	4.75	0.0247
JFA-SVM	5.39	0.0275	-	-	-	-	5.11	0.0256	5.32	0.0266	4.58	0.0232
JFA-SRC	5.30	0.0276	-	-	-	-	-	-	5.10	0.0246	**4.50**	0.0244
JFA-CDS	5.40	0.0270	-	-	-	-	-	-	-	-	4.57	**0.0226**
JFA	4.96	0.0251	-	-	-	-	-	-	-	-	-	-

5.5 Speaker Recognition Experiments on NIST 2010 SRE

5.5.1 Experimental setup

In this section, the classifiers were evaluated using the larger and more contemporary NIST 2010 database, in order to see the database independency of the results. Results are reported for the five evaluation conditions with normal vocal effort, corresponding to det conditions 1-5 in the SRE’10 evaluation plan [145], which include *int-int*, *int-tel*, *int-mic* and *tel-tel*.

We used exactly the same UBM and joint factor analysis configuration (300 speaker factors, 100 channel factors) as in Section 5.4. The only difference lay in the amount of data used to train the JFA hyperparameters, NAP, WCCN, LDA and SVM impostor with respect to the evaluation conditions. We added the Mixer 5 and interview data taken from the follow-up corpus of the NIST 2008 SRE for interview (*int*) conditions, NIST 2005 and 2006 SRE microphone segments for microphone (*mic*) conditions and NIST 2006 SRE for telephone (*tel*) conditions. Table 5.8 summarises the datasets used to estimate our system parameters. Similarly to the previous setup (in Section 5.4.1), any common utterances from speakers in the NIST 2008 follow up and NIST 2010 databases have been excluded from the training set.

Table 5.8: Corpora used to estimate UBM, JFA hyperparameters, WCCN, LDA, SVM impostors, Z- and T-norm data for evaluation on the NIST 2010 SRE.

		Switchboard II	Mixer 5	NIST 2004	NIST 2005	NIST 2006	NIST 2008 follow up
UBM				x			
JFA	V	x		x			
	D			x			
	U		x	x	x	x	x
T-norm				x			
Z-norm					x		
NAP			x	x	x	x	x
WCCN			x	x	x	x	x
LDA		x	x	x	x	x	x
SVM-Impostor				x	x	x	x

5.5.2 Single-system Speaker Verification Results

The performance of each classifier for each condition is given in Table 5.9. The results show that JFA obtained the best performance, followed by JFA-SRC ($\lambda = 0.8$) in the speaker factor space, JFA-SVM and JFA-CDS in the speaker factor space and GMM-

SVM, all of which are consistent with the findings reported in [189]. Interestingly, the JFA-SRC gave the best minDCF for all conditions. From Table 5.9, the JFA-SRC approach performs better than all SVM variants in all conditions with just a single dictionary, designed according to the column vector frequency (X = 500) in Section 5.4.3.3, which indicates that the dictionary generalises well to different type of common conditions. On the other hand, for SVM-based systems, it has been observed that different background data sets need to be constructed separately for different conditions (i.e *int-int, int-tel, int-mic and tel-tel*) to achieve good performance (where the results with best configuration are reported in Table 5.9), which is similar to the observations in [13, 14, 212].This is most probably because the SVM system relies heavily on the background observations to provide most of the observable discriminatory information. The background dataset must, therefore, consist of suitable impostor examples to ensure good classification performance [38] as opposed to SRC, which relies less on model selection as discussed in [175]. On the whole, the experiment shows that the sparse representation approach can outperform the best performance achieved by SVM. In addition, the JFA-SRC outperforms the JFA-CDS, which is of interest since both do not require a training phase and additionally do not require any form of score normalisation to achieve good performance. On the whole, the experiment shows that the sparse representation approach can outperform the best performance achieved by SVM and CDS without the need for score normalisation as shown in Table 5.9.

Table 5.9: Speaker verification performance on the NIST 2010 SRE database with speaker detection cost model parameters of $C_{Miss} = 1$, $C_{FalseAlarm} = 1$, $P_{Target} = 0.001$ (EERx100, minDCFx1000)

Common Condition	GMM-SVM		JFA		JFA-SRC		JFA-SVM		JFA-CDS	
	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
1 (*int-int*)	4.81	0.548	**3.86**	0.566	4.00	**0.515**	4.32	0.536	3.88	0.619
2 (*int-int*)	7.71	0.672	**5.58**	0.696	6.10	**0.615**	7.04	0.664	6.19	0.709
3 (*int-tel*)	4.64	0.587	**4.61**	0.654	4.65	**0.586**	4.70	0.644	4.80	0.689
4 (*int-mic*)	6.31	0.704	**5.00**	0.774	5.40	**0.642**	6.04	0.672	5.61	0.684
5 (*tel-tel*)	3.66	0.530	3.38	0.549	**3.20**	**0.436**	3.74	0.498	4.08	0.583

5.5.3 Fused Speaker Verification Results

The fused results of the baseline system (JFA) with JFA-SVM, JFA-CDS or JFA-SRC are shown in Table 5.10. The results are consistent with those shown in Section 5.4.4, demonstrating that the fusion of JFA and JFA-SRC is better than the fusion of JFA and JFA-SVM, and the fusion of JFA and JFA-CDS. The fusion of JFA and JFA-SRC achieves an improvement of 5.4-18% relative reduction in EER and 2.4%-30% relative reduction in minDCF over the baseline as shown in Figure 5.10. Furthermore, the fusion of JFA and JFA-SRC gave a relative improvement on the fusion of JFA and JFA-SVM by 1.3-7.6% in terms of EER and 0.7-24% in terms of minDCF. . In contrast, the fusion of JFA-SRC and JFA-SVM (shown in Table 5.10) results in minimal improvement in EER since both of the classifiers have almost similar classification decisions for most of the test point as explained in Section 5.2.

Table 5.10: Fused speaker verification performance of JFA-SVM, JFA-CDS or JFA-SRC with JFA on the NIST 2010 SRE database with speaker detection cost model parameters of $C_{Miss} = 1$, $C_{FalseAlarm} = 1$, $P_{Target} = 0.001$ (EERx100, minDCFx1000)

System	Common Condition 1		Common Condition 2		Common Condition 3		Common Condition 4		Common Condition 5	
	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
JFA-SVM + JFA	3.19	0.475	5.21	0.593	4.10	0.564	**4.73**	0.661	3.01	0.450
JFA-CDS + JFA	**3.04**	0.442	**5.03**	0.608	3.95	0.568	4.77	0.653	3.16	0.513
JFA-SRC + JFA	3.15	**0.360**	**5.03**	**0.571**	**3.79**	**0.560**	**4.73**	**0.617**	**2.82**	**0.395**
JFA-SVM + JFA-SRC	3.70	0.459	5.99	0.585	4.60	0.580	5.33	0.638	3.20	0.411

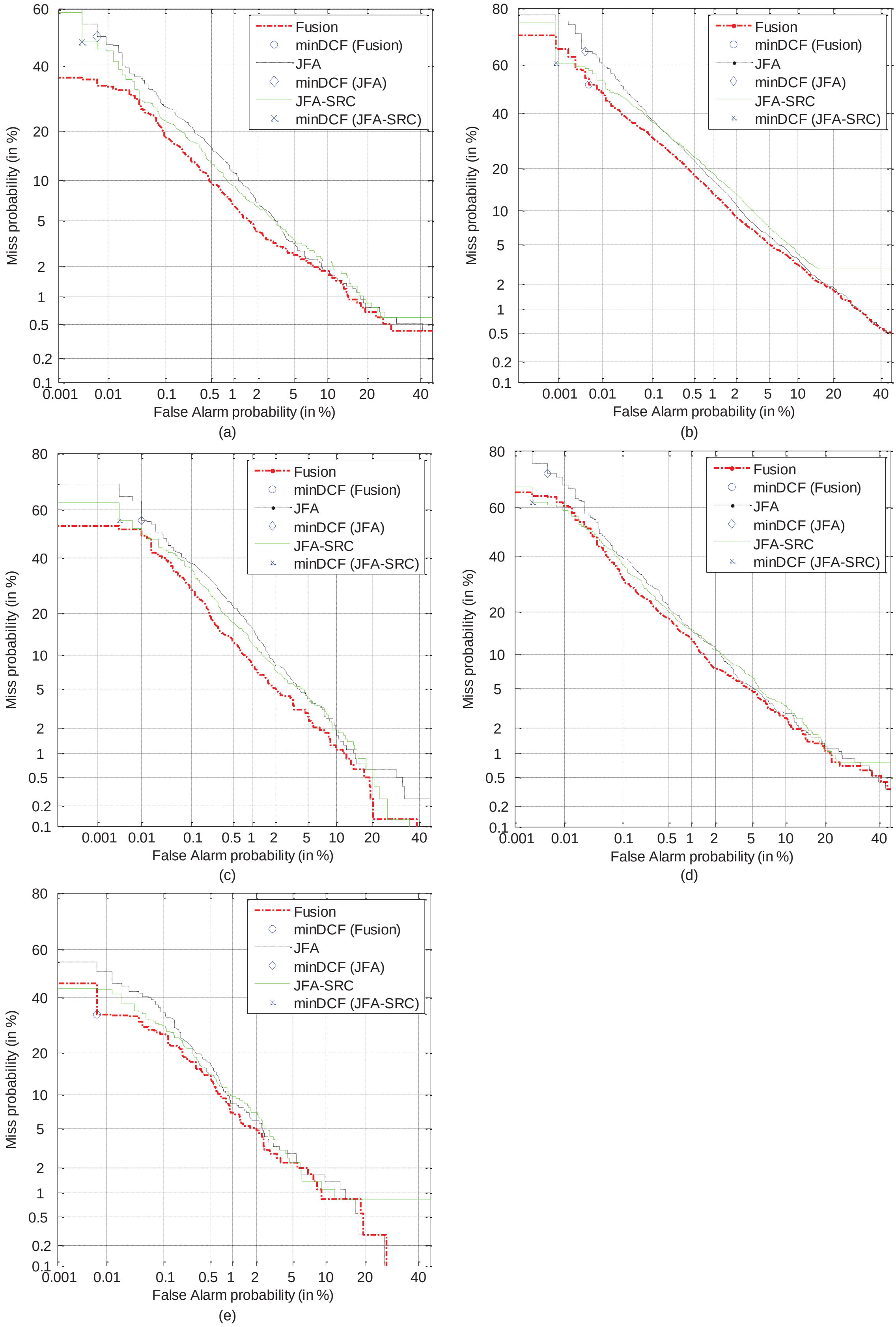


Figure 5.10 DET curves showing the speaker recognition results of JFA and JFA-SRC on the NIST 2010 SRE database for Condition 1 – 5 as shown in (a) – (e) respectively

5.5.4 Complementary Information of Features and Classifiers

Up until now, the proposed features and proposed classification methods presented in Chapter 3 and this chapter respectively have been analysed separately. It is expected that the features and classification contain complementary information, which may lead to an improvement in verification performance when they are combined. In this section, all possible pairwise combinations of MFCC, LogLSGD, SCM and SCF with JFA or JFA-SRC are evaluated on the NIST 2010 SRE database (Condition 5) and fused at the score level after S-calibration. The recognition performance of the features and classification methods, without fusion, are given in Table 5.11. Similar to the results reported in section 5.5.2, we observed that the JFA-SRC achieved better minDCF than the JFA for each individual feature. Furthermore, it can be observed that the trends of the individual performance were consistent with those reported in the foregoing chapters, with MFCC being the best performing feature, followed by SCM, SCF and LogLSGD.

Table 5.11: Speaker verification results of individual systems based on various features and classification when evaluated on the NIST 2010 SRE database (Condition 5)

Systems	EER (%)	minDCF
MFCC JFA	3.38	0.549
MFCC JFA-SRC	**3.20**	**0.436**
SCF JFA	3.81	0.608
SCF JFA-SRC	4.50	0.512
SCM JFA	3.50	0.614
SCM JFA-SRC	4.22	0.589
LogLSGD JFA	4.78	0.647
LogLSGD JFA-SRC	5.39	0.629

Table 5.12 summarises the recognition performance of the various features and classification methods after fusion. In general, it can be observed that the recognition performance improved for all pairwise combinations. Notably, the fusion of MFCC JFA-SRC and SCF JFA gave the best fused performance, with a relative EER reduction of

38% and relative minDCF reduction of 25% to MFCC JFA-SRC, outperforming the fusion of MFCC JFA and MFCC JFA-SRC (in section 5.5.3).

In summary, the recognition results in this section not only indicate that SRC-based classifiers contain complementary information, they also show that recognition performance can be *significantly* improved through score-level fusion of different features and classification methods.

Table 5.12: Speaker verification results of fused systems based on various features and classification when evaluated on the NIST 2010 SRE database (Condition 5)

	Individual Performance		Fused Performance													
Systems			**MFCC JFA-SRC**		**SCF JFA**		**SCF JFA-SRC**		**SCM JFA**		**SCM JFA-SRC**		**LogLSGD JFA**		**LogLSGD JFA-SRC**	
	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
MFCC JFA	3.38	0.549	2.82	0.395	2.31	0.442	3.08	0.419	2.34	0.467	3.09	0.405	2.81	0.470	3.09	0.462
MFCC JFA-SRC	**3.20**	**0.436**	-	-	**1.97**	**0.327**	2.85	0.338	**1.97**	0.346	2.92	0.346	2.41	0.354	3.09	0.338
SCF JFA	3.81	0.608	-	-	-	-	3.66	0.490	2.78	0.579	3.09	0.536	2.95	0.574	3.45	0.485
SCF JFA-SRC	4.50	0.512	-	-	-	-	-	-	2.50	0.528	3.90	0.467	3.61	0.490	3.38	0.495
SCM JFA	3.50	0.614	-	-	-	-	-	-	-	-	3.38	0.545	2.81	0.571	3.17	0.503
SCM JFA-SRC	4.22	0.589	-	-	-	-	-	-	-	-	-	-	3.38	0.566	3.66	0.523
LogLSGD JFA	4.78	0.647	-	-	-	-	-	-	-	-	-	-	-	-	4.55	0.560
LogLSGD JFA-SRC	5.39	0.629	-	-	-	-	-	-	-	-	-	-	-	-	-	-

5.6 Summary

In this chapter, we investigated the discriminative nature of the sparse representation classification (SRC) for speaker verification using supervectors and speaker factors from the joint factor analysis using GMM-sparse representation classification (GMM-SRC) and joint factor analysis-sparse representation classification (JFA-SRC) systems respectively. Inspired by the principles of the sparse representation model and based on the intuitive hypothesis that a speaker can be represented by a linear combination of training samples from the same speaker, we first compute the sparse representation through ℓ_1-minimisation and classification is achieved based on target ℓ_1-norm. Our initial investigation with GMM-SRC showed promising results. Then, we proposed the inclusion of inter-session variability compensation, NAP, on the supervectors before sparse representation classification resulting in an approach we termed GMM-SRC-NAP to take into account the effect of inter-session variation in NIST SRE database. Although the GMM-SRC-NAP was able to achieve comparable performance to GMM-SVM-NAP system without the need for a training phase (training of a target model before scoring), it has a significantly slower recognition process as compared with the GMM-SVM-NAP system due to size of the over-complete dictionary.

In an attempt to resolve the problem as mentioned above, we instead adopted the speaker factors from the JFA model as feature vectors for the SRC resulting as an approach we termed JFA-SRC. However the initial evaluation of JFA-SRC did not show promising results. Given that SRC has only recently appeared in the context of speaker recognition, we evaluated a range of existing techniques for sparse representation classification and examined the effect on speaker recognition performance. The techniques considered include the augmentation of the dictionary with an identity matrix and a sparseness method that uses a combination of $\ell 1$ and $\ell 2$ minimisation (Elastic net). A combination of both techniques achieved a 24% relative reduction in EER over a SRC system based on $\ell 1$ minimization and without identity matrix, suggesting that a high degree of sparseness (by $\ell 1$ constraint) leads to a decrease in accuracy.

Then, motivated by background speaker selection for the SVM-based system, we proposed the SRC background dataset selection based on column vector frequency. Although no improvement in terms of EER was observed, we demonstrated that a smaller dictionary refined by column vector frequency could be used, allowing a faster verification process (79% relative reduction in computation time). Furthermore, we showed that the dictionary chosen for development on NIST 2006 SRE generalised well to the evaluation on NIST 2010 SRE corpus for different evaluation condition as opposed to SVM background data, which requires significant amount of tuning based on the evaluation condition.

Experimental results on the NIST 2010 database validated the findings that the sparse representation approach can match and/or outperform the best performance achieved by SVM. The fusion of JFA-SRC and JFA system gave a relative reduction in EER of 5.4 – 18% over JFA alone, and the fusion of JFA with JFA-SRC outperformed the fusion of JFA with JFA-SVM in the range of 1.3-7.6% relative reduction in EER and 0.7-24% in minDCF.

Finally, this chapter evaluated the performance of various complementary/alternative features (LogLSGD, SCM and SCF) and classification methods (JFA-SRC) discussed in this thesis on the NIST 2010 SRE database. The fused results demonstrated that the fusion of systems with different features and classification improve the verification performance significantly, strongly supporting the hypothesis that the features and classifiers carry complementary information.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis has reported research conducted into automatic speaker recognition with the aim of: (i) investigating and developing complementary features for magnitude-based speaker recognition systems; (ii) developing an understanding of the relative contributions of the acoustic and speaker modelling ‘stages’ and the benefits brought by fusing systems based on different acoustic features; (iii) exploring different classification approaches to verify speakers.

6.1.1 Investigation of Novel Features

In chapter 3 of this thesis, we investigated and developed phase and frequency based features to complement the magnitude information captured by features such as MFCC. In regards to phase-based feature, an alternative group delay features regularised using a least squares approach termed the log-compressed least squares group delay (LogLSGD) was proposed. Interestingly, the experimental results indicated that the proposed LogLSGD is a simple and effective way to reduce the dynamic range of modified group delay (MODGD) features (caused by strong excitation components which mask the formant peaks in group delay function), by alleviating the ill-conditioning of the MODGDF calculation and also removes the need for data-dependent empirical parameters in the GD feature extraction algorithm. Furthermore, LogLSGD improved on a 5.09% EER MFCC baseline to 4.54% after fusion on the NIST 2006 SRE database.

These results not only showed that group delay carry complementary information to MFCC but also indicates that the performance of the group delay could be improved significantly when the masking effect of the strong excitation components are suppressed.

Next, two spectral centroid features, namely spectral centroid frequency (SCF) and spectral centroid magnitude (SCM) were presented in section 3.2. The SCF was recently used for speaker recognition to complement magnitude-based features (MFCC) but with a little success. In our research, we proposed an improved implementation of SCF by increasing the number of FFT points (to 2048 FFT points) and extracting it using Bark scale Gabor filterbank, which was found to have a relative improvement of 22% in terms of EER on the SCF performance, significantly outperforming the previous extraction configuration (with 160 FFT points and mel scale triangular filterbank). Then, a complementary feature to SCF, which captures the distribution of energy in a subband, termed the Spectral Centroid Magnitude was proposed. Experimental results showed that the fusion of the SCF-based system and SCM-based system outperformed a solely MFCC-based system on the NIST 2006 SRE database (relative improvement of 13.5% in terms of EER). This is most likely because SCM and SCF collectively capture information concerning both the average energy (similar to MFCC) and the distribution of energy (frequency bias of the SCM) within each subband respectively. Furthermore, we showed that the combination of the (FFT-based) SCM and SCF outperformed and was more computationally efficient than the alternative feature combination of MFCC and frame-averaged FM, where FM extraction occurs in the time domain. Summing up, these results not only highlight SCF as a promising feature, for the first time, in speaker recognition system, but also demonstrate the ability of SCM and SCF to provide a better approximation of the speech spectrum as compared with the MFCC.

6.1.2 Investigation of Front-end Diversity

Chapter 4 described an effort to determine the extent to which performance improvements in fused systems based on different types of features is achieved purely through different 'partitioning' of the acoustic space (UBM), rather than primarily through speaker modelling (MAP adaptation). Two different types of experiments were considered: ensembles of different variants of MFCCs, and the use of clustering comparison measures.

A number of variations on MFCC features were employed to investigate the range of performance when systems based on features with minimal or no speaker-related information with respect to MFCCs were fused with an MFCC-based system. Any improvement observed in the fused speaker recognition performance was assumed to be attributable to acoustic modelling differences. The variations considered include different subband energy weighting, different filterbank, leaving out one band of filterbank enegies and drop-one-out cepstral coefficient. Interestingly, evaluations on the NIST 2006 SRE database showed relative improvements of up to 17% in EER when one modified MFCC subsystem (26 Gammatone MFCC) was fused with a conventional MFCC-based system, and improvements of up to 22% when two modified MFCC subsystems (14 Gabor MFCC + SCM) were fused. This seems to support the hypothesis that system diversity can be achieved purely through different 'partitioning' of the acoustic space.

Next, we introduced a novel approach, based on clustering comparison measures (normalised information distance) to separately investigate the acoustic and speaker modelling 'stages' of the GMM-UBM based systems, towards determining the contributions of each stage (acoustic and speaker modelling stage) to the speaker recognition performance across different features. First, we investigated the contribution of the acoustic modeling in fused systems towards the speaker recognition performance, in particular for systems fused with MFCC based systems. Interestingly, the results indicated that there is a high correlation between the NID of the acoustic clustering and reduction in EER for the fusion of systems with different front-end features, where a higher NID consistently correspond to a greater reduction in EER for the fused system. This strongly suggests that the NID between the UBMs could be utilised as an initial indicator of the amount of complementary information between two front-ends before computing any speaker recognition results, which could be time-consuming. Also, the fact that the UBM alone (with no speaker modelling) could be somewhat predictive of EER seems to be an interesting new perspective on speaker recognition using the GMM-UBM approach.

Next, we investigated the contribution of the acoustic modeling in single-feature systems through the use of resampling. It was observed from the experimental results that features that exhibit good ‘stability’ with respect to repeated clustering are shown to give good EER performance in speaker recognition, demonstrating the importance of stability in acoustic modelling and explaining in clustering terms why MFCC usually outperforms alternative features. Then, the extent to which different features give different clustering after MAP adaptation (speaker modelling) was also investigated. Surprisingly, we observed that all feature sets have almost equal amounts of adaptation in both feature domains, with respect to the UBM. This suggests that different features do not differ very much in how much they model individual speakers, or that speaker modelling may not be as important as acoustic modelling.

Finally, based on the findings on the importance of the acoustic modeling, a novel utterance selection algorithm on training a “stable” UBM was presented and evaluated on the NIST 2006 database. Results showed that using NID-based resampling to select utterances during UBM training can improve speaker recognition performance up to an

11% relative reduction in EER despite employing a smaller set of training data (20-30% of the usual UBM training data set). This could be due to the fact that introducing more utterances to a "stable" UBM may simply increase the variability within the UBM which might not be desirable in terms of acoustic space modelling.

Notably, these findings represent the first direct observation relating the changes in speaker recognition performances to changes in acoustic modeling and the importance of the acoustic space modelling (UBM) where focused research on UBM training has not yet been conducted in the literature. Moreover, it was possibly the first attempt to investigate the relative contributions of the acoustic and speaker modelling aspects of the speaker recognition task on the speaker recognition performance separately in *any* speaker recognition system. Summarising this last contribution, we have introduced to the speaker recognition community a clustering comparison approach that facilitates the development of stable single-UBM systems and highly complementary multiple-UBM systems.

6.1.3 Investigating Classification Approaches

Investigation into an alternative and complementary classification method, namely the sparse representation classification (SRC), for speaker recognition was reported in Chapter 5. In our research, we investigated the discriminative nature of the sparse representation classification using supervectors (GMM-SRC) for speaker verification on the contemporary NIST SRE databases and proposed the inclusion of nuisance attribute projection (NAP) to reduce or compensate the effect of inter-session variability in SRC-based systems. In addition, in an attempt to understand the similarities and differences between SRC and SVM, a comparison of SVM and SRC in terms of classification performance on 2-dimensional data (for easy visualisation) was conducted. We observed that SRC has the advantage of allowing a more adaptive classification with respect to the test sample by changing the number and type of support training samples for each test sample as opposed to SVM, which fixes the number and type of supports after the training procedure. Experimental results on the NIST 2006 SRE database indicated that the GMM-SRC was able to achieve comparable EER performance to GMM-SVM. On the other hand, despite the comparable performance, previously published SRC-based systems exhibited a significantly slower recognition process compared with SVM-based systems, because the sparse representation of large dimension supervectors requires a large amount of memory due to the over-complete dictionary. Therefore, we instead adopted the speaker factors (from JFA model) as feature vectors for the SRC, producing an approach we termed joint factor analysis-sparse representation classification (JFA-SRC).

However, the initial performance of the JFA-SRC was significantly poorer than that of JFA-SVM. Therefore, we evaluated a range of existing techniques for sparse representation classification and examined the effect on speaker recognition performance. First, we observed that the inclusion of the identity matrix in the dictionary helps to remove sensitivity to outliers and appears to be an essential aspect of the dictionary composition, giving a relative reduction of 8.6% in EER on the NIST 2006 SRE. Next, a sparseness method that uses a combination of ℓ_1 and ℓ_2 (Elastic net), offers better performance than one with only a ℓ_1 constraint, since the latter enforces a high degree of sparseness which leads to a decrease in accuracy. Notably, the combination of both techniques (inclusion of the identity matrix and $\ell_1+\ell_2$ constraint) results in a relative reduction of 24% in EER on the NIST 2006 SRE, achieving comparable performance with JFA-SVM.

In an attempt to further increase the computational efficiency of SRC-based systems, a novel SRC background dataset selection based on column vector frequency was presented. Although no improvement in terms of EER was observed, we demonstrated that a smaller dictionary refined by column vector frequency could be used, allowing a faster verification process. Furthermore, we showed that the dictionary chosen for development on the NIST 2006 SRE generalised well to the evaluation on the NIST 2010 SRE corpus for different evaluation condition, as opposed to SVM background data which requires a significant amount of tuning based on the evaluation condition.

Finally, a detailed comparison of JFA-SRC across various state-of-the-art classifiers used in speaker recognition systems was conducted on the NIST 2010 SRE databases (conditions 1-5). These included the GMM-SVM, JFA, JFA-SVM and JFA-CDS (joint factor analysis–cosine distance scoring) configurations. Experimental results on the NIST 2010 SRE show that JFA-SRC consistently outperformed JFA-SVM and JFA-CDS in EER in the range of 0.05–0.94% (absolute) and minDCF in the range of 0.021-0.147 (absolute). Furthermore, the fusion of JFA and JFA-SRC achieved a minimum relative EER reduction of 5.4% and minimum relative minDCF reduction of 2.4% over JFA alone. Interestingly, the JFA-SRC achieved the best minDCF both as an individual (among GMM-SVM, JFA, JFA-SVM and JFA-CDS) and fused system. These results highlight the usefulness of SRC-based systems when combined with other systems.

6.1.4 Multi-feature and Multi-classification Evaluation of Improved Verification System

In Chapter 5, the various proposed features and classification method were evaluated on the contemporary NIST 2010 SRE database. The features include the LogLSGD, SCM, SCF and MFCC (baseline). The classification methods include the JFA-SRC and JFA (baseline). The systems were compared both as an individual system and fused system. It has been observed that the best performing feature is MFCC (which is expected), followed by SCM, SCF and LogLSGD; these results demonstrate the consistency in performance across databases. Next, all possible pairwise combination of the features and classification were conducted. Notably, the fusion of MFCC JFA-SRC and SCF JFA gave the best fused performance reported in this thesis with a EER of 1.97% and a minDCF of 0.327 (a relative EER reduction of 38% and relative minDCF reduction of 25%), outperforming the fusion of MFCC JFA and MFCC JFA-SRC (EER of 2.82% and minDCF of 0.395).

In summary, these results formally validate that the alternative features considered carry complementary information leading to improved performance of the speaker verification systems. Furthermore, they also show that recognition performance can be *significantly* improved through score-level fusion of different classification systems using different features.

6.2 Future Work

The research discussed in this thesis has provided a number of avenues for future work as outlined below:

- The performance of various features on a common backend (Chapter 3 and 4) or MFCC on a different backend (Chapter 5) is included in this thesis and is useful for comparison purposes. Although different combinations of feature with JFA and/or JFA-SRC were considered in section 5.5.4, the various combinations are not comprehensive. It is therefore interesting to further investigate if a specific classifier works better for a feature as compared to another.
- Although different frequency scales were used to evaluate the proposed spectral centroid feature systems, the optimal number of subbands were not

considered in this thesis. This is because spectral centroid frequency (SCF) are mainly proposed for comparisons with the all-pole FM, so feature dimensions of SCF are fixed at 14 (same as all-pole FM). An investigation into the optimal number of filters for SCF and SCM can be conducted in the future since it has been shown in [76, 213] that the number of filters have a significant effect on the performance for various applications. In [76], it was observed that the EER vary in the range of 14% to 21% for the extraction of the all-pole FM with 6 to 26 filters (optimal number of filter = 16). For cognitive load classification in [213], the accuracy varies between 49% and 72% for the extraction of SCF and SCM with 2 to 20 filters (optimal number of filter = 6).

- The work reported in this thesis indicates that front-end diversity is achieved purely through different ’partitioning’ of the acoustic space. Furthermore, the fusion of systems with alternative clustering on the acoustic space leads to an improvement in system performance (section 4.2.6). However not all fusion of systems with alternative clustering leads to an improvement. This might be related to the stability issue of alternative clustering methods. This suggests an investigation to determine the optimal alternative clustering representation, if any, and suitable clustering approache is warranted.
- Finally, although care has been taken in this thesis to investigate many aspects of SRC-based speaker recognition, it is highly possible that the results can be further improved with more research, for example into areas such as score normalisation techniques for sparse representation, which remains an underexplored problem in the literature for SRC-based recognition

applications [188], and on improving the processing time taken by SRC-based classification [214].

Appendix A

Feature-domain Speaker Dependency Experiments for Group Delay Feature

As a preliminary investigation, the invariant cluster separation index [215] J can be employed to determine the degree to which speakers can be separated using the various type of features and for parameters setting:

$$J = trace(S_w^{-1}.S_B) \tag{A.1}$$

where S_w and S_B are the within-cluster and between-cluster scatter matrix for a data matrices whose rows $X_i, i = 1,2,\dots,N$ comprise features of dimensionality d and overall mean m respectively. The data were partitioned into N_c clusters representing N_c speakers, each with N_j features and mean m_j for the j^{th} speaker.

$$S_w = \sum_{j=1}^{N_C}\sum_{i=1}^{N_j}\left(\left(X_i - m_j\right)^T.\left(X_i - m_j\right)\right) \tag{A.2}$$

$$S_B = \sum_{j=1}^{N_C}\left(N_j.\left(m_j - m\right)^T.\left(m_j - m\right)\right) \tag{A.3}$$

Experiments were conducted on the group delay (GD) features obtained from various extraction techniques on NIST 2001 SRE database. The J-scores for various GD extraction techniques are shown in Figure A.1. These results suggest that a window length of L = 3 provides a good trade-off between regularising the GD estimate and preserving the frequency resolution. A longer window gives a smoother and more robust estimate but blurs the frequency resolution and introduces unwanted correlations between samples as shown in Figure 3.2(d).

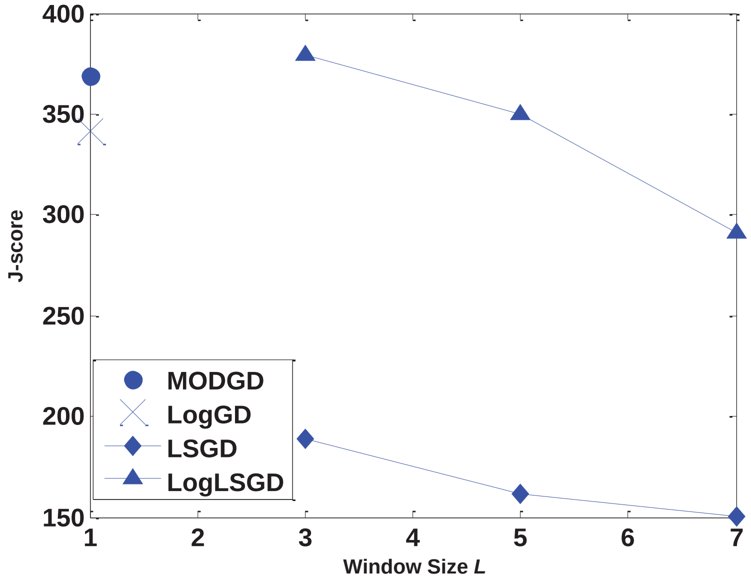


Figure A.1 Comparison of J-score for various GD feature extraction techniques

Appendix B

Frequency Band Allocation for Spectral Centroid Frequency Feature

B.1 Distribution of Speaker-Specific Information

In order to investigate the distribution of speaker-specific information of SCF in each frequency band of speech based on its speaker discriminative ability, speech is decomposed into subbands using 14 Gabor filters uniformly spaced across the telephone bandwidth. A uniform filterbank is used in this analysis to avoid the variability caused by different bandwidths of other filterbanks. The variation of speaker separation across different bands is studied in the model space using Kullback-Leibler (KL) distance [216] as follows

$$KL(f||g) \approx \sum_{i=1}^{N} \alpha_i KL(f_i||g_i) \tag{B.1}$$

where f and g are the two GMMs considered, N is the total number of mixtures, α_i is the weight of the i^{th} mixture. The symmetric version of the KL distance between two single mixtures is [217]

$$\begin{aligned} KL(f_i||g_i) \approx\ & 0.5\left(\vec{\boldsymbol{u}}_i^g - \vec{\boldsymbol{u}}_i^f\right)^T \left(\frac{1}{\boldsymbol{\Sigma}^g} + \frac{1}{\boldsymbol{\Sigma}^f}\right)\left(\vec{\boldsymbol{u}}_i^g - \vec{\boldsymbol{u}}_i^f\right) + 0.5 \\ & \cdot tr\left(\frac{\boldsymbol{\Sigma}^f}{\boldsymbol{\Sigma}^g} + \frac{\boldsymbol{\Sigma}^g}{\boldsymbol{\Sigma}^f} - 2 \cdot \mathbf{I}\right) \end{aligned} \tag{B.2}$$

where $\mathbf{\Sigma}$ is the (full) covariance matrix, $\boldsymbol{\mu}$ is the mean vector and $\mathbf{I}$ is the identity matrix. The pairwise KL distance in equation (B.2) is computed for every two speakers, with GMMs f and g, and then the final KL distance as shown in Figure B.1 is obtained by averaging across all pairwise distances (calculated on 174 speakers in the NIST 2001 training data) where higher values of KL distance indicates higher speaker discrimination information. It can be observed that the speaker discrimination is higher around the range of frequencies approximately 300 – 700 Hz and 2200 – 2800 Hz. It is, in general, consistent with the observation reported in [157]. The observation worth noting in this section and in [85, 157] is that there is a prominent speaker-specific region approximately above 2.4kHz after a dip approximately around 1.5kHz. This supports the results in [40] and discussed in section 2.1.1 that inter-speaker variation of the hypopharynx affects spectra approximately around 2.5kHz, obtained using morphological analysis. On the other hand, auditory scales such as mel and Bark give more importance to the lower frequency area and gradually reduce the importance to higher frequency area, in contrast to the above observation.

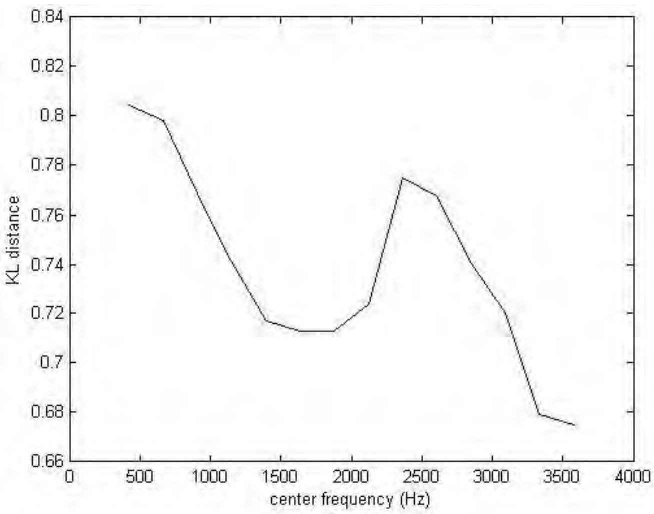


Figure B.1 KL distance for different frequency bands of SCF features averaged across the NIST 2001 database

B.2 Filterbank Design

Motivated by [85, 157], a filter bank was designed for SCF features to capture speaker specific information effectively. In this work, instead of reallocating the bandwidth with respect to the EER reduction curve reported in [157] for all-pole FM, the KL distance curve is used for simplicity since it avoids the need for repeating a series of leave-one-band-out speaker recognition experiments.

Having selected the KL distance curve as the basis curve, the next step is to use a method to allocate bandwidth. In [157] the area under the curve is equally divided while in [85] the inverse of the basis curve is used. As these two methods do not have a parameter to determine the amount of emphasis, we used the method in [218] as follows:

1) The KL distance curve is offset so that the minimum point of the curve has a weight of 1.
2) The bandwidths of the i^{th} filters are allocated with respect to the shifted KL distance curve:

$$BW_i \approx \frac{1/(KL_i)^{\alpha}}{\sum_{i=1}^{N} 1/(KL_i)^{\alpha}} BW \qquad \text{(B.3)}$$

where N is the number of bands, BW is the full speech bandwidth and α is a positive constant which determines the amount of emphasis (or de-emphasis) on high (or low) KL distance region of the filterbank.

3) The center frequency is allocated as the center within that bandwidth where the lower and upper cutoff frequencies of the i^{th} filters, f_i^l and f_i^u are defined as

$$f_i^l = f_{i-1}^u \qquad \text{(B.4)}$$

The bandwidth and center frequencies of the proposed filterbank for various values of α together with mel and uniform scale filterbank are given in Figure B.2. It can be observed that by increasing the value of α, the bandwidth of the filters in the high KL distance region are reduced, increasing the number of filters, but the bandwidth of the filters in the low KL distance region increases, decreasing the number of filters. This emphasises a high KL distance frequency region by allocating more filters in those regions and deemphasising lower KL distance regions by allocating a fewer number of filters.

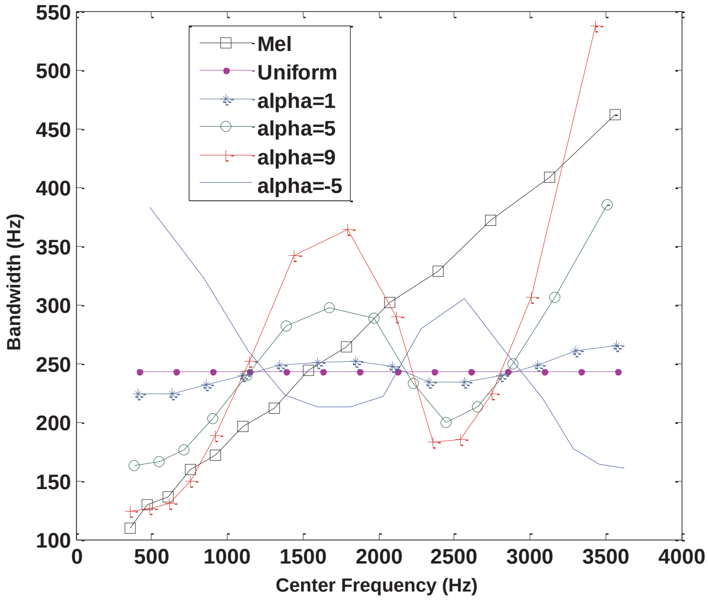


Figure B.2 Bandwidth and center frequencies of the filter banks based on KL-distance

B.3 Evaluation

As discussed in section B.2, increasing the value of α, increases the amount of emphasis on a high KL region. In order to determine the optimal amount of emphasis required for speaker recognition using SCF, speaker recognition experiments were conducted with values of α varying from 1 to 10 with a step size of 2 (our informal preliminary experiments have found that α beyond this range degrades the system performance).

Results for these experiments on the NIST 2001 SRE database are given in Table B-1. These show that α = 5 gave the best performance for the proposed KL distance based filterbank, outperforming the SCF feature extracted using auditory motivated mel-scale filters, which achieves an EER of 8.83%.

Above, an implicit assumption was made that to effectively capture speaker specific information, more filters need to be assigned by reducing their bandwidths in prominent speaker-specific regions rather than allocating a fewer number of larger bandwidth filters. In order to verify this assumption another filter was designed with α = -5, whereby a fewer number of filters with a larger bandwidth were assigned in the prominent speaker specific regions, as can be seen from Figure B.2. Experiments with this filter produced an EER of 10.51%, which did not perform better than other KL based filters, validating the assumption.

Table B-1: Development results of SCF features extracted using the proposed filterbank with various values of α on the NIST 2001 database.

α	EER (%)
1	9.38
3	8.82
5	**8.44**
7	8.82
9	9.03

As the NIST 2001 database was used in the development of the proposed filterbank, in order to ensure data-independency, the proposed filterbank with α = 5 was evaluated on the NIST 2006 database (1conv4w-1conv4w). The proposed filter (EER=6.06%) produced a 6% relative reduction in EER compared with SCF extracted using mel-scale (EER=6.45%). These results show that auditory motivated filters may not be the optimal filters for speaker recognition when the SCF feature is used and suggesting that feature-dependent filterbank design can be useful for speaker recognition.

References

[1] Haizhou Li, Kar-Ann Toh, and Liyuan Li, *Advanced Topics in Biometrics*: World Scientific, 2011.

[2] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 4-20, 2004.

[3] D. James, H. P. Hutter, and F. Bimbot, "CAVE–Speaker Verification in Banking and Telecommunications," in *Proceedings of the Ubilab Conference*, 1996.

[4] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, pp. 12-40, 2010.

[5] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, and A. Stolcke, "The SRI NIST 2010 speaker recognition evaluation system," in *Proc. of ICASSP*, 2011, pp. 5292-5295.

[6] D. Sturim, W. Campbell, N. Dehak, Z. Karam, A. McCree, D. Reynolds, F. Richardson, P. Torres-Carrasquillo, and S. Shum, "The MIT LL 2010 speaker recognition evaluation system: Scalable language-independent speaker recognition," in *Proc. of ICASSP*, 2011, pp. 5272-5275.

[7] A. Fazel and S. Chakrabartty, "An Overview of Statistical Pattern Recognition Techniques for Speaker Verification," *Circuits and Systems Magazine, IEEE*, vol. 11, pp. 62-81, 2011.

[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, 2000, pp. 19-41.

[9] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308-11, 2006.

[10] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," *Tech Report Online: http://www.crim.ca/perso/patrick.kenny*, 2005.

[11] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. of INTERSPEECH*, 2009.

[12] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. of ICASSP*, 2006, pp. 97-100.

[13] N. Brummer, L. Burget, and P. Kenny, "ABC system description for NIST SRE 2010," *Proc. NIST 2010 Speaker Recognition Evaluation*, 2010.

[14] J. Villalba, C. Vaquero, E. Lleida, and A. Ortega, "I3A NIST SRE2010 System Description."

[15] J. Weiwu, M. Man-Wai, R. Wei, and H. Meng, "The HKCUPU system for the NIST 2010 speaker recognition evaluation," in *Proc. of ICASSP*, 2011, pp. 5288-5291.

[16] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling: application to speaker verification," Doctoral Dissertation, Ecole de Technologie Superieure 2009.

[17] R. J. Vogt, S. Kajarekar, and S. Sridharan, "Discriminant NAP for SVM Speaker Recognition," in *Proc Odyssey: Speaker and Language Recognition Workshop*, 2008, pp. 629-632.

[18] L. Burget, N. Brümmer, D. Reynolds, P. Kenny, J. Pelecanos, R. Vogt, F. Castaldo, N. Dehak, R. Dehak, and O. Glembek, "Robust speaker recognition over varying channels–report from JHU workshop 2008," Technical report, http://www.clsp.jhu.edu/workshops/ws08/documents/jhu_report_main.pdf2009.

[19] R. P. Ramachandran, K. R. Farrell, R. Ramachandran, and R. J. Mammone, "Speaker recognition--general classifier approaches and data fusion methods," *Pattern Recognition*, vol. 35, pp. 2801-2821, 2002.

[20] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, pp. 210-229, 2006.

[21] H. Li, B. Ma, K. A. Lee, H. Sun, D. Zhu, K. C. Sim, C. You, R. Tong, I. Karkkainen, and C. L. Huang, "The I4U system in NIST 2008 speaker recognition evaluation," in *Proc. of ICASSP*, 2009, pp. 4201-4204.

[22] M. Nosratighods, T. Thiruvaran, J. Epps, E. Ambikairajah, M. Bin, and L. Haizhou, "Evaluation of a fused FM and cepstral-based speaker recognition system on the NIST 2008 SRE," in *Proc. of ICASSP*, 2009, pp. 4233-4236.

[23] N. Brummer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2072-2084, 2007.

[24] R. M. Hegde, H. A. Murthy, and G. V. Ramana Rao, "Application of the modified group delay function to speaker identification and discrimination," in *Proc. of ICASSP*, 2004, pp. 517-520.

[25] T. Thiruvaran, E. Ambikairajah, and M. Epps, "Group delay features for speaker recognition," in *Proc. of ICICSP*, 2007, pp. 1-5.

[26] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Extraction of FM components from speech signals using all-pole model," *Electronics Letters*, vol. 44, pp. 449-50, 2008.

[27] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, pp. 1289-1306, 2006.

[28] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210-227, 2009.

[29] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse Representation for Speaker Identification," in *Proc. of ICPR*, 2010, pp. 4460-4463.

[30] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, pp. 3736-3745, 2006.

[31] K. Huang and S. Aviyente, "Sparse representation for signal classification," *Advances in Neural Information Processing Systems*, vol. 19, p. 609, 2007.

[32] H. Xiyi and W. Fang-Xiang, "Sparse representation for classification of tumors using gene expression data," *Journal of Biomedicine and Biotechnology*, vol. 2009, 2009.

[33] Ming Li and S. Narayanan, "Robust Talking Face Video Verification Using Joint Factor Analysis And Sparse Representation On GMM Mean Shifted Supervectors," in *Proc. of ICASSP*, 2011.
[34] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Proc. of ICASSP*, 2010, pp. 4370-4373.
[35] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. of ICASSP*, 2002, pp. 617-620.
[36] T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Computationally efficient frame-averaged FM feature extraction for speaker recognition," *Electronics Letters*, vol. 45, pp. 335-337, 2009.
[37] T. Dietterich, "Ensemble Methods in Machine Learning Multiple Classifier Systems," in *Proc. of the First International Workshop on Multiple Classifier Systems*, 2000, pp. 1-15.
[38] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "Data-driven impostor selection for T-norm score normalisation and the background dataset in SVM-based speaker verification," *Advances in Biometrics*, pp. 474-483, 2009.
[39] L. Rabiner and B. Juang, *Fundamentals of speech recognition*: Prentice Hall, 1993.
[40] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustical Science and Technology*, vol. 26, pp. 16-26, 2005.
[41] K. Honda, "Individuality of orofacial form reflected in vowel spaces," in *Proc. Spring Meet. Acoust. Soc. Jpn.*, 1997, pp. 237–238.
[42] W. Fitch, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, p. 1511, 1999.
[43] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans," *The Journal of the Acoustical Society of America*, vol. 101, pp. 456-465, 1997.
[44] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high-and low-level features for speaker recognition," in *Proc. of EUROSPEECH*, 2003, pp. 2665-2668.
[45] R. J. Mammone, Z. Xiaoyu, and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," *Signal Processing Magazine, IEEE*, vol. 13, p. 58, 1996.
[46] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 639-643, 1994.
[47] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, J. Qin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and X. Bing, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," in *Proc. of ICASSP*, 2003, pp. IV-784-7 vol.4.
[48] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, pp. 357-66, 1980.
[49] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561-580, 1975.

[50] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America,* vol. 87, pp. 1738-52, 1990.
[51] J. Harrington and S. Cassidy, *Techniques in speech acoustics*: Kluwer Academic Publisher, 1999.
[52] H. Helmholtz, *On the Sensations of Tone*: New York: Dover Publications Inc, 1954.
[53] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication,* vol. 22, pp. 403-417, 1997.
[54] K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. of EUROSPEECH*, 2003, pp. 2117–2120.
[55] K. K. Paliwal and L. Alsteris, "Usefulness of phase in speech processing," in *Proc. IPSJ Spoken Language Processing Workshop,* 2003, pp. 1-6.
[56] L. D. Alsteris and K. K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication,* vol. 48, pp. 727-736, 2006.
[57] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *Speech Communication,* vol. 10, pp. 209-221, 1991.
[58] J. Tribolet, "A new phase unwrapping algorithm," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 25, pp. 170-177, 1977.
[59] V. V. Volkov and Y. Zhu, "Deterministic phase unwrapping in the presence of noise," *Optics letters,* vol. 28, pp. 2156-2158, 2003.
[60] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing: A Review Journal,* vol. 17, pp. 578-616, 2007.
[61] H. Banno, J. Lu, S. Nakamura, K. Shikano, and H. Kawahara, "Efficient representation of short-time phase based on group delay," in *Proc. of ICASSP*, 1998, pp. 861-864.
[62] M. H. Rajesh, A. M. Hema, and G. Venkata Ramana Rao, "Significance of the Modified Group Delay Feature in Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 190-202, 2007.
[63] H. A. Murthy and B. Yeganarayana, "Speech processing using group delay functions," *Signal Processing,* vol. 22, pp. 259-67, 1991.
[64] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. of ICASSP*, 2003, pp. 68-71.
[65] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing,* vol. 40, pp. 2281-2289, 1992.
[66] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals* vol. 100: Prentice-hall Englewood Cliffs, NJ, 1978.
[67] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Appropriate windowing for group delay analysis and roots of z-transform of speech signals," in *Proc. of EUSPICO*, 2004.
[68] L. D. Alsteris and K. K. Paliwal, "Evaluation of the modified group delay feature for isolated word recognition," in *Proc. of the Eighth International Symposium on Signal Processing and Its Applications*, 2005, pp. 715-718.
[69] K. R. Rao and P. Yip, *Discrete cosine transform: algorithms, advantages, applications* vol. 4: Academic Press London, 1990.

[70] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *Signal Processing, IEEE Transactions on*, vol. 41, pp. 3024-3051, 1993.
[71] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *Proc. of ICASSP*, 1991, pp. 421-424 vol.1.
[72] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On separating amplitude from frequency modulations using energy operators," in *Proc. of ICASSP*, 1992, pp. 1-4 vol.2.
[73] G. Fant, *Acoustic Theory of Speech Production*: The Hague: Mouton, 1960.
[74] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP28, pp. 599-601, 1980.
[75] D. Dimitriadis and P. Maragos, "Robust energy demodulation based on continuous models with application to speech recognition," in *Proc. of EUROSPEECH*, 2003, pp. 2853-2856.
[76] T. Thiruvaran, "Automatic Speaker Recognition Using Phase Based Features," Doctor of Philosophy, School of Electrical Engineering and Telecommunications, The University of New South Wales, 2009.
[77] A. Potamianos, "Speech Processing Applications Using an AM-FM Modulation Model," Doctor of Philosophy, Division of Applied Sciences, Harvard University, 1995.
[78] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, pp. 95-120, 1994.
[79] N. Kaibao, G. Stickney, and Z. Fan-Gang, "Encoding frequency Modulation to improve cochlear implant performance in noise," *IEEE Transactions on Biomedical Engineering*, vol. 52, pp. 64-73, 2005.
[80] J. L. Flanagan, D. I. S. Meinhart, R. M. Golden, and M. M. Sondhi, "Phase Vocoder," *The Journal of the Acoustical Society of America*, vol. 38, pp. 939-940, 1965.
[81] D. Hosseinzadeh and S. Krishnan, "On the use of complementary spectral features for speaker recognition," *Eurasip Journal on Advances in Signal Processing*, vol. 2008, p. 10, 2008.
[82] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304-1312, 1974.
[83] T. Kinnunen, "Spectral features for automatic text-independent speaker recognition," *Licentiates thesis, Department of computer science, University of Joensuu*, 2003.
[84] A. Lawson, P. Vabishchevich, M. Huggins, P. Ardis, B. Battles, and A. Stauffer, "Survey and evaluation of acoustic features for speaker recognition," in *Proc. of ICASSP*, 2011, pp. 5444-5447.
[85] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Communication*, vol. 50, pp. 312-322, 2008.
[86] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, 1995.

[87] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," in *IEEE Workshop Neural Networks for Signal Processing*, 2000, pp. 775-784.
[88] D. Reynolds, "A gaussian mixture model approach to text-independent speaker identification," Doctor of Philosophy, Georgia Institute of Technology, 1992.
[89] G. J. McLachlan and K. E. Basford, *Mixture models: Inference and applications to clustering*: New York: Marcel Dekke, 1988.
[90] J. L. Gauvain and L. Chin-Hui, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291-298, 1994.
[91] R. Togneri and D. Pullella, "An Overview of Speaker Identification: Accuracy and Robustness Issues," *IEEE Circuits and Systems Magazine*, vol. 11, pp. 23-61, 2011.
[92] R. Faltlhauser and G. Ruske, "Improving Speaker Recognition Using Phonetically Structured Gaussian Mixture Models," in *Proc. of EUROSPEECH*, 2001, pp. 751-754.
[93] A. Park and T. J. Hazen, "ASR dependent techniques for speaker identification," in *Proc. of ICSLP*, 2002, pp. 2521–2524.
[94] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Loquendo-Politecnico di Torino’s 2006 NIST speaker recognition evaluation system," in *Proc. of INTERSPEECH*, 2007, pp. 1238-1241.
[95] E. G. Hansen, R. E. Slyh, and T. R. Anderson, "Speaker recognition using phoneme-specific GMMs," in *Proc Odyssey: Speaker and Language Recognition Workshop*, 2004, pp. 179-184.
[96] W. M. Campbell, "A SVM/HMM system for speaker recognition," in *Proc. of ICASSP*, 2003, pp. 209-212.
[97] V. N. Vapnik, *The nature of statistical learning theory*: Springer, 1995.
[98] R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *The Journal of Machine Learning Research*, vol. 1, p. 160, 2001.
[99] N. List and H. U. Simon, "SVM-optimization and steepest-descent line search," in *Proceedings of the 22nd Annual Conference on Computational Learning Theory*, 2009.
[100] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, pp. 121-167, 1998.
[101] N. A. Gunasekara, "Meta learning on string kernel SVMs for string categorization," Master of Computer and Information Sciences, AUT University, 2010.
[102] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, pp. 131-159, 2002.
[103] H. Frohlich and A. Zell, "Efficient parameter selection for support vector machines in classification and regression via model-based global optimization," in *International Joint Conference on Neural Networks*, 2005, pp. 1431-1436.
[104] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. of ICASSP*, 2002, pp. 161-164.
[105] P. J. Moreno and P. P. Ho, "A new SVM approach to speaker identification and verification using probabilistic distance kernels," in *Proc. of EUROSPEECH*, 2003, pp. 2965-2968.

[106] Z. N. Karam and W. M. Campbell, "A multi-class MLLR kernel for SVM speaker recognition," in *Proc. of ICASSP*, 2008, pp. 4117-4120.
[107] V. Wan and S. Renals, "Evaluation of kernel methods for speaker verification and identification," in *Proc. of ICASSP*, 2002, pp. 669-672.
[108] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 203-210, 2005.
[109] Y. Lei, T. Hasan, J. W. Suh, A. Sangwan, H. Boril, G. Liu, K. Godin, C. Zhang, and J. H. L. Hansen, "The CRSS systems for the 2010 NIST speaker recognition evaluation."
[110] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Improved SVM speaker verification through data-driven background dataset collection," in *Proc. of ICASSP*, 2009, pp. 4041-4044.
[111] J. W. Suh, Y. Lei, W. Kim, and J. H. L. Hansen, "Effective background data selection in SVM speaker recognition for unseen test environment: more is not always better," in *Proc. of ICASSP*, 2011.
[112] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Exploiting multiple feature sets in data-driven impostor dataset selection for speaker verification," in *Proc. of ICASSP*, 2010, pp. 4434-4437.
[113] E. J. Candès, "Compressive sampling," in *Proc. Int'l Congress of Mathematicians*, 2006.
[114] A. Y. Yang, M. Gastpar, R. Bajcsy, and S. S. Sastry, "Distributed sensor perception via sparse representation," *Proceedings of the IEEE*, vol. 98, pp. 1077–1088.
[115] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, pp. 489-509, 2006.
[116] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, p. 118, 2007.
[117] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, pp. 237-260, 1998.
[118] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, pp. 797-829, 2006.
[119] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, pp. 1207-1223, 2006.
[120] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Transactions on Information Theory*, vol. 52, pp. 5406-5425, 2006.
[121] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 643-660, 2001.
[122] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," ed, 1993.
[123] A. R. Webb, *Statistical pattern recognition*: John Wiley & Sons Inc, 2002.

[124] C. Q. Alex Solomonoff, and William M. Campbell, "Channel Compensation for SVM Speaker Recognition," in *Proc Odyssey: Speaker and Language Recognition Workshop*, 2004, pp. 57-62.
[125] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. of ICASSP*, 2005, pp. 629-32.
[126] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133-147, 1998.
[127] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578-589, 1994.
[128] K.-K. Yiu, M.-W. Mak, and S.-Y. Kung, "Environment Adaptation for Robust Speaker Verification," in *Proc. of EUROSPEECH*, 2003, pp. 2973-2976.
[129] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc Odyssey: Speaker and Language Recognition Workshop*, 2001, pp. 213 - 218.
[130] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. of ICASSP*, 2003, pp. II-53-6 vol.2.
[131] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Improvements in factor analysis based speaker verification," in *Proc. of ICASSP*, 2006, pp. 113-116.
[132] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788-798, 2011.
[133] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1435-47, 2007.
[134] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1448-1460, 2007.
[135] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. of INTERSPEECH*, 2006, pp. 1471 - 1474.
[136] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "i-vector based speaker recognition on short utterances," in *Proc. of INTERSPEECH*, 2011, pp. 2341-2344.
[137] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing: A Review Journal*, vol. 10, pp. 42-54, 2000.
[138] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. of EUROSPEECH*, 1997, pp. 963-966.
[139] R. Zheng, S. Zhang, and B. Xu, "A Comparative Study of Feature and Score Normalization for Speaker Verification," in *Advances in Biometrics*. vol. 3832, D. Zhang and A. Jain, Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 531-538.
[140] D. J. Mashao and M. Skosan, "Combining classifier decisions for robust speaker identification," *Pattern Recognition*, vol. 39, pp. 147-155, 2006.
[141] Tomi Kinnunen, Ville Hautamaki, and P. Franti, "Fusion of Spectral Feature Sets for Accurate Speaker Identification," in *Proc. of SPECOM*, 2004, pp. 361 - 365.
[142] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, pp. 2115-2125, 2003.

[143] N. Brummer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, pp. 230-275, 2006.
[144] B. Ma, H. Li, and R. Tong, "Spoken language recognition using ensemble classifiers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2053-2062, 2007.
[145] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Proc. of INTERSPEECH*, 2010, pp. 2726-2729.
[146] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. of EUROSPEECH*, 1997, pp. 1895–1898.
[147] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proc. of ICASSP*, 1992, pp. 517-520.
[148] D. Graff, A. Canavan, and G. Zipperlen, "Switchboard-2 Phase I," ed: Linguistic Data Consortium, Philadelphia, 1998.
[149] D. Graff, K. Walker, and A. Canavan, "Switchboard-2 Phase II," ed: Linguistic Data Consortium, Philadelphia, 1999.
[150] D. Graff, D. Miller, and K. Walker, "Switchboard-2 Phase III Audio," ed: Linguistic Data Consortium, Philadelphia, 2002.
[151] D. Graff, K. Walker, and D. Miller, "Switchboard Cellular Part 1 Audio," ed: Linguistic Data Consortium, Philadelphia, 2001.
[152] D. Graff, K. Walker, and D. Miller, "Switchboard Cellular Part 2 Audio," ed: Linguistic Data Consortium, Philadelphia, 2004.
[153] C. Cieri, W. Andrews, J. P. Campbell, G. Doddington, J. Godfrey, S. Huang, M. Liberman, A. Martin, H. Nakasone, and M. Przybocki, "The Mixer and Transcript Reading Corpora: Resources for Multilingual," in *Proc. Int. Conf. Lang. Resources Evaluation (LREC)*, 2006, pp. 117-120.
[154] NIST. (2012). *NIST speech group website*. Available: http://www.itl.nist.gov/iad/mig/tests/spk/
[155] N. Thian, C. Sanderson, and S. Bengio, "Spectral Subband Centroids as Complementary Features for Speaker Authentication Biometric Authentication," in *Proc. of First International Conference on Biometric Authentication*, 2004, pp. 631-639.
[156] T. Kinnunen, B. Zhang, J. Zhu, and Y. Wang, "Speaker Verification with Adaptive Spectral Subband Centroids," in *Proc. International Conference on Biometrics*, 2007, pp. 58-66.
[157] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Analysis of band structures for speaker-specific information in FM feature extraction," in *Proc. of INTERSPEECH*, 2009, pp. 1111-1114.
[158] K. Sri Rama Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, pp. 52-55, 2006.
[159] E. Ambikairajah, "Emerging features for speaker recognition," in *Proc. of ICICS*, 2007, pp. 1-7.
[160] T. Kinnunen, "Joint Acoustic-Modulation Frequency for Speaker Recognition," in *Proc. of ICASSP*, 2006, pp. I-I.
[161] Z. Nengheng, L. Tan, and P. C. Ching, "Integration of Complementary Acoustic Features for Speaker Recognition," *IEEE Signal Processing Letters*, vol. 14, pp. 181-184, 2007.

[162] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt, and R. R. Gadde, "Speaker recognition using prosodic and lexical features," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 19-24.

[163] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification,* vol. 2, pp. 193-218, 1985.

[164] M. Meil and D. Heckerman, "An experimental comparison of model-based clustering methods," *Machine Learning,* vol. 42, pp. 9-29, 2001.

[165] T. M. Cover and J. A. Thomas, *Elements of Information Theory*: John Wiley & Sons, Inc., 1991.

[166] N. X. Vinh, J. Epps, J. Bailey, and M. Houle, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *The Journal of Machine Learning Research,* pp. 2837-2854, 2010.

[167] A. Strehl and J. Ghosh, "Cluster ensembles---a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research,* vol. 3, pp. 583-617, 2003.

[168] Christopher M. Bishop, Markus Svensen, and Christopher K. I.Williams, "GTM: The Generative Topographic Mapping," *Neural Computation,* vol. 10, pp. 215-234, 1997.

[169] N. X. Vinh, "Information Theoretic Methods for Clustering with Applications to Microarray Data," PhD, School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia, 2010.

[170] S. P. Smith and R. Dubes, "Stability of a hierarchical clustering," *Pattern Recognition,* vol. 12, pp. 177-187, 1980.

[171] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning,* vol. 52, pp. 91-118, 2003.

[172] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Proc. of PSB*, 2002, p. 6.

[173] V. Nguyen Xuan and J. Epps, "minCEntropy: A Novel Information Theoretic Approach for the Generation of Alternative Clusterings," in *IEEE 10th International Conference on Data Mining* 2010, pp. 521-530.

[174] T. Hasan, Y. Lei, A. Chandrasekaran, and J. H. L. Hansen, "A novel feature sub-sampling method for efficient universal background model training in speaker verification," in *Proc. of ICASSP*, 2010, pp. 4494-4497.

[175] H. Xiyi, "Cancer classification by sparse representation using microarray gene expression data," in *IEEE International Conference on Bioinformatics and Biomeidcine Workshops*, 2008, pp. 174-177.

[176] F. Girosi, "An equivalence between sparse approximation and support vector machines," *Neural computation,* vol. 10, pp. 1455-1480, 1998.

[177] C. Blake and C. J. Merz, "UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California," *Department of Information and Computer Science,* vol. 460, 1998.

[178] T. N. Sainath, D. Nahamoo, B. Ramabhadran, and D. Kanevsky, "Sparse representation phone identification features for speech recognition," *Speech and Language Algorithms Group, IBM, Tech. Rep,* 2010.

[179] B. G. Park, K. M. Lee, and S. U. Lee, "Face recognition using face-arg matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1982-1988, 2005.

[180] P. J. Bickel, B. Li, A. B. Tsybakov, S. A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. van der Vaart, "Regularization in statistics," *Test*, vol. 15, pp. 271-344, 2006.
[181] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, pp. 1-32, 2000.
[182] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?," *Database Theory—ICDT'99*, pp. 217-235, 1999.
[183] J. M. K. Kua, E. Ambikairajah, J. Epps, and R. Togneri, "Speaker verification using sparse representation classification," in *Proc. of ICASSP*, 2011, pp. 4548-4551.
[184] Y. Ariki and K. Doi, "Speaker recognition based on subspace methods," in *lCSLP*, 1994, pp. 1859 - 1862.
[185] Y. Ariki, S. Tagashira, and M. Nishijima, "Speaker recognition and speaker normalization by projection to speaker subspace," in *Proc. of ICASSP*, 1996, pp. 319-322.
[186] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1088-1099, 2006.
[187] P. Campadelli, E. Casiraghi, and G. Valentini, "Support vector machines for candidate nodules classification," *Neurocomputing*, vol. 68, pp. 281-288, 2005.
[188] M. Li, X. Zhang, Y. Yan, and S. Narayanan, "Speaker Verification using Sparse Representations on Total Variability I-Vectors," in *Proc. of INTERSPEECH*, 2011.
[189] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *Proc. of ICASSP*, 2009, pp. 4237-4240.
[190] N. Dehak. (2011). *Low-dimensional speech representation based on Factor Analysis and its applications*. Available: http://www.clsp.jhu.edu/news-events/abstract.php?sid=20110812
[191] N. Dehak, Z. N. Karam, D. A. Reynolds, R. Dehak, W. M. Campbell, and J. R. Glass, "A channel-blind system for speaker verification," in *Proc. of ICASSP*, 2011, pp. 4536-4539.
[192] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 980-988, 2008.
[193] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, pp. 586-597, 2007.
[194] E. Candes and J. Romberg. *1-MAGIC: Recovery of Sparse Signals via Convex Programming, 2005*. Available: http://www.acm.caltech.edu/l1magic.
[195] D. Donoho, V. Stodden, and Y. Tsaig, "Sparselab," *Software: http://sparselab. stanford. edu*, vol. 25, 2005.
[196] K. Koh, S. Kim, and S. Boyd, "l1 ls: A matlab solver for large-scale l1-regularized least squares problems," ed: Stanford University, Mar, 2007.
[197] C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," *Database Theory—ICDT 2001*, pp. 420-434, 2001.

[198] D. Kanevsky, T. N. Sainath, B. Ramabhadran, and D. Nahamoo, "An analysis of sparseness and regularization in exemplar-based methods for speech classification," in *Proc. of INTERSPEECH*, 2010.

[199] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review,* vol. 51, pp. 34–81, 2009.

[200] J. F. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves, "Compressive Sensing for Missing Data Imputation in Noise Robust Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 272-287, 2010.

[201] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, pp. 267-288, 1996.

[202] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2346-2356, 2008.

[203] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, pp. 301-320, 2005.

[204] A. Carmi and P.Gurfil, "Convex Feasibility Programming for Compressed Sensing," Technion 2009.

[205] A. N. Tikhonov and V. I. A. Arsenin, *Solutions of ill-posed problems*: Winston Washington, DC:, 1977.

[206] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, p. 1, 2010.

[207] M. Aharon, M. Elad, and A. Bruckstein, "K SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311-4322, 2006.

[208] M. D. Plumbley, "Dictionary learning for l1-exact sparse coding," in *International Conference on Independent Component Analysis and Signal Separation*, 2007, pp. 406-413.

[209] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *Advances in Neural Information Processing Systems*, vol. 21, pp. 1033-1040, 2009b.

[210] D. Vainsencher, S. Mannor, and A. M. Bruckstein, "The Sample Complexity of Dictionary Learning," *Arxiv preprint arXiv:1011.5395*, 2010.

[211] F. Sedlák, T. Kinnunen, V. Hautamäki, K. A. Lee, and H. Li, "Classifier subset selection and fusion for speaker verification," in *Proc. of ICASSP*, 2011, pp. 4544 - 4547.

[212] R. Matejka, L. Burget, P. Schwarz, O. Glembek, M. Karafiat, F. Grezl, J. Cernocky, D. A. van Leeuwen, N. Brummer, and A. Strasheim, "STBU System for the NIST 2006 Speaker Recognition Evaluation," in *Proc. of ICASSP*, 2007, pp. IV-221-IV-224.

[213] P. N. Le, "The use of Spectral Information in the Development of Novel Techniques for Speech-based Cognitive load classification," PhD, School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia, 2012.

[214] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, pp. 245-266, 2012.

[215] P. J. G. Lisboa, I. O. Ellis, A. R. Green, F. Ambrogi, and M. B. Dias, "Cluster-based visualisation with scatter matrices," *Pattern Recognition Letters*, vol. 29, pp. 1814-1823, 2008.
[216] J. Goldberger and H. Aronowitz, "A distance measure between gmms based on the unscented transform and its application to speaker recognition," in *Proc. of INTERSPEECH*, 2005, pp. 1985–1988.
[217] T. Stadelmann and B. Freisleben, "Fast and Robust Speaker Clustering Using the Earth Mover'S Distance and Mixmax Models," in *Proc. of ICASSP*, 2006, pp. I-I.
[218] L. Phu Ngoc, E. Ambikairajah, E. H. C. Choi, and J. Epps, "A non-uniform subband approach to speech-based cognitive load classification," in *ICICS*, 2009, pp. 1-5.