VOICE BIOMETRICS UNDER MISMATCHED NOISE CONDITIONS

Surosh G. Pillay

A thesis submitted in partial fulfilment of the requirements of the University of Hertfordshire for the degree of Doctor of Philosophy

The programme of research was carried out in the
Faculty of Science, Technology and Creative Arts, University of Hertfordshire,
Hatfield, Hertfordshire, AL10 9AB, United Kingdom

September 2010

"Genius is one percent inspiration and 99 percent perspiration."
Thomas Edison

ABSTRACT

This thesis describes research into effective voice biometrics (speaker recognition) under mismatched noise conditions. Over the last two decades, this class of biometrics has been the subject of considerable research due to its various applications in such areas as telephone banking, remote access control and surveillance. One of the main challenges associated with the deployment of voice biometrics in practice is that of undesired variations in speech characteristics caused by environmental noise. Such variations can in turn lead to a mismatch between the corresponding test and reference material from the same speaker. This is found to adversely affect the performance of speaker recognition in terms of accuracy.

To address the above problem, a novel approach is introduced and investigated. The proposed method is based on minimising the noise mismatch between reference speaker models and the given test utterance, and involves a new form of Test-Normalisation (T-Norm) for further enhancing matching scores under the aforementioned adverse operating conditions. Through experimental investigations, based on the two main classes of speaker recognition (i.e. verification/ open-set identification), it is shown that the proposed approach can significantly improve the performance accuracy under mismatched noise conditions.

In order to further improve the recognition accuracy in severe mismatch conditions, an approach to enhancing the above stated method is proposed. This, which involves providing a closer adjustment of the reference speaker models to the noise condition in the test utterance, is shown to considerably increase the accuracy in extreme cases of noisy test data. Moreover, to tackle the computational burden associated with the use of the enhanced approach with open-set identification, an efficient algorithm for its realisation in this context is introduced and evaluated.

The thesis presents a detailed description of the research undertaken, describes the experimental investigations and provides a thorough analysis of the outcomes.

ACKNOWLEDGEMENTS

This thesis is the culmination of a long journey of hard work, dedication and determination that has undoubtedly shaped me into the person I am today. However, this would not have been possible without the support and encouragement of some very important people to whom I will forever be indebted.

First and foremost, I would like to express my deep and sincere gratitude to my principal supervisor, Prof. Aladdin Ariyaeeinia, for his patience, enthusiasm and the invaluable discussions throughout this research programme. His methodological approach to research, attention to detail and high standards have been a great inspiration and motivation to me. I consider myself very lucky to have had not only an excellent supervisor, but also a mentor, and a role-model.

I would like to extend my gratitude to my industrial supervisor, Mr Mark Pawlewski for his continuous support and exciting insights into the ‘real world’. I would also like to thank British Telecom (BT) for providing the funding which has allowed me to carry out research in this field.

I am also very thankful to my third supervisor, Dr. Perasiriyan Sivakumaran for always finding time out of his busy schedule to provide expert advice and technical guidance. I have undoubtedly learnt a lot from our discussions and I will always aspire to develop his ability to be so precise and thorough while always keeping an open mind for new ideas and suggestions.

Thanks also go to all the members of staff and technicians of the School of Engineering and Technology. I am also thankful to my fellow researchers and friends for providing such a fun and stimulating environment to work in. In particular, my special thanks go to: Milos Milosavljevic, Sat Juttla, Stratis Sofianos and Ali Bakhsh.

Finally, I would like to thank my parents for their long-distance support and for always believing in me. Last, but in no way least, my deepest note of gratitude goes to my fiancée Nishna for her love, incredible patience, encouragement, and for everything that we went through together to make this thesis a reality.

LIST OF ABBREVIATIONS

ANN - Artificial Neural Networks

CMN - Cepstral Mean Normalisation

CMS - Cepstral Mean Subtraction

CN - Cohort Normalisation

DCT - Discrete Cosine Transform

DDCS - Distortion Driven Cluster Splitting

DET - Detection Error Trade-off

DFT - Discrete Fourier Transform

EER - Equal Error Rate

EM - Expectation-Maximisation

FA - False Alarm

FAR - Rate of FA

FFT - Fast-Fourier Transform

FFTC - FFT Derived Cepstrum

GLDS - Generalised Linear Discriminant Sequence

GMM - Gaussian Mixture Model

HMM - Hidden Markov Model

IDFT - Inverse Discrete Fourier Transform

IER- Identification Error Rate

JFA- Joint Factor Analysis

L-D - Levinson-Durbin

LFCC - Linear Frequency-based Cepstral Coefficients

LGB - Linde-Buzo-Gray

LP - Linear Prediction

LPC - Linear Predictive Coding

LPCC - Linear Prediction-based Cepstral Coefficients

LSF -Line Spectral Frequency

LSP - Line Spectral Pair

MAP - Maximum a posteriori

MD - Missed Detection

MDR - Rate of MD

MFCC - Mel Filter bank - based Cepstral Coefficient

MFT- Missing Feature Theory

ML - Maximum Likelihood

MLLR - Maximum Likelihood Linear Regression

MLP - Multi-Layer Perceptron

M-Norm – Model-Normalisation

NAP- Nuisance Attribute Projection

NPA- Non Parametric Analysis

OSI - Open-Set Identification

OSIE - Open-Set Identification Error

OSTI-SI - Open-Set, Text-Independent Speaker Identification

PA- Parametric Analysis

PDF - Probability Density Function

PLP - Perceptual Linear Prediction

PLPC - Perceptual Linear Prediction based Coefficients

PMC-Parallel Model Combination

RASTA - Relative Spectral

RBF- Radial Basis Function

ROC- Receiver Operating Characteristics

SCD - Speaker Change Detection

SRE- Speaker Recognition Evaluation

SS- Spectral Subtraction

SV - Speaker Verification

SVM - Support Vector Machines

T-Norm- Test Normalisation

UBM - Universal Background Model

UCN - Unconstrained Cohort Normalisation

VAD - Voice Activity Detection

VQ - Vector Quantisation

WBLS - Weighted bilateral scoring

WM - World Model

WMN - World Model Normalisation

Z-Norm- Zero Normalisation

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Voice Biometrics

The ability to automatically, reliably and efficiently verify individuals' identities has nowadays become an essential requirement in many real-world applications. This is mainly attributed to the growing need to combat the alarming cases of identity theft, financial fraud and international terrorism. Examples of such applications include telephone banking, forensics, immigration control and online security. Traditionally, recognition systems have relied on what the individual knows, e.g. a password or a Personal Identification Number (PIN) and/or what the individual has, e.g. a key, token or a personal card [1-3]. However, such approaches have a number of limitations which are primarily due to the fact that they focus on identifying an object, phrase or set of numbers instead of the person. As a result, security breaches can easily occur if a person's card or key is lost, stolen or copied. On the other hand, passwords and PINs can be forgotten by a legitimate user if they are too difficult, guessed by an impostor if they are too simple or even cracked using sophisticated software technologies. Moreover, such methods have limited use in surveillance applications which involve operating in a surreptitious manner and therefore do not require user cooperation during the authentication process. The use of biometrics offers an alternative to the conventional methods of authentication, which helps to avoid the aforementioned problems [1, 2, 4]. Biometrics or biometric recognition is best defined as the process of automatically authenticating individuals based on their physiological (e.g. fingerprint, face, iris) and/or behavioural characteristics (e.g. handwriting, keystrokes, gait). These two categories are also referred to as intrinsic and extrinsic biometrics.

Voice biometrics or speaker recognition, is described as the process of recognising a person based on the unique characteristics of his/her voice. This can be considered as a hybrid type of biometrics since a speaker's voice is defined by the structure of the vocal tract (i.e. physiological component) as well as the way that person talks (i.e. behavioural component) [5-7]. These unique (speaker-specific) characteristics are usually exploited by individuals to recognise their friends or families over the telephone; an intrinsic ability which can also be considered as a naïve form of speaker recognition. As such, voice biometrics applications are, in general, not regarded as intrusive and users are not usually reluctant to provide a speech sample for recognition purposes. Furthermore, systems based on voice biometrics do not require any specialised hardware for capturing the speech signal. For instance, telephone based applications only require the user to have a telephone handset or a mobile phone, while non-telephone based applications involve the use of microphones and soundcards: technologies which these days are readily available at a very low-price [5]. Furthermore, in some cases, depending on the nature of applications, voice biometrics may be the only feasible option which is available for recognising individuals. For instance, in telephone banking applications where users need to be remotely authenticated prior to allowing access to their bank details. For all these reasons, speaker recognition is usually considered as one of the most attractive forms of biometric authentication.

1.2 General Approach

In general, voice biometrics (speaker recognition) can operate in one of the two main modes of verification and identification. Speaker verification is defined as the task of determining whether a speaker is who (s)he claims to be, based on a given test utterance [5-7]. In other words, this process can be considered as a 1:1 matching between a claimed identity and given voice sample. Such an identity claim may be made verbally, by typing-in a personal identity number, or by some other means. The speaker verification operation comprises two stages. These are the training (or enrolment) stage and the testing (or matching/recognition) stage. Figure 1.1 shows a block diagram of the general approach to speaker verification.

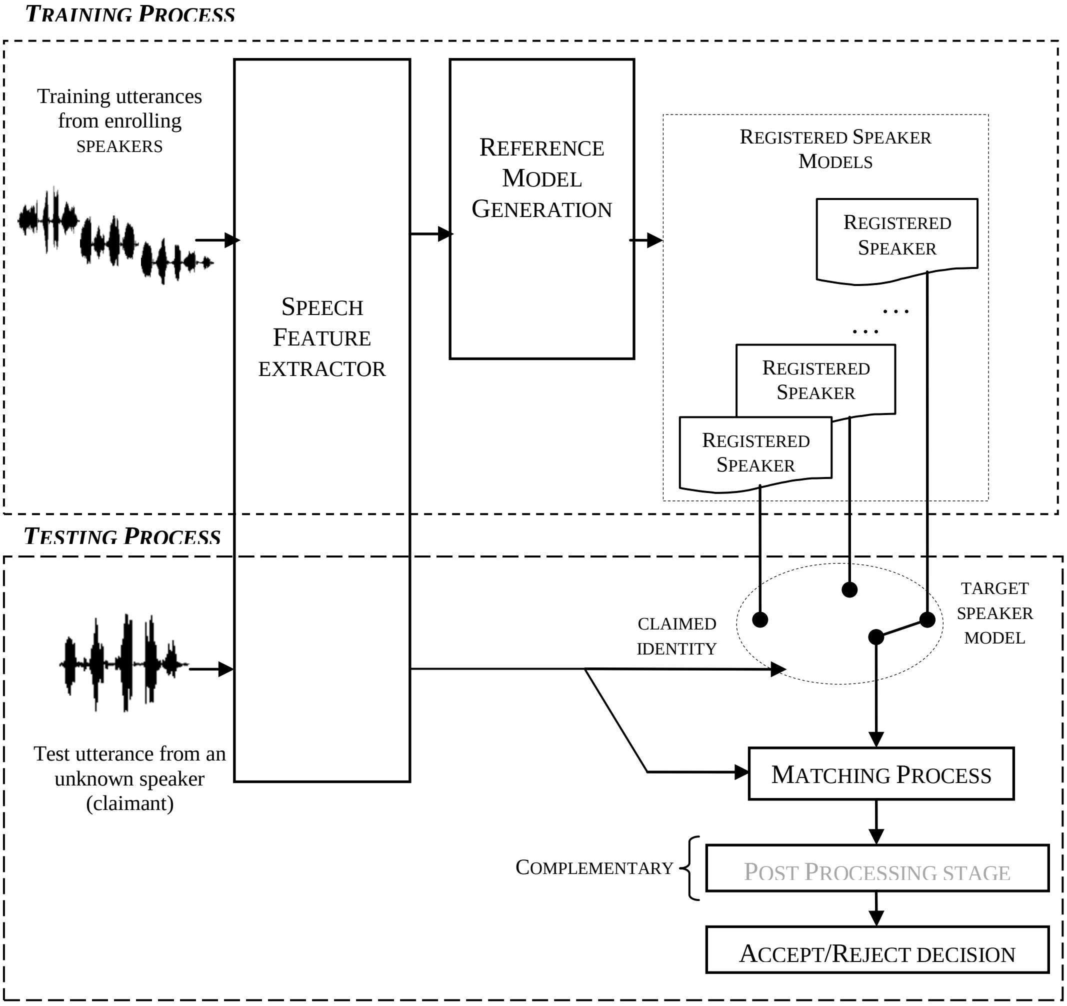


Figure 1.1: General approach to the speaker verification task.

As illustrated in this figure, the first step in the training phase consists of extracting parametric speech features from the enrolment utterances of registered speakers. These parameters provide a more stable, robust and compact representation of the input speech signal in a form which is suitable for the subsequent stages. The second step involves creating reference models of the registered speakers using their extracted feature vectors. Details of the feature extraction and speaker modelling processes are presented in chapters 2 and 3 respectively. During the test phase, speech sample(s) are obtained from an unknown speaker together with a claimed identity. As in the training stage, speech feature parameters are extracted from the given test utterance(s). Then, the extracted speech parameters are tested against the claimed speaker model to obtain a match score. This score indicates the degree of closeness (similarity) between the test utterance(s) and the target speaker model. Following this, the match score can be fed into a complementary post-processing stage in order to increase robustness and reliability. One such post-processing approach, which is commonly employed in speaker recognition, is that of score normalisation. This is carried out to alleviate the impact of adverse operating conditions. The score normalisation process is reviewed in more detail in the next two chapters. Finally, a decision to accept or reject the claimant is made, depending on whether the (normalised) score is higher or lower than a pre-defined threshold.

Speaker identification, on the other hand, is defined as the process of determining the correct speaker from a population of registered speakers. To be precise, this can be considered as a 1: N process where an unknown speaker is compared against a database of N registered speakers to find the best matching speaker. If the process includes the option of declaring that the test utterance does not belong to any of the registered speakers, it is termed open-set speaker identification. Otherwise, it is a closed-set identification process [8-10]. In principle, the process of open-set speaker identification consists of two successive stages of identification and verification. In other words, first, it is required to identify the speaker model in the set, which best matches the given test utterance. Then, it must be verified whether the test utterance has actually been spoken by the speaker associated with the best-matched model or by some unknown speaker outside the registered set. For this reason, open set speaker identification is usually considered as the most challenging subclass of speaker recognition. A modular representation of the open-set speaker identification process is shown in Figure 1.2. The feature extraction, reference speaker model generation, and post-processing modules are identical to the ones used for speaker verification.

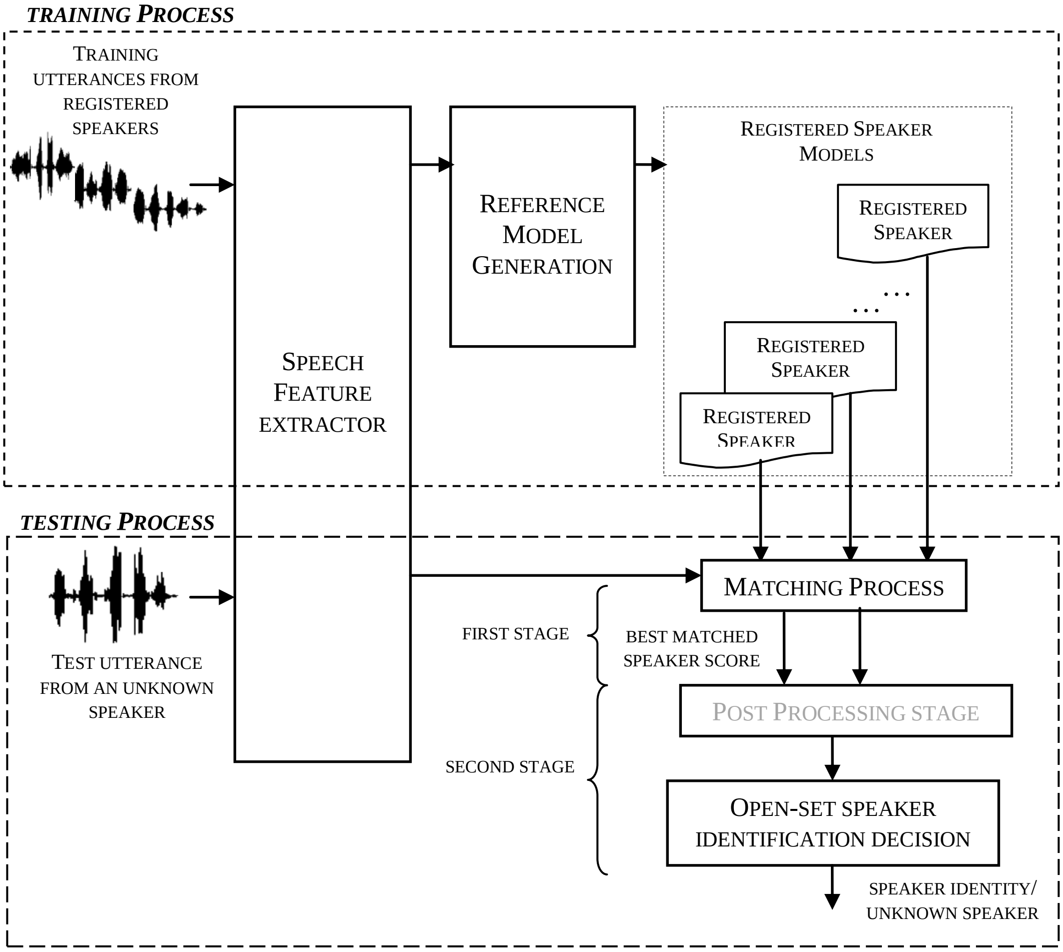

TRAINING PROCESS
TRAINING UTTERANCES FROM REGISTERED SPEAKERS
SPEECH FEATURE EXTRACTOR
REFERENCE MODEL GENERATION
REGISTERED SPEAKER MODELS
REGISTERED SPEAKER
REGISTERED SPEAKER
REGISTERED SPEAKER
TESTING PROCESS
TEST UTTERANCE FROM AN UNKNOWN SPEAKER
MATCHING PROCESS
FIRST STAGE
BEST MATCHED SPEAKER SCORE
POST PROCESSING STAGE
SECOND STAGE
OPEN-SET SPEAKER IDENTIFICATION DECISION
SPEAKER IDENTITY/ UNKNOWN SPEAKER

Figure 1.2: General approach of the open-set speaker identification task.

The speaker recognition process (speaker verification and speaker identification) can be further classified into text-dependent and text-independent tasks. In the former scenario, the utterance which is presented to the system is constrained to a specific textual content (e.g. fixed password or phrase). Conversely, in the latter case, there are no constraints on the textual content of test utterance(s) and the speaker is given full flexibility to say anything (s)he wants.

1.3 Challenges

Over the past two decades, research into voice biometrics has attracted a great deal of interest from the research community [8, 11-16]. This, as discussed in Section 1.1, is mainly attributed to the advantages that speaker recognition applications can offer. However, one of the major problems in speaker recognition remains that of variations in speech characteristics. Such variations can usually be divided into two categories: speaker dependent (or intra-speaker) and speaker independent variations. Speaker dependent variations occur due to various causes such as uncharacteristic sounds by the speakers (e.g. lip smacks, breaths, dry mouth) or physiological factors (e.g. illness, surgery, ageing).

Conversely, speaker-independent variations arise primarily due to technological factors (e.g. channel conditions) and environmental factors (e.g. additive noise) which affect the characteristics of the speech signal when operating under practical conditions. To date, considerable research efforts have been put into developing effective approaches for dealing with variations due to the former. This has recently led to the introduction of methods that have the capabilities of explicitly modelling the effects of channel variability on the given utterances. These approaches therefore allow the effects of channel variations to be compensated during both the training and testing phases, resulting in significant improvements in performance when operating under such adverse conditions

On the other hand, variations due to environmental (additive) noise occur when the background condition of the reference material is different from that of the test material. In practice, the latter problem is usually exacerbated by the mobile nature of many speaker verification applications which in general, considerably increases the likelihood that the speech material may be contaminated by various unpredictable and/or time-varying sources of noise. One typical example is when a user tries to automatically access his/her bank details (i.e. telephone banking) over a handheld device. In this scenario, environmental noises (e.g. phone ringing in the office or door closing in the car) which are not experienced during the training stage can be quite common. These can in turn significantly degrade the test utterances originating from true speakers.

The net result of such variations is a mismatch between the reference (training) and the test material of the same speaker, which can potentially lead to cases of false rejection/false acceptance and therefore affect the overall performance of speaker recognition applications in practice.

1.4 Aim of the research

The aim of this research is to develop effective approaches for speaker recognition under mismatched noise conditions. The theoretical and experimental efforts involved in achieving this goal are based on the specific objectives described below.

The literature review which is carried out in the next chapter reveals that one of the most popular and widely used approaches in the field for dealing with the effects of environmental noise contamination is that of score normalisation. The literature also appears to lack sufficient information on the performance of this method on state-of-the-art speaker verification approaches, when there is a considerable difference between the types and levels of noise degradation in the training and testing data. Hence, the first objective in this study is to thoroughly investigate the effectiveness of score normalisation under these more realistic and practical operating conditions.

Another major objective of the work described in this thesis is that of enhancing the effectiveness of score normalisation under various levels and types of mismatched noise conditions. This is carried out in the context of both text-independent speaker verification and open-set text-independent speaker identification (OSTI-SI). As mentioned earlier, the latter is considered as the most challenging class of speaker recognition because of its additional complexity. Due to the specific characteristics of OSTI-SI, the realisation of the above objective can become computationally expensive when the population of registered speakers grows significantly. Thus, although the main focus of the research work is that of enhancing the effectiveness of speaker recognition under mismatched noise conditions, investigations into approaches for retaining the computational efficiency of OSTI-SI will also be considered in this study.

1.5 Thesis layout

The thesis is organised into seven chapters. A brief description of each of these chapters is given below

Chapter 2: Literature Review

In this chapter, a review of the literature in the area of automatic speaker recognition is presented. This includes a review of the various approaches proposed for feature extraction, speaker modelling and enhancing effectiveness under mismatched noise conditions. A description of the evaluation techniques together with the speech corpora used for the purpose of investigations in this study is also included.

Chapter 3: Techniques for Speaker Verification

This chapter focuses on the techniques which are important in the context of the present study and describes them in detail. This includes a description of the operations involved in the extraction of Linear Prediction-based Cepstral Coefficients (LPCC) together with the details of Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) speaker modelling approaches. The latter part of the chapter also provides a thorough description of the most popular techniques for dealing with mismatched noise conditions.

Chapter 4: Investigations into state-of-the-art Speaker Verification approaches

In this chapter, the state-of-art-the-art techniques which are considered in this thesis (i.e. GMM-UBM and GMM-SVM) are investigated for their effectiveness. The Chapter starts with a description of these techniques and details complementary methods which help to enhance the speaker verification performance. Details of the experimental setup used for comparing the relative effectiveness of these speaker verification approaches are then given. This is followed by a description of the experimental investigations into the relative effectiveness of GMM-UBM and GMM-SVM under both matched and mismatched data conditions.

Chapter 5: Improving the Speaker Recognition accuracy under mismatch conditions

This chapter proposes a new approach for speaker recognition operating under mismatched noise conditions. An account of the motivation behind the proposed approach for speaker verification is given together with details of the experimental investigations to examine its effectiveness in relation to state-of-the-art approaches in the field. Furthermore, the Chapter provides implementation details of the proposed approach and analyses its performance in the open-set text-independent speaker identification (OSTI-SI) context.

Chapter 6: Multi SNR CT-Norm for Speaker Recognition

This chapter presents a new approach for speaker recognition under significant mismatched noise conditions. The Chapter clearly explains the motivations behind this approach and demonstrates its effectiveness in relation to other important methods. The use of the proposed approach is then considered in the context of OSTI-SI. This includes the introduction of a fast realisation of the proposed method in order to enhance the computational efficiency in this case.

Chapter 7: Summary, Conclusions and Future Work

The final chapter summarises the main outcomes in this study and draws overall conclusions. A number of suggestions for future research in the field are also included in this chapter.

CHAPTER 2

LITERATURE REVIEW

Chapter Overview

This chapter provides a background review of various approaches used for the speaker recognition task. It also includes details of techniques used for evaluating the recognition performance. The chapter starts with a brief description of the human speech production process which is considered to be useful in identifying the discriminative characteristics between different speaker voices. Section 2.2 presents a detailed discussion of the commonly used speech features for distinguishing between speakers. The discussion in Section 2.3 is focused on the major speaker modelling and classification techniques currently used for automatic speaker recognition. Section 2.4 then presents a review of commonly used approaches in speaker recognition to achieve robustness against environmental noise. Finally, a description of the speaker recognition evaluation techniques and the speech corpora used in this study are given in sections 2.5 and 2.6 respectively.

2.1 Human speech production

The human speech production system is a complex procedure consisting of closely intertwined psychological and physical aspects. For simplicity, the psychological part can be considered as a three stage procedure. This process is initiated when a speaker decides to transmit a message to the listener(s). The message is then converted into a form (language) that can be understood by the listener(s). Finally, a set of neuromuscular commands is executed in order to control the physical structure shown below [17].

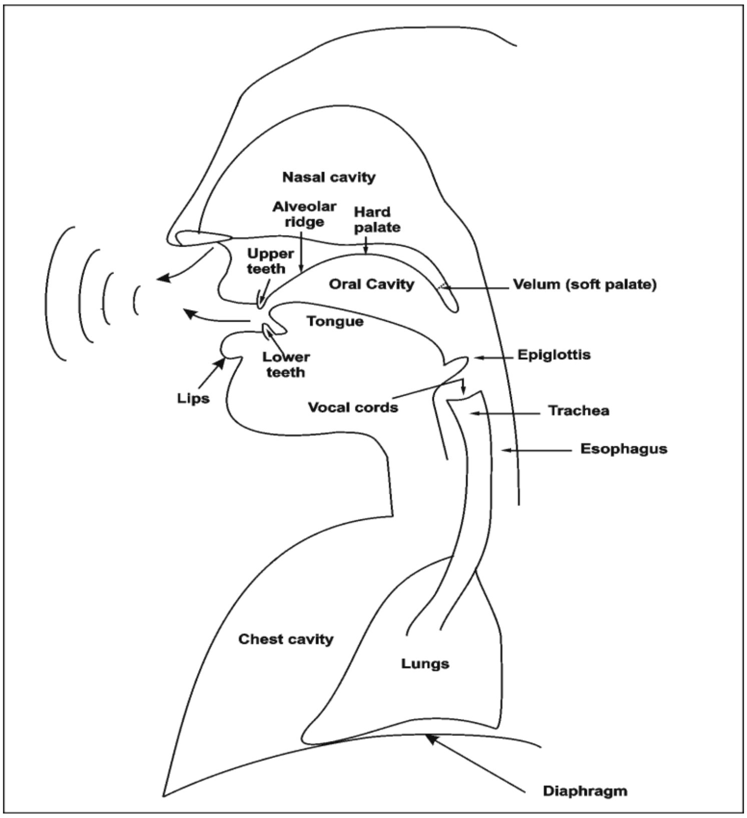


Figure 2.1: Schematic view of the anatomy of the human speech production system [17].

Figure 2.2 illustrates a simplified version of the underlying mechanism involved in the physical aspect of the speech production system [17]. This process can be broken down into three components, namely: power production, tone production and tone resonance.

The power production (or initiation) part is carried out through the respiratory system. During inhalation, the diaphragm is contracted and air is drawn inside the lungs. In order for speech to be produced, the diaphragm is relaxed causing the lungs to recoil and as a result, air is forced out through the bronchi and trachea. The larynx which is made up of the vocal cords[1] is then responsible for the tone production (or phonation) part. When the vocal cords are tensed, the airflow causes them to vibrate at a rate dependent on their length, thickness and tension. This results in a quasi-periodic speech waveform, known as voiced sounds. On the other hand, when the vocal cords are relaxed, the airflow passes through a constriction in the vocal tract which results in an aperiodic (random) speech waveform (commonly known as unvoiced sounds). For voiced sounds, the generated pulse waveform is filtered in the vocal tract to have the harmonics near the natural resonance of the tract. Different sounds are produced by changing the shape of the vocal tract (e.g. by moving the tongue, lips, jaw and velum[2]) so that the natural resonances occur at different frequencies. These resonant frequencies are commonly known as the formants. Finally, the resultant acoustic wave is radiated from the lips.

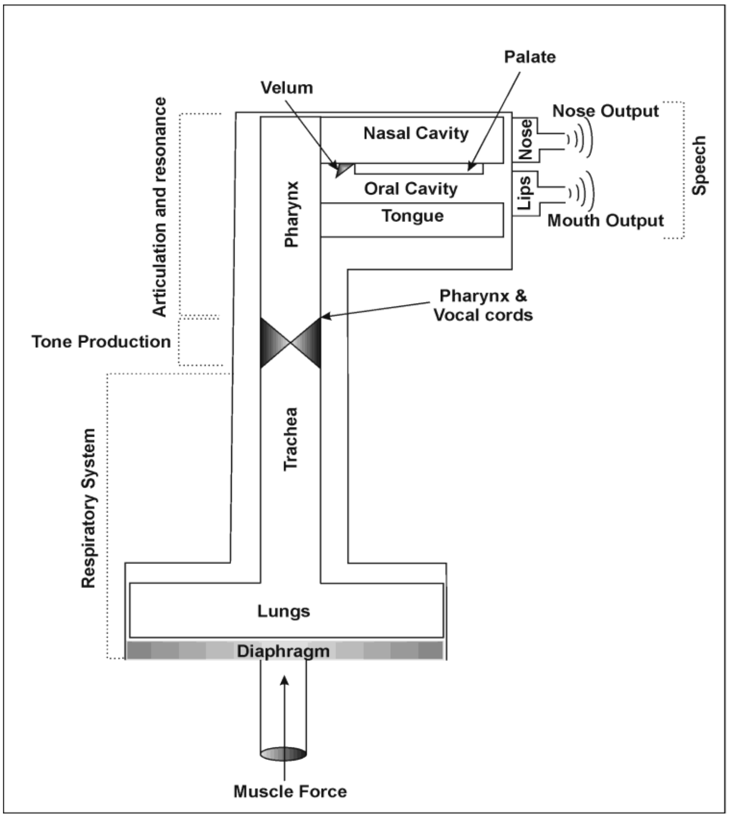
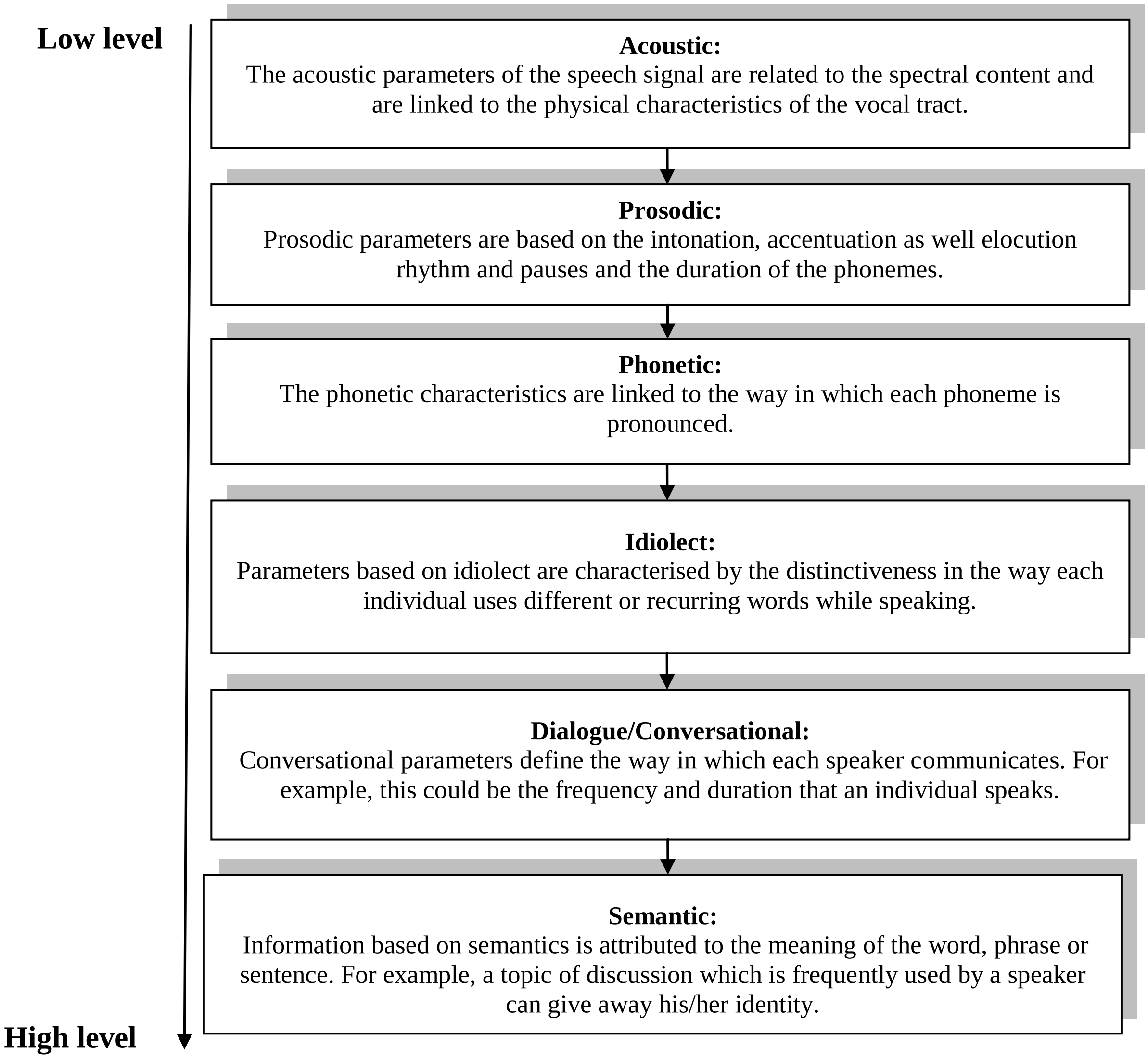


Figure 2.2: Mechanical representation of the human vocal tract [17].

[1] This is also commonly referred to as the vocal folds

[2] The velum, also referred to as the soft palate, can retract or elevate to separate the oral cavity from the nasal cavity.

The above described speech production process is identical for all speakers. In general, however, there are different factors that are based on the speaker's physical and behavioural characteristics, which allow speech waveforms to be discriminative between different speakers. Thus, various levels of information can be extracted from a speaker's utterance to represent these differences. For instance, information based on the natural anatomic variation of the components involved in speech production is considered to be low level. Conversely, information based on acquired traits such as learning and practical use of a language is usually classified as high level [18]. Figure 2.3 illustrates the different levels of information (from low to high) which are represented in a speech signal.



Acoustic:
The acoustic parameters of the speech signal are related to the spectral content and are linked to the physical characteristics of the vocal tract.

Prosodic:
Prosodic parameters are based on the intonation, accentuation as well elocution rhythm and pauses and the duration of the phonemes.

Phonetic:
The phonetic characteristics are linked to the way in which each phoneme is pronounced.

Idiolect:
Parameters based on idiolect are characterised by the distinctiveness in the way each individual uses different or recurring words while speaking.

Dialogue/Conversational:
Conversational parameters define the way in which each speaker communicates. For example, this could be the frequency and duration that an individual speaks.

Semantic:
Information based on semantics is attributed to the meaning of the word, phrase or sentence. For example, a topic of discussion which is frequently used by a speaker can give away his/her identity.

Figure 2.3: Description of the different levels of information which can be extracted from a speech signal.

Since the early days [19-21], features based on the acoustic content of the speech waveform have been the predominant means for tackling the problem of automatic speaker recognition. Recent years have, however, seen the emergence of new speaker recognition systems which utilise high level features together with low level features [22-26]. Although, such systems have shown some relative improvements over traditional speaker recognition applications, they usually require heavy computational front-end processing. For this reason, speech features based on acoustical content, remain the most widely used parameters in most speaker recognition systems [14, 27-29] and are adopted in this research study.

2.2 Speech features for Automatic Speaker Recognition

As mentioned in the previous section, to date, the most commonly used speech features for speaker recognition are based on physical differences of the components involved in the speech production process. Based on the model shown in Figure 2.2, it can therefore be expected that the vocal cords and the configuration of the vocal tract should both contain speaker-dependent information which are useful for discriminating between different speakers. As such, it can be argued that the rate of vibration of the vocal cords, which is characterised by the fundamental frequency[3], F0, of the voiced speech sounds for each individual speaker, should provide a strong set of speech features for speaker recognition. It has, however, been demonstrated in earlier studies that a speaker's pitch can vary considerably due to non-physiological factors such as the emotional state or stress level of the individual [30]. In addition, it has also been reported that a reliable estimation of the fundamental frequency is very difficult to obtain because of its lack of robustness against noise corruption [17, 30].

The configuration of the vocal tract, on the other hand, contains important speaker-discriminative characteristics which are represented in the form of frequency components in the speech spectrum of each speaker. As a result, parameters which represent the vocal tract structure are more robust to adverse factors when compared to those which characterise vocal cords vibrations. In general, in order to extract features which represent the vocal tract configuration, the speech signal is

[3] The fundamental frequency is also commonly referred to as the pitch.

examined within windows of short duration. The reason for this is the slow time varying aspects of the vocal tract which can be considered stationary during short period of time (between 5ms and 100ms) [17, 30].

As shown in Figure 2.4, the short-time spectrum of speech is made up of a convolution of two components. The first component is the spectral envelope which changes slowly as a function of frequency. This is associated with the resonances of the vocal tract as well as the radiation characteristics at the lips and nostrils. The second component which is the spectral fine structure changes rapidly and is associated with the excitation source (or vocal cords vibrations). The aim of most speaker recognition systems is therefore to extract the spectral envelope from the short-term speech spectrum [31].

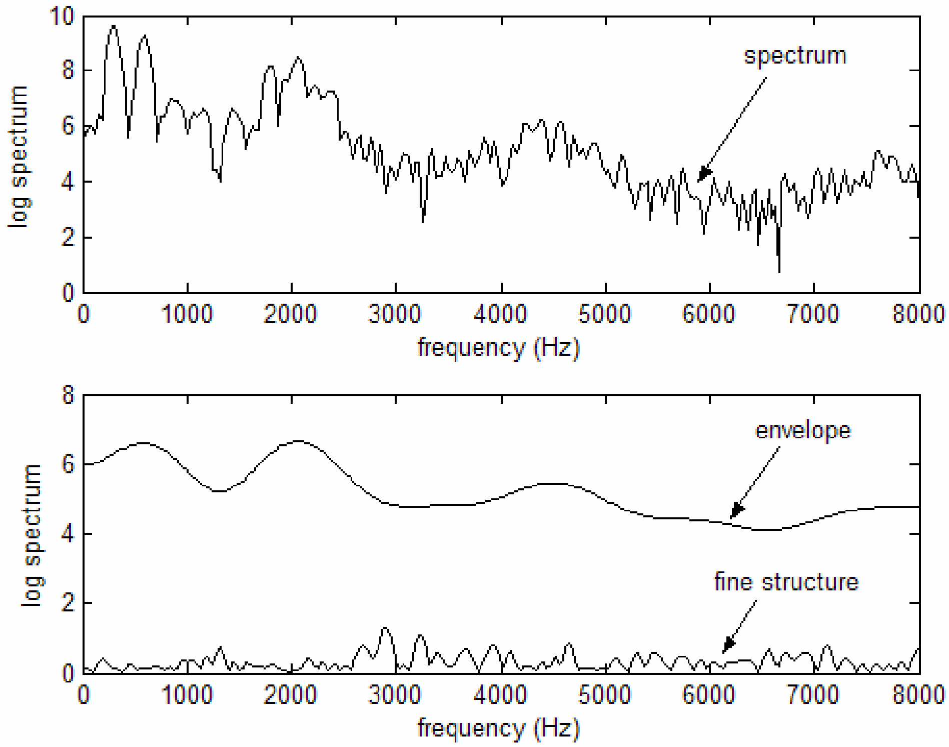


Figure 2.4: Structure of the short-term speech spectrum and the components within it [31].

2.2.1 Commonly used features

The most dominant spectral analysis techniques used to date for automatic speaker recognition are the linear predictive coding (LPC) [32], cepstral analysis and the filter-banks spectrum analysis model [17]. In the LPC analysis approach, the vocal tract is modelled as an all-pole filter. This is based on the assumption that for voiced sounds, the excitation can be represented as an impulse train generator which represents the series of nearly periodic glottal pulses generated by the vocal cords. For unvoiced sounds, a random noise generator is used to represent turbulent air flowing through a constriction along the vocal tract. Each given speech sample is then approximated as a linear combination of the past p samples. The output of the LPC analysis is then given by a vector of predictor coefficients (or LP coefficients), which represent the parameters of the vocal tract configuration for each speech frame. These are obtained by minimising the predictor error.

As mentioned earlier, the speech signal is considered to be a convolution between excitation of the vocal cords (fine structure) and the impulse response of the vocal tract (spectral envelope). Cepstral analysis which is based on the principle of homomorphic[4] signal processing provides an intuitive way of converting the convolutive relationship between the fast and slow varying aspects of the speech spectrum into a summation, thus, allowing easier separation of these two components. As such, the Linear Prediction-based Cepstral Coefficients (LPCC) which can be directly derived from the LPC analysis is widely used to characterise the vocal tract [17, 32].

The Mel frequency-based cepstral analysis provides an alternative approach to obtaining cepstral features. These speech features are referred to as Mel Frequency-based Cepstral Coefficients (MFCC). In this approach, a filter bank is used such that each filter is applied to a different frequency band of the given short-term speech spectrum. The logarithm of the energy in each filter is then computed and accumulated before the Discrete Cosine Transform (DCF) is applied to obtain the cepstral coefficients. There are two different types of cepstral features which can be

[4] This is a generalised term used to describe techniques which involve a non-linear mapping of the signal to a different domain in which linear filter techniques are applied. This is then followed by mapping to the original domain.

obtained using this approach. This depends on the configuration of the processing filters. When the bandpass filters are linearly distributed in frequency, the resulting parameters are known as Linear Frequency-based Cepstral Coefficients (LFCC). In the second, more popular approach, the arrangement of the filters is based on the human perception of speech. This involves the spacing of bandpass filters according to the Mel-scale. As shown in Figure 2.5, on this scale, there is a near linear correspondence between real frequencies and perceived frequencies up to 1 kHz and a logarithmic correspondence for higher frequencies [32]. The feature parameters extracted using this approach are called Mel Frequency based Cepstral Coefficients (MFCC) [17].

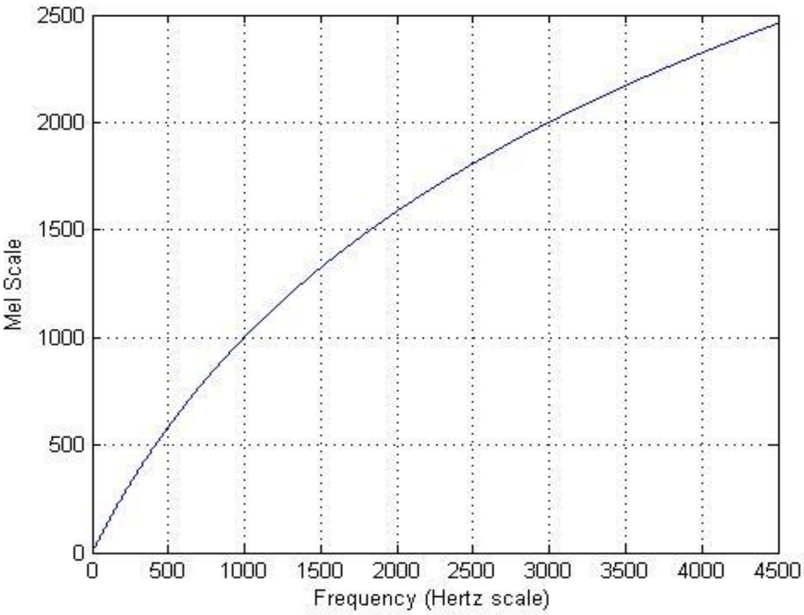


Figure 2.5: Mel-scale representation [17].

Another method for generating perceptually motivated features is through perceptual linear prediction (PLP) [33]. This is carried out in a three-stage process. Similarly to the extraction of MFCC, the short-term speech spectrum is first processed according to the human perception of tones. In this case, however, the centre frequencies of the filters are spaced equally on the Bark scale [34]. The motivation behind the Bark scale is based on the masking phenomenon which is known to affect the hearing of a tone in the presence of another adjacent tone. In the second stage, the PLP analysis compensates for differences between the actual and perceived loudness of tones which occur at different frequencies. The final operation is then based on the all-pole modelling (using the autocorrelation method) of the resulting spectrum to obtain the PLP parameters. Similarly to the LPC approach, the PLP parameters can then be transformed into their cepstral derivatives by using the recursive relationship between the prediction coefficients and the cepstral coefficients [17]. These are known as Perceptual Linear Prediction Coefficients (PLPC) [34]. It should however be pointed out that the PLP analysis has been reported [30] to suppress essential speaker-specific characteristics from the speech spectrum and hence is not a popular choice for speaker recognition applications.

To date, there is no agreement in the literature in relation to the choice of the best set of features for speech applications. This is because speech feature extraction techniques are highly dependent on the specific context in which the features are used. In general, MFCC and PLPC are widely used in speech recognition [35, 36] while LPCC and MFCC are popular choices in speaker recognition. Moreover, studies in speaker recognition have shown that LPCC exhibit better performance when compared to MFCC [13]. These observations are in agreement with the work carried out in [34]. For this reason, in this research work, LPCC is adopted for the parametric representation of speech. A detailed description of the processes involved in obtaining the LPCC is therefore presented in the next chapter.

2.3 Speaker modelling and classification techniques

Given a sequence of feature vectors produced by an unknown speaker, the task of a speaker recognition system is to identify whether that sequence has originated from one of the registered speakers (i.e. speaker identification) or to verify if the sequence has been pronounced by the claimed speaker (i.e. speaker verification). To achieve either of these, speaker models are usually constructed, during the training stage, using the features obtained from the speech signal of the registered population of speakers. During the classification stage, the test utterance from an unknown speaker is then matched against the registered speaker model(s) to obtain an utterance score which indicates the degree of correspondence. This section presents a general description of the various modelling techniques used in speaker recognition systems.

2.3.1 Vector Quantisation (VQ)

Vector quantisation, which is also known as a centroid model, can be considered as one of the simplest classification models for speaker recognition [37, 38]. The approach involves building speaker models by partitioning the feature vectors into *K* non-overlapping clusters which individually represent different acoustic classes. Each cluster is represented by a code vector which is the centroid (average vector) of that cluster. A speaker model in the VQ approach is therefore, a collection of centroid vectors which is commonly referred to as a codebook. This approach provides an effective way of reducing the data storage requirements while preserving the fundamental aspect of the original distribution [37]. The two most effective algorithms for generating the codebook are based on the Linde-Buzo-Gray (LBG) algorithm [39] and the Distortion Driven Cluster Splitting algorithm [34]. During the classification stage, the distance of each of the extracted feature vectors of the test utterance to its nearest codebook vector is accumulated to obtain an utterance score.

2.3.2 Gaussian Mixture Model (GMM)

A Gaussian Mixture Model is the representation of various acoustic classes in a speaker's voice using a linear combination of Gaussian Probability Density Functions (or components/mixtures). This can be considered as an extension of the VQ approach, in which the clusters are allowed to overlap with each other. Each speaker GMM is represented by the mean and covariance statistics of the mixture densities, and the weight associated with each of them. There are two commonly used approaches for obtaining these parameters from the registered speaker's training data. The first method is that of computing the model parameters using the iterative Expectation-Maximisation algorithm (EM) [40]. The second approach involves developing a Universal Background Model (UBM) and then adapting this using the given training data, and through a modified realisation of the Maximum a Posteriori (MAP) [12]. The UBM development is based on the EM algorithm and the use of utterances from a large population of speakers. During the classification (or testing) stage, the test data is compared to the claimed speaker model or the registered set of speaker models using the maximum likelihood rule. Over the last decade, the said technique, which is commonly referred to as GMM-UBM, has been one of the predominant speaker modelling approach for text-independent speaker recognition [5, 6, 11, 12]. For this reason, the approach is adopted in this study. Further details of the GMM modelling approach are provided in the next chapter.

2.3.3 Hidden Markov Model (HMM)

The GMM approach described above can be considered as a static model which does not model variations in time. The Hidden Markov Model (HMM), on the other hand, has the additional capability of being able to model the temporal variations between the various acoustic classes [17]. A HMM may be described as a finite state generator. In speaker recognition, each of these states may represent phones or larger units of speech. At discrete times, the system undergoes a change of state according to a set of probabilities associated with it. After each transition, an output is emitted from the current state. Although such outputs can be observed, the associated states are ‘hidden’ and can only be inferred from the available outputs. The temporal information between the acoustic classes is encoded by moving from state to state along the allowed transitions. The amount of time spent in each state accounts for variability in speaking rate and is therefore dependent on the training data. For a thorough review of the theory and implementation of HMM, the interested reader is referred to [17].

In general, HMM has been mainly used in speech recognition applications [17, 41-44]. However, to date, several studies have considered the use of HMM for text-dependent and text-independent speaker recognition [45-47]. The study in [48] has shown that in the text-independent scenarios, the sequencing of acoustic classes is not important since it contains limited speaker-dependent information. Such findings have also been confirmed in the experimental studies in [49] and [50] which have found that the text-independent performance is unaffected by discarding the temporal information in the HMMs.

2.3.4 Artificial Neural Networks (ANN)

An artificial neural network (ANN) [51-53] is a discriminative classifier which is made up of a collection of simple adaptive processing units (or nodes) that can collectively accomplish complex machine learning tasks. These nodes can be considered as analogous to the neurones which are present in the human central nervous system, although as expected, the complexity of the artificial neural network is far less than that of the human brain. Each processing unit computes the weighted sum of the inputs and passes the results through a sigmoid-like nonlinearity. ANN is a powerful tool which can be used for both regression and classification tasks. Although there are many different types of ANN, to date, the multi-layer perceptrons (MLP) has been the most commonly used architecture for speaker recognition [54-56]. As shown in Figure 2.6, an MLP is made up of a network (multi-layers) of simple nodes which are known as perceptrons. The underlying concept of the MLP is based on a two-stage process. First, a linear weighted sum of its input connections is computed. Second, a non-linear function (also known as activation function) is applied in order to compute the output of the node. It is generally acknowledged that, given a sufficiently large number of nodes in the hidden layer, an MLP with a non-linear activation function can approximate any non-linear mapping between the input and output [53-55]. For the speaker verification task, an MLP has only one output node. This is because in this case, the objective is to obtain a score over all the frames of the given test utterance.

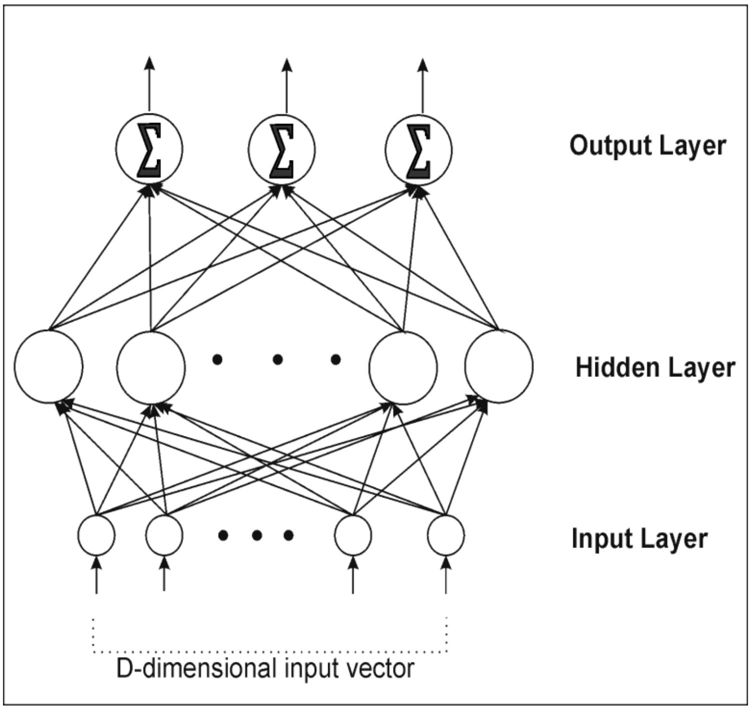


Figure 2.6: Multi-layer perceptron architecture.

2.3.5 Support Vector Machines (SVMs)

The support vector machine (SVM) is another discriminative binary classifier which involves modelling the linear boundary between two classes as a separating hyperplane [57-59]. In speaker verification, one class consists of the target speaker's training vectors (labelled as +1) and the other class consists of training vectors from a large number of background speakers (labelled as -1). SVMs can also learn non-linear boundary regions between samples by mapping the input samples into a higher dimensional space. This is carried out through the use of kernel functions. A separating hyperplane is then chosen (in the higher dimensional space) in such a way as to maximise its distance from the closest training samples, known as support vectors. During the test stage, a classification score is then obtained by evaluating the distance of the test sample in relation to the hyperplane. This approach has been increasingly used in recent years for the speaker verification task [22, 60, 61] and has been shown to give the state-of-the-art performance. For this reason, SVM is adopted in this research work and a detailed description of its fundamental concepts is given in Chapter 3

2.3.6 Hybrid modelling techniques

In general, the underlying concept of generative approaches such as GMMs and HMMs is that of estimating probability densities to model the underlying characteristics of the speaker's voice based on the given training data. On the other hand, discriminative approaches such as SVMs and MLPs usually involve modelling the boundary between classes and discard any information which is not considered to be useful for classification. It has been reported in the literature [31, 61] that generative modelling approaches have important features which discriminative modelling approaches do not possess and vice-versa. Table 2.1 presents a comparison of those complementary features.

Favourable Features	**Generative Model**		**Discriminative Model**	
1. Ability to deal with impostors not present during the training stage.	Creates a full model of the registered speaker voice independent of the availability of impostor utterances. This allows the model to be more robust to impostor attacks.	✓	Discards information which is considered unnecessary for modelling the boundary. This process makes the model vulnerable to impostors not present during the training process.	✗
2. Ability to deal with data of arbitrary length	A generative model is built by clustering feature vectors irrespective of the length of training data.	✓	Discriminative models cannot deal with sequences of varying length during the training or testing stages.	✗
3. Small storage capacities	Obtaining a full model of a speaker's voice requires large storage capacities[5].	✗	Modelling only class boundaries results in smaller more compact models.	✓
4. Modelling algorithm should NOT over-tune to the training data.	Generative approaches attempt to model all the underlying variations of the training data.	✗	Discriminative models focus on modelling the boundary between classes.	✓

Table 2.1: Advantages and disadvantages of the generative model and the discriminative modelling methods.

It is therefore not surprising that a significant amount of work has been carried out over the last few years into approaches for combining the two modelling strategies to obtain a robust classification method. The most popular techniques to achieve this involve either a combination of a generative model with a discriminative classifier or using a discriminative objective function to adjust the parameters of a generative model. Such approaches include the Radial basis function (RBF)

[5] For GMMs, the storage requirements depend on the number of mixtures used for modelling the speaker's data.

networks [62], GMM/SVM combinations [28, 63] and HMM/MLP hybrids [64]. A brief description of each of the said approaches is provided below.

a) Radial Basis Function (RBF) networks

The RBF network combines the generative modelling strategy of GMMs with the discrimination capabilities of the MLP [53, 62]. Mathematically, an RBF network is almost identical to that of several GMMs. It has a two layer topology, similar to that of the MLP. The output layer of an RBF network is exactly the same as a MLP. In this case, however, the nodes of the hidden layer each consist of a unimodal Gaussian (or Gaussian basis functions). An example of an RBF with 1 output (applicable to the speaker verification task) is illustrated in Figure 2.7.

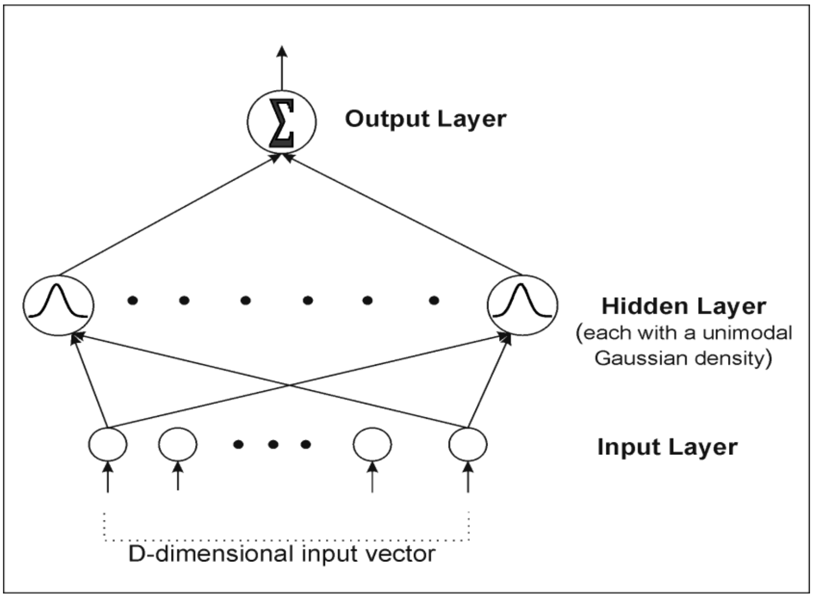

Output Layer
Hidden Layer
(each with a unimodal Gaussian density)
Input Layer
D-dimensional input vector

Figure 2.7: A radial basis function (RBF) network with one output.

There are two ways in which an RBF network can be trained. Firstly, it can be trained entirely by minimising the empirical risk using the gradient descent algorithm [53]. This results in a completely discriminative RBF model. Conversely, the network can also be trained using a combination of gradient descent and the expectation-maximisation algorithm. The former approach is used for learning the weights of the output layer while the latter technique is employed to obtain the means and covariances of each Gaussian. Thus, the second approach creates a network which benefits from the generative nature of the Gaussians while the preserving the discriminative nature of the output weights.

b) HMM/MLP

The HMM/MLP hybrid, which has been proposed in [54, 64, 65], combines the efficient temporal processing features of HMMs with the discriminative capabilities of the MLP. It is shown in these studies that such a combination results in a system which yields better recognition performance than approaches based on only the MLP or HMM. This is because the MLP has limited segmentation capabilities which restrict its effectiveness for speech/speaker recognition tasks. Various approaches have been proposed in the literature to tackle this issue for both speech and speaker recognition systems [54, 64, 66]. The fundamental concept behind most of these techniques is to replace the HMM state observation probabilities (or likelihood) with scaled probabilities estimated using an MLP. In other words, the MLP is used to estimate posterior probabilities and these are then scaled by the prior probability for each of the HMM states (or GMMs) and incorporated into the training scheme. During the classification stage, the posterior probability of the utterance is then obtained instead of the likelihood.

c) GMM/ SVM

To date, several approaches which combine the GMM and SVM modelling strategies [28, 63, 67-70] have been proposed in the literature. This section presents a review of two popular approaches which employ such combinative techniques for speaker recognition. The reason behind the first approach is based on the assumption that the conventional computation of the log-likelihood ratio [67] is not optimal because the probabilities cannot be estimated accurately. Thus, in this approach, during the training phase, the GMM log-likelihood scores which are obtained from the registered speaker and the UBM are fed as a two-dimensional vector into an SVM. Adjustable parameters are then obtained as the output of the SVM and these are then incorporated into the computation of the GMM log-likelihood ratio to obtain a more reliable utterance score during the test phase.

Another widely used approach proposed in [28] is based on the use of a concatenation of the means (known as supervectors) from the registered speaker’s GMM and the GMMs for a large set of impostors to train the SVM speaker model. During the classification phase, a supervector of means is extracted from the test model and this is then compared to the client's SVM model to obtain a classification score. This approach which is also known as SVM based on GMM supervectors of means has been shown to give the state-of-the art speaker recognition performance [14, 28, 71]. For this reason, this approach is adopted for the purposes of the work described in this thesis. An account of the procedures involved in GMM/SVM approach is given in Chapter 3.

2.4 Noise robustness techniques in speaker recognition

A factor adversely affecting the accuracy of speaker recognition systems in practice is that of variations in speech characteristics. Such variations occur due to various causes such as environmental noise, channel effects, or uncharacteristic sounds by speakers (e.g. lip smacks). The net result is a mismatch between the corresponding test and reference material for the same speaker, which in turn reduces the accuracy in speaker recognition. As mentioned in Chapter 1, the main focus of the work described in this thesis is to deal with mismatch conditions which result from environmental noise (additive noise). To date, several approaches have been proposed in the literature to tackle the impact of environmental noise on speaker recognition. An overview of the most commonly used approaches, which can be categorised based on the level at which they operate, is given in the next sub-sections. For an extensive review of the techniques given in each category, the interested reader is referred to [72-74].

2.4.1 Speech level approaches

Approaches which operate at the speech or acoustical level have been originally proposed in the speech enhancement literature and later used in speaker recognition in order to achieve robustness under noise conditions [72]. Such approaches, which aim to improve the signal-to-noise ratio (SNR) of the input speech signal, can be further classified into single-channel methods and multi-channel methods. The former category assumes that the speech and noise data are available in a single mixed form (e.g. single microphone). On the other hand, the latter category, assumes that the speech and noise are available in various combinations due to the availability of multiple signal inputs. The primary focus of the work carried out in this thesis is that based on the assumption that the speech signal is captured using a single microphone. As such, multi-channel approaches are considered to be outside the scope of this thesis.

In general, research into single channel speech based approaches has targeted the impact of environmental noise on speaker recognition through filtering techniques [72, 75]. These approaches usually produce estimates of the 'enhanced' short-time speech spectra by filtering out the noise components. This is carried out by using *a priori* knowledge of the statistics (e.g. power spectra or variances, or signal-to-noise ratio) of the noise and clean speech signal. Some commonly used techniques in this category include spectral subtraction [76, 77], Wiener filtering [78, 79] or Kalman filtering [80-82]. Although these approaches have been reported to be effective when dealing with stationary and slowly-varying types of noise, they are usually less reliable for non-stationary noise.

2.4.2 Feature level approaches

Feature level approaches for tackling environmental noise are based on the general assumption that in practice, the features representing the speech signal can be divided into two different subspaces. The first, 'noisy' speech subspace represents unreliable or missing features while the other, 'speech' subspace, consists of reliable or present features. The aim of feature based approaches is therefore that of estimating or detecting those missing/unreliable features in order to compensate, discard or deemphasise them during the recognition process. To date, several feature level methods have been proposed for this purpose [76, 83-87]. Such approaches can be grouped into two main categories. The first category involves the estimation of the noise signal in order compensate for the unreliable features. Recently, a study of this category of approaches has also suggested that when knowledge of the noise is insufficient or cannot be reliably estimated for enhancing the speech data, an alternative approach is to completely ignore the severely corrupted speech data segments. The recognition process is then solely based on the portion of the speech signal which is considered to contain little or no contamination [85]. Other approaches have shown that, during the matching stage, only speech features vectors which generate reliable scores should be kept for computing the overall likelihood score to improve accuracy under noisy conditions. Commonly used feature-level approaches in each category include missing feature theory [76, 83, 84, 88] and feature score pruning techniques [85, 86, 89] respectively. However, these approaches tend to be effective only when there is partial noise corruption of the signal and in some cases they can also lead to the removal of useful speaker discriminative information.

2.4.3 Model level approaches

In general, approaches which operate at the model level tackle the effects of noise conditions on speaker recognition by minimising the mismatch between the reference model for the target speaker and the test material such that they have the same noise characteristics. These techniques can be grouped into two categories. The first category includes approaches which are based on an estimation of the noise characteristics during the training and/or testing stages in order to minimise the mismatch. On the other hand, the other category consists of approaches which rely on multiple training which represent various noisy conditions to build several statistical models for the same speaker. During the test phase, the model which best matches the characteristics of the speech signal and therefore yields the highest likelihood score is chosen for recognition. Commonly used approaches in each group include parallel model combination (PMC) [90-92] and multi-SNR methods [87, 93, 94] respectively. Techniques in the former category have been reported to provide significant improvements in the relative effectiveness of speaker recognition applications when operating under mismatch conditions between the training and test material. For this reason, a PMC approach is adopted in this research study and further discussed in Chapter 3.

2.4.4 Score level approaches

In general, approaches which operate in the score domain aim to alleviate the effects of variations in the characteristics of the speech signal caused by environmental noise by reducing the overlap in the score distributions for the target speaker and impostors. To date, the most widely adopted approaches in this category have been based on score normalisation [5, 73, 95, 96]. Such approaches can be further divided into two distinct groups.

The techniques in the first category are derived using the Bayesian equation for likelihood estimation. In this approach, the matching score obtained from a registered speaker model is normalised with the score obtained from a UBM [96] or a cohort of background speaker models [73, 97]. The second set of approaches is based on standardisation (or distribution scaling) of the score distribution. Two of the most popular techniques in this category are Zero Normalisation (Z-Norm) and Test Normalisation (T-Norm) [73, 96, 98].

Score normalisation has been reported to be highly effective under practical operating conditions particularly when accurate information about the existence, level and nature of variations in speech characteristics is unavailable [73, 95, 99]. For this reason, this approach is adopted in this study and a mathematical perspective of the above score normalisation techniques is given in Chapter 3.

2.5 Speaker Recognition Evaluation Techniques

In Section 2.3, various techniques for obtaining registered speaker models from their training speech utterances have been reviewed. For speaker verification, during the classification stage, the pattern matching algorithm compares the test utterance against the claimed speaker model. A measure of similarity, which is usually given in terms of an utterance score, is then computed. This score is then used to decide whether to accept or reject the identity claim. In the speaker identification scenario, the test utterance is compared against all the registered speaker models in order to determine the identity of the speaker.

In theory, the ideal speaker verification system needs to be able to accept all identity claims made by clients and reject all those made by impostors. In reality, however, due to various adverse factors (described in Chapter 1), this does not always occur. In fact, there are four different decisions which are usually made. As shown in Table 2.2, based on statistical hypothesis testing, this can result in two types of errors: type I (False Acceptance) and type II (False Rejection).

Possible Decisions		Type of Errors
1. Accept a client	✓	N/A
2. Accept an impostor	×	**Type I**: False Acceptance (FA)
3. Reject a client	×	**Type II**: False Rejection (FA)
4. Reject an impostor	✓	N/A

Table 2.2: Speaker recognition decisions.

The utterance scores of a client model are usually made up of two overlapping Probability Distribution Functions (PDF). The first PDF represents the scores obtained when the client targets his/her own model while the other represents scores obtained when impostors target the registered client model. A threshold must then be set such that it attempts to minimise the number of errors (FA or FR) made by the system.

In order to quantify the system performance into a single measure, the verification performance is obtained in terms of Equal Error Rates (EER) [100]. This is the error rate that occurs when the threshold is set such that the rate of false-accepts (2.1) is equal to the rate of false rejects (2.2).

$$\text{False Acceptance Rate (FAR)} = \frac{\text{Number of FAs}}{\text{Number of impostor trials}} \tag{2.1}$$

$$\text{False Rejection Rate (FRR)} = \frac{\text{Number of FRs}}{\text{Number of client trials}} \tag{2.2}$$

The trade-off between FAR and FRR can be graphically represented by a Receiver Operating Characteristics (ROC) curve [101]. As illustrated in Figure 2.8, in this curve, the FAR is plotted on the horizontal axis while the true acceptance rate, (equivalent to the FRR subtracted from one hundred), is given on the vertical axis. The area under the curve is a measure of the performance of the system. Another approach for illustrating the system performance is the Detection Error Trade-Off (DET) plot. An example of the DET plot is shown in Figure 2.9. In this case, the FAR is plotted on the horizontal axis while the FRR is represented on the vertical axis. The curves are plotted using the normal deviate scale [101]. As a result, approximately linear curves are produced, making it easier to visualise relative differences between different classifiers. For this reason, the DET plot is adopted in this research study.

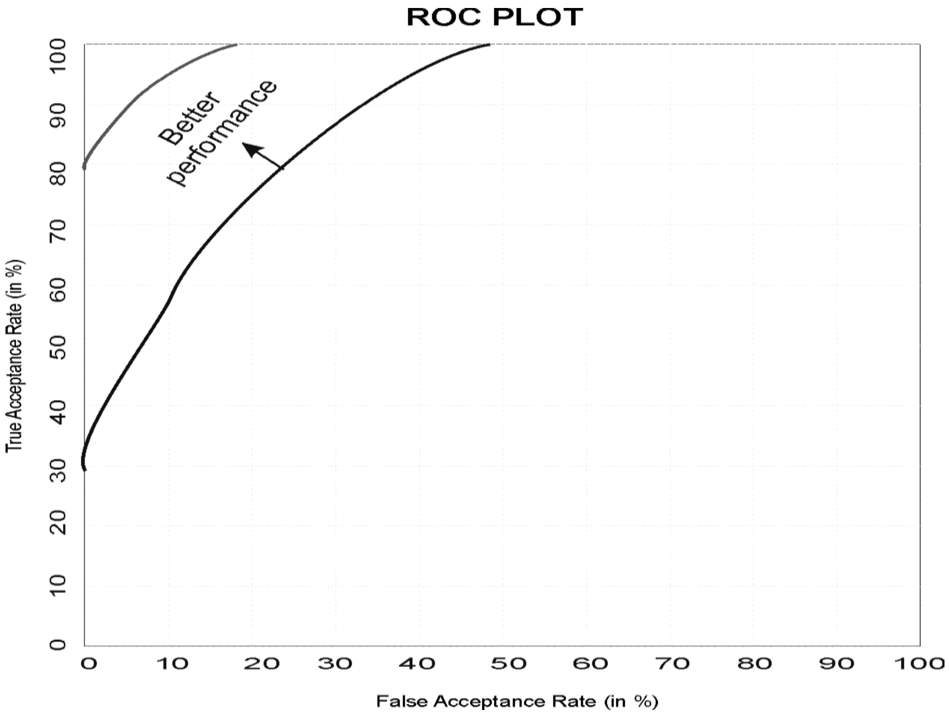


Figure 2.8: Illustration of ROC curves [101].

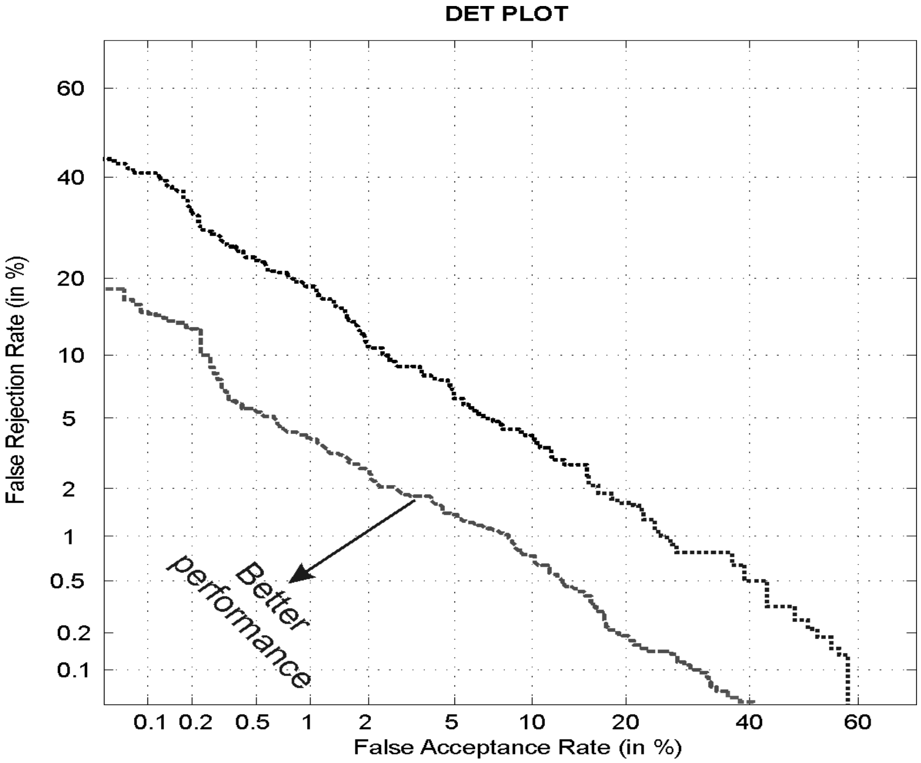


Figure 2.9: Illustration of DET plots [101].

As discussed in Chapter 1, the speaker identification task can be subdivided into two categories of closed-set and open-set identification respectively. The closed-set identification is the process of identifying a person from a group of known (registered) speakers. On the other hand, in the open-set identification problem, the test utterance may or may not belong to one of the known (registered) speakers.

Open-set identification consists of two stages of closed-set identification and verification. The performance of the verification stage is evaluated using the approach discussed above for speaker verification. In this case, the verification performance is expressed in terms of Open-Set Identification Equal Error Rate (OSI-EER) while the identification performance is expressed in terms of Identification Error Rate (IER).

This is evaluated as follows:

$$\text{IER} = \frac{\text{Number of incorrectly identified client speakers}}{\text{Number of client trials}} \text{ x } 100\ \% \qquad (2.3)$$

Finally, it should be noted that an estimate of the 95% confidence interval (CI_{95}) is also presented for the various EERs and OSI-EERs obtained as a result of the experimental investigations in this study. This is given by [102]

$$CI_{95} = \varepsilon \pm 1.96 \sqrt{\varepsilon\,(100 - \varepsilon\,)/\tau} \quad , \qquad (2.4)$$

where ε is the EER or OSI-EER in percentage and τ is the number of true speaker tests.

2.6 Speech and Noise databases

This section reviews the two speech corpora which have been adopted for the purposes of the research study in this thesis. These are the NIST Speaker Recognition Evaluation (SRE) 2003 database [103] and the TIMIT database [104].

2.6.1 NIST SRE 2003

The NIST SRE 2003 [103] is part of the ongoing evaluation databases developed for conducting yearly evaluations of the state-of-the art speaker recognition systems. A brief summary of the NIST SRE 2003 is given below:

1. Each speech file is recorded on one side of a telephone conversation with a sample rate of 8 kHz.
2. The database is made up of 11,839 speech utterances which amount to around forty-six hours of speech.
3. The speech data was compiled from the LDC’s CALLFRIEND, CALLHOME and Switchboard-2 corpora.
4. The training utterances are about three minutes long while the test utterances are between three and thirty seconds in duration.

2.6.2 TIMIT

The TIMIT corpus [104] is designed to provide speech data for the development and evaluation of automatic speech recognition systems. The database has been recorded at Texas Instruments (TI), transcribed at Massachusetts Institute of Technology (MIT) and verified and prepared by NIST. The database can be summaries as follows:

1. It contains recordings of 630 speakers of eight major dialects of American English.
2. Each speaker pronounces ten phonetically rich sentences.
3. The speech data have been recorded at a sampling rate of 16 kHz.
4. All the utterances are gathered under clean (noise-free) environment.

It is also important to point out that it would have been more beneficial to use a larger database for the purposes of the experimental investigations in this work. The main reason for using the TIMIT database is that it remains amongst the only widely used and readily available speech corpora which comprises speech utterances recorded under clean conditions (i.e. without noise contamination or handset variability) [88, 105, 106]. Thus, the TIMIT database offers the flexibility required for investigating the relative effectiveness of the proposed approach under controlled noise conditions while enabling the fast and easy replication of the results. In addition, since the focus in this study is the effects of background (additive) noise on the speaker verification accuracy, the choice of dataset must be such that it allows freedom from convolutive noise (e.g. channel noise), which is the case with the TIMIT database.

Moreover, in this thesis, the NOISEX 92 [138] and the BT Piper [34] databases are utilised for simulating the effects of mismatched noise conditions caused by additive noise on the speaker recognition accuracy. These databases contain various types of noises, recorded in real-life situations using either a land or a cellular telephone. The digitisation of these databases is based on the use of a sampling frequency of 16 kHz. The actual noise files deployed are car and office noise from the NOISEX92 database and factory noise from the Piper database. These provide a general representation of the stationary and non-stationary nature of additive noises which can be expected during either the training or test stages.

2.7 Chapter Summary

This chapter has presented an overview of the major techniques used in speaker recognition. A description of the human speech production mechanism is given, which provides an overview of the speaker-specific characteristics of voice. The literature review has revealed that most of the speech parametric representations are based on the short-term spectral analysis. The most appropriate parametric representations have been reviewed. The Linear Prediction-based Cepstral Coefficients (LPCC) approach is chosen in this study as a suitable representation.

The literature review has also covered various approaches used for representing registered speakers and classifying the test utterance in speaker recognition. From the literature, it is clear that hybrid modelling approaches have become the most commonly used classifiers. Amongst these, the “SVM based on GMM supervectors of means” approach is found to be the most popular and effective. The literature review has also covered various techniques that are typically used for dealing with degradation in speech caused by environmental noise. It is also clear from the literature that model and score level approaches are amongst the most appropriate methods for introducing robustness when dealing with mismatched data conditions.

Finally, a description of the techniques which are commonly used for evaluating the performance of a speaker recognition system is presented. This is then followed by a review of the speech corpora used in this study.

CHAPTER 3

TECHNIQUES FOR SPEAKER VERIFICATION

Chapter Overview

The previous chapter has presented a review of various important techniques in speaker recognition. This chapter focuses on the techniques which are important in the context of the present study and describes them in detail. It is clear from the literature review that the most popular parametric representation of speech for speaker discrimination is the cepstrum. The Chapter starts with a description of the pre-processing requirements for the purpose of speech feature extraction. The operations involved in the extraction of Linear Prediction-based Cepstral Coefficients (LPCC) which is the choice for parametric representation of speech in this study are then covered in Section 3.2. The discussions in Section 3.3 are focussed on the techniques used for speaker modelling and classification. This is followed by a description in Section 3.4 of the techniques adopted in this study for dealing with mismatched noise conditions in the context of speaker verification.

3.1 Front end processing

As discussed in the previous chapter, the short-term spectrum is the most appropriate and widely adopted speech representation for use in speaker recognition. In general, most speaker recognition applications employ a front-end processing unit in order to characterise the speech signal in this manner. This, as shown in Figure 3.1, consists of a series of pre-processing steps followed by a speech feature extraction unit. This section focuses on the various operations involved in the extraction of Linear Prediction-based Cepstral Coefficients (LPCC), which is employed in the present study. The discussion starts with a brief description of the pre-processing stages.

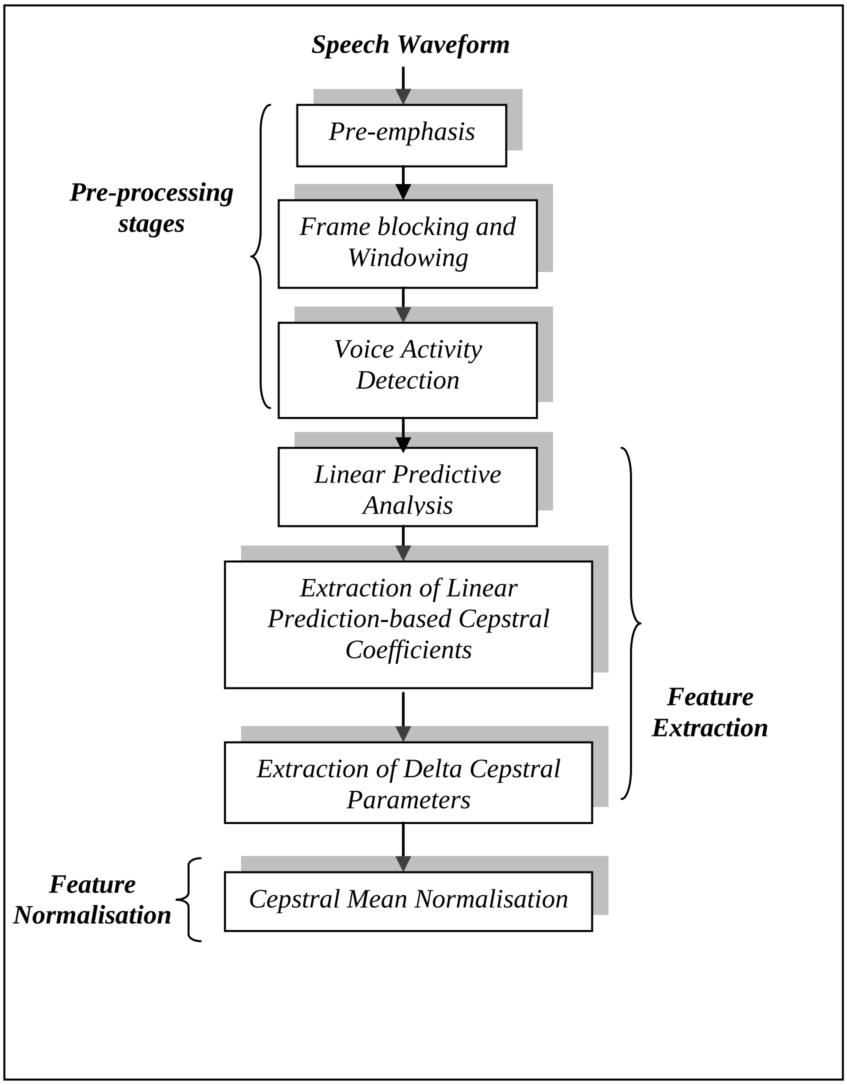

Speech Waveform
Pre-processing stages
Pre-emphasis
Frame blocking and Windowing
Voice Activity Detection
Linear Predictive Analysis
Extraction of Linear Prediction-based Cepstral Coefficients
Extraction of Delta Cepstral Parameters
Feature Extraction
Feature Normalisation
Cepstral Mean Normalisation

Figure 3.1: Steps involved in the LPCC feature extraction process.

3.1.1 Pre-processing

It is known that, due to the physiological characteristics of the human speech production system, the speech signal experiences a spectral roll-off of about -20dB/decade [17]. It is, therefore, desirable to compensate for this degradation by pre-processing the speech signal. This involves filtering the sampled speech signal using a first-order Finite Impulse Response (FIR) high-pass filter. The application of this filter results in a spectral lift of the short term spectrum of speech for the high frequency components. Additionally, another important motivation for the use of the said filter is the prevention of numerical instability in the LP analysis [107].

The transfer function of the high-pass filter is given as [108]

$$H(z) = 1 - \alpha z^{-1} \tag{3.1}$$

The constant α controls the degree of emphasis. A typical value of 0.95 is chosen in this study.

In the next step, the speech samples are grouped into frames of about 10-30 ms where the signal is considered to be stationary. This operation is known as frame blocking and reflects the short-term nature of the speech signal under analysis. The frame blocking process can be considered as multiplying the speech signal by a rectangular window which is zero everywhere except during the analysis period. The problem with this approach is that it introduces discontinuities at the edges of the frame which in turn leads to the distortion of the short-term speech spectrum by unwanted high frequency components. In order to minimise these adverse effects, a better approach is to multiply the speech signal by a Hamming window [17]. This is defined as

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), & for\ 0 \leq n \ \leq N-1 \\ 0 & , \ otherwise \end{cases} \tag{3.2}$$

In this approach, however, a low weighting is applied to samples that lie near the ends of the Hamming window, regardless of whether they represent a significant speech event or not. In such cases, the speech events in question will not be effectively featured in the speech analysis. To overcome this issue, the adjacent segments are usually overlapped so that any event will be covered by at least two overlapping windows. The typical duration of overlap is 50% of the length of the window. It is clear that this approach allows a speech event which is at the edge of one window to be weighted appropriately in the following window.

Finally, the last pre-processing step involves removing frames which contain silence from the input signal. This process, which is commonly known as Voice Activity Detection (VAD) [109, 110], is very important as it allows the speaker verification application to focus on speaker-dependent speech segments only and therefore and is not adversely affected by low energy frames or non-speech frames. In this work, an energy-based VAD which is detailed in [110] is utilised.

3.2.2. Linear Prediction (LP) Analysis

The LPC model is based on an all-pole implementation of the vocal tract response to an excitation of a series of nearly periodic glottal pulses generated by vocal cords (for voiced sounds) or turbulence flow of air passing through a constriction along vocal tract (for unvoiced sounds). In this model (Figure 3.2) the speech output at n^{th} sampling instant is given by [17]

$$s[n] = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \quad , \tag{3.3}$$

where, p is the prediction order, a_k are the predictor coefficients (LPC coefficients), $s(n-k)$ are the past p samples, G is a gain term and $u(n)$ is the appropriate input excitation.

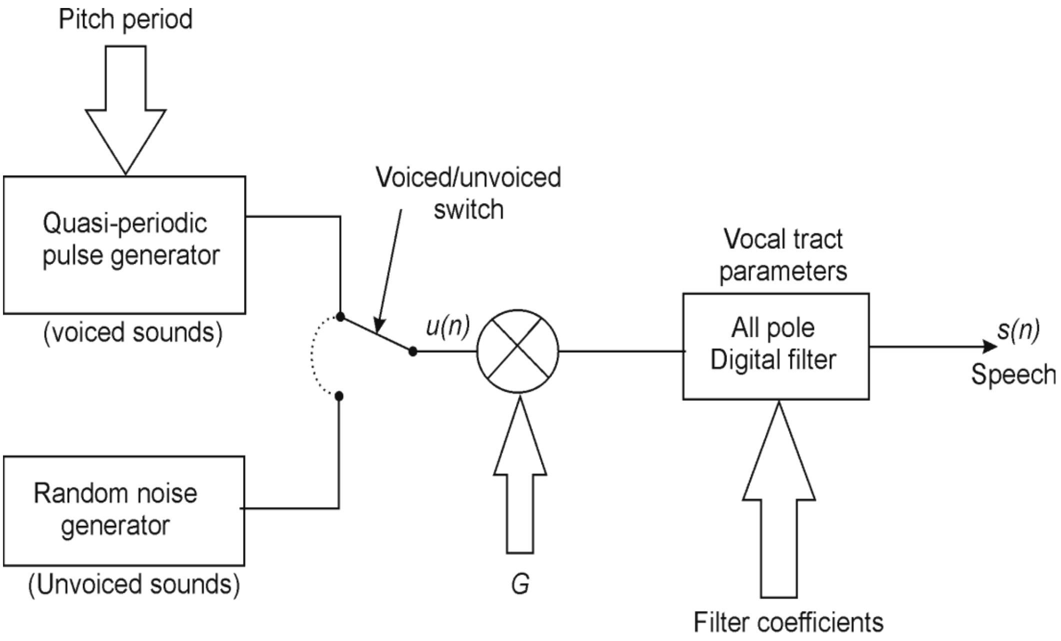


Figure 3.2: LPC model of speech. The excitation source for the voiced and unvoiced sounds is represented as an impulse train generator and a random noise generator respectively. [108]

Applying the z-transform and rearranging the terms of the above equation yields the transfer function of the all-pole filter [17].

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-1}} \tag{3.4}$$

In general, for speech applications, $s(n)$ is estimated using the past p samples. This is given as

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k)\,, \tag{3.5}$$

where $\tilde{s}(n)$ is the approximation of $s(n)$.

The prediction error, $e(n)$ between the actual speech sample, $s(n)$, and the predicted speech sample is then given as.

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)\,, \tag{3.6}$$

The aim of LP analysis is to obtain a set of predictor coefficients, a_k , directly from the short-term speech frame so that the spectral properties of the digital filter of Figure 3.2 match those of the speech frame within the analysis window. The approach to the computation of the LP coefficients is through the minimisation of the mean-squared prediction error, ε, for the frame under investigation. This is given as

$$\varepsilon = \sum_{n=0}^{N-1+p} \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right]^2 , \tag{3.7}$$

where, N is the number of samples in the given speech frame and all the other symbols have the same meaning as in the above equations.

The values of a_k that lead to the minimisation of ε are then obtained by differentiating equation (3.7) with respect to each coefficient and equating the result to zero, i.e.

$$\frac{\partial \varepsilon}{\partial a_k} = 0 , \qquad for\ k = 1,2 \ldots ., p \tag{3.8}$$

This yields the following set of p simultaneous linear equations

$$\sum_{k=1}^{p} a_k \sum_{n=0}^{N-1+p} s(n-k)s(n-i) = \sum_{n=0}^{N-1+p} s(n)s(n-i) \quad for\ i = 1,2 \ldots , p \tag{3.9}$$

It can be seen from the above expression that both the second and the third summation terms are equivalent the short-term autocorrelation values of $s(n)$ at lags$(k-i)$ and (i) respectively. These are given by

$$R(i) = \sum_{n=0}^{N-1+p} s(n)s(n-i) \tag{3.10}$$

$$R(k-i) = \sum_{n=0}^{N-1+p} s(n-k)s(n-i) \tag{3.11}$$

Hence, substituting equations (3.10) and (3.11) into equation (3.9) gives

$$\sum_{k=1}^{p} a_k R(k-i) = R(i) \qquad for\ \ i = 1{,}2 \ldots, p \tag{3.12}$$

The above equation may also be expressed in the matrix form as [17]

$$\begin{bmatrix} R(0) & R(1) & R(2) & \ldots & R(p-1) \\ R(1) & R(0) & R(1) & \ldots & R(p-2) \\ R(2) & R(1) & R(0) & \ldots & R(p-3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \ldots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix} \tag{3.13}$$

It can immediately be seen that this matrix is a Toeplitz matrix since it is symmetrical and has equal diagonal elements. This can be efficiently solved through a widely known procedure known as the Levinson-Durbin (L-D) recursion [17, 79].

In speech analysis, the above method of computing the LPC parameters is known as the autocorrelation method. Other approaches such as the covariance method can also be used to compute these parameters. However, these approaches are not as computationally efficient as the autocorrelation method and do not offer the same inherent numerical stability [17].

The magnitude response of the LPC all-pole filter, gives a smoothed spectral envelope of the short term speech spectrum being analysed. The accuracy of this approximation is related to the number of poles *p* in equation (3.4). As illustrated in Figure 3.3, increasing p leads to a better approximation of the model but at the expense of increased memory requirements and computation. A typical choice for *p* is (F + 4), where F is the sampling frequency of the speech signal in kHz [111].

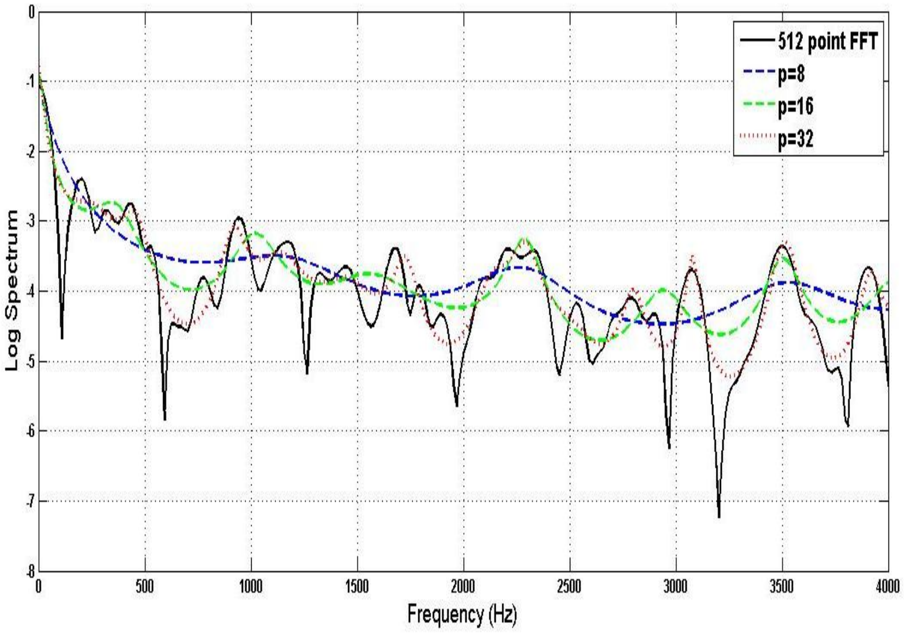


Figure 3.3: Illustration of the effect of increasing the number of LPC coefficients (with p =8, 16 and 32) for a 30 ms speech frame sampled at 8 kHz.

3.2.3 LP-Based Cepstral Analysis

The speech production model in the previous section consists of a vocal tract filter which is driven by an impulse train generator (for voiced sounds) and a random noise generator (for unvoiced speech). Hence, for voiced sounds, the short-term spectrum of speech consists of both a slowly varying spectral envelope and a rapidly varying fine structure. The former corresponds to the vocal tract filter while the second component (for voiced sounds) corresponds to the periodic excitation and its harmonics. The observed output speech sequence is therefore a result of the convolution of these two components in the time domain.

The objective of the cepstral analysis is to separate these two components by transforming their convolutional relationship into a summation. In the frequency domain, the convolution is transformed into a multiplication. This can in turn be transformed into a summation by using the logarithmic operation. A transformation back into a time-like domain, known as the quefrency domain (anagram for frequency), results in the cepstrum (anagram for spectrum). In this domain, the excitation and vocal tract components appear at high and low quefrencies respectively. The vocal tract component which provides a useful speaker representation can then by separated by truncating the series of cepstral coefficients through a process known as liftering (an anagram for filtering). It should be pointed out the cepstral analysis process, which is shown in Figure 3.4, forms part of the family of homomorphic filtering techniques [79]. This, as mentioned in Chapter 2 (Section 2.2.1), is a general term given to any technique which involves a nonlinear mapping to a different domain, followed by a reverse mapping to the original domain.

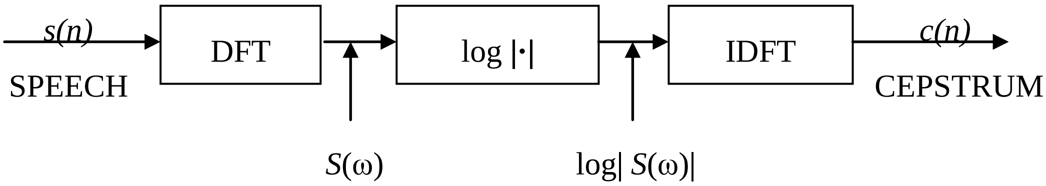


Figure 3.4: Sequences involved in the cepstral analysis process [31]

As shown in the above block diagram, the discrete Fourier transform (DFT) is applied to the incoming speech samples to obtain the short-term spectrum S(ω). The logarithm operation is then applied to the modulus of S(ω), and this is followed by the inverse Fourier transform (IDFT) operation. It should be noted that with the logarithm function being real and even (for the discrete case), the cepstrum can be computed using the Discrete Cosine Transform (DCT) instead of the IDFT [17, 20]. This results in the real cepstrum which is the most popular type for speech processing applications [79]. The cepstral coefficients obtained in this manner are known as fast Fourier transform derived cepstra (FFTC).

The cepstral coefficients can also be obtained directly from the LPC coefficients. In this approach, the Z transform is applied to the speech signal modelled by the LP analysis. This is obtained as [112].

$$\log H(z) = C(z) = \sum_{k=1}^{+\infty} c_k z^{-k} \tag{3.14}$$

The relationship between the parameters c_k and the LP coefficients a_k is found by taking the derivatives on both sides of Equation (3.14) with respect to z^{-1} and equating the terms with equal powers of z^{-1}. The resulting recursive relationship is shown below.

$$c_1 = -a_1 \tag{3.15}$$

$$c_n = -a_n - \sum_{i=1}^{n-1}\left(1-\frac{i}{n}\right)a_k c_{n-k} \quad ,n = 2,3,\ldots.,p \tag{3.16}$$

$$c_n = \sum_{i=1}^{n-1}\left(1-\frac{i}{n}\right)a_k c_{n-k} \quad , \quad n > p \quad , \tag{3.17}$$

where p is the order of the LP analysis and a_k are the LPC coefficients. It should be noted that although the above recursion implies that the sequence of cepstral parameters is of infinite length, in practice, only the first p terms are used. This type of cepstral analysis is known as LPC derived cepstra (LPCC). Figure 3.5 shows the FFT, LPC and LPCC based spectra for a given speech frame. It can immediately be seen that the LPCC spectrum, which is a truncation of the cepstral sequence, results in a smoothing of the spectral envelope.

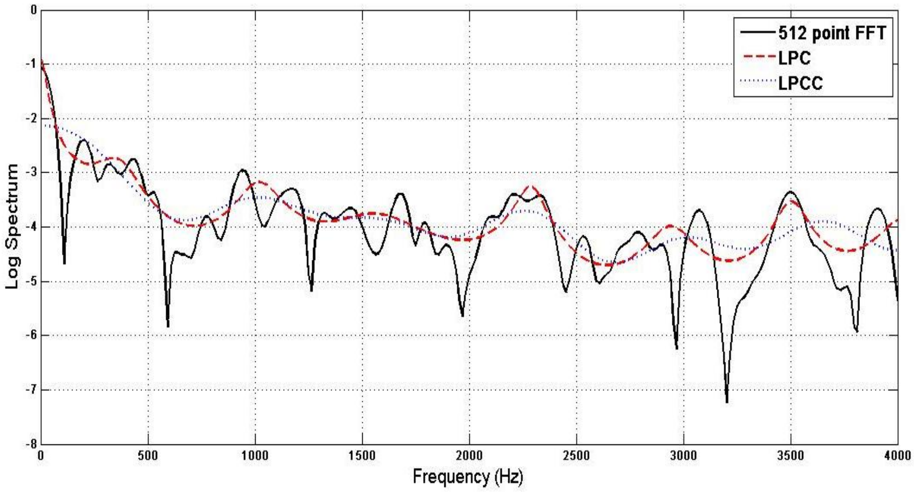


Figure 3.5: Illustration of the cepstral analysis which results in a smoother spectral representation.

3.2.4 Delta Cepstrum

The LPCC coefficients are called static features because they give a representation of the properties of the spectral envelope for a fixed period in time. In order to include information about the slow-moving vocal tract dynamics, transitional cepstral coefficients, also referred as delta coefficients, can be computed. It has been shown that such aspects can be useful in discriminating between different speaker utterances [113]. The delta parameters are approximated by the finite time difference [79]. This is given as

$$\Delta \boldsymbol{c_m}(t) = \boldsymbol{c_m}(t+\delta) - \boldsymbol{c_m}(t-\delta)\,, \tag{3.18}$$

where $\Delta \boldsymbol{c_m}(t)$ is the m^{th} coefficient of the i^{th} cepstral feature vector and δ represents the number of frames which are included in the analysis (backward and forward in time). A typical value for δ is 1 or 2. It is, however, argued in [114] that the delta features obtained using Equation (3.18) are inherently noisy. An alternative method based on fitting each coefficient's trajectory with a first or second order polynomial function over a finite length window has therefore been proposed [17].

$$\frac{\partial \boldsymbol{c_m}(t)}{\partial t} \approx \Delta \boldsymbol{c_m}(t) = \frac{\sum_{k=-K}^{K} k h_k \boldsymbol{c_m}(t+k)}{\sum_{k=-K}^{K} h_k\, k^2} \quad , \tag{3.19}$$

where h_k is a symmetric window of length 2K+1 frames. A value of $K=3$ has been found to be appropriate for an estimate of the first order delta feature [17]. In general, LPCC coefficients are concatenated with the delta cepstrum to obtain a better performing feature vector [113].

3.2.5 Cepstral Mean Normalisation (CMN)

Cepstral Mean Normalisation (CMN), also known as Cepstral Mean Subtraction (CMS) is a feature normalisation approach which aims to reduce the effects of different communication channels on the speech signal [115]. This is carried out by estimating a mean vector for the extracted set of cepstral features and subtracting it from all the feature vectors. This is given as

$$C_{CMN} = c_e(t) - \frac{1}{T}\sum_{t=1}^{T} c_e(t) \qquad \text{for } t = 1, \dots., T \quad , \tag{3.20}$$

where c_e is the extracted cepstral vector, T is the total number of cepstral vectors and t is the frame index.

It has also been shown in [116] that CMN can also help reduce inter-session speaker variation for clean speech and does not necessarily discard important speaker-discriminative information. It should be noted that the reason for performing a subtraction in the cepstral domain is that, in this domain, the channel noise becomes additive.

3.3 Speaker Modelling

In speaker recognition, as explained in Chapter 1 and Chapter 2, the cepstral speech features (LPCC in this case) which are obtained from the registered speaker's training utterance(s) are used to obtain a speaker model. In the test phase, the speaker model is then compared against the test utterance to obtain a similarity score. As discussed in the literature review, the most popular modelling approaches, to date, are based on GMM and SVM. A description of the said modelling strategies is presented in this section.

3.3.1 Gaussian Mixture Model (GMM)

A GMM is a weighted sum of *C* components (or mixtures) Gaussian Probability Density Functions (PDFs). This summation is given as [117, 118]

$$p(\mathbf{o}|\lambda) = \sum_{i=1}^{C} w_i \mathcal{N}(\,\mathbf{o}|\boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_i) \quad , \tag{3.21}$$

where **o** is a *F*-dimensional feature vector, w_i, *i* =1,…..,*C*, are the weights of each of the *C* components. These are constrained by $\sum_i w_i$ =1. $\mathcal{N}(\mathbf{o}|\boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_i)$ are the *F*-variate Gaussian density functions given by

$$\mathcal{N}(\mathbf{o}|\boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{d}{2}}\,|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu}_\mathrm{i})'\boldsymbol{\Sigma}_i{}^{-1}(\mathbf{o}-\boldsymbol{\mu}_\mathrm{i})\right\} , \tag{3.22}$$

for i={ 1,2,…,*C* }. |.| and (.)$'$ indicate the determinant and transpose operation respectively.

The weights, means and covariance parameters are collectively represented by the notation $\lambda = \{w_i, \boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_i\}$. An example of a GMM (with four mixture densities) obtained using two-dimensional LPC-derived cepstral coefficients (LPCC) is shown in Figure 3.6.

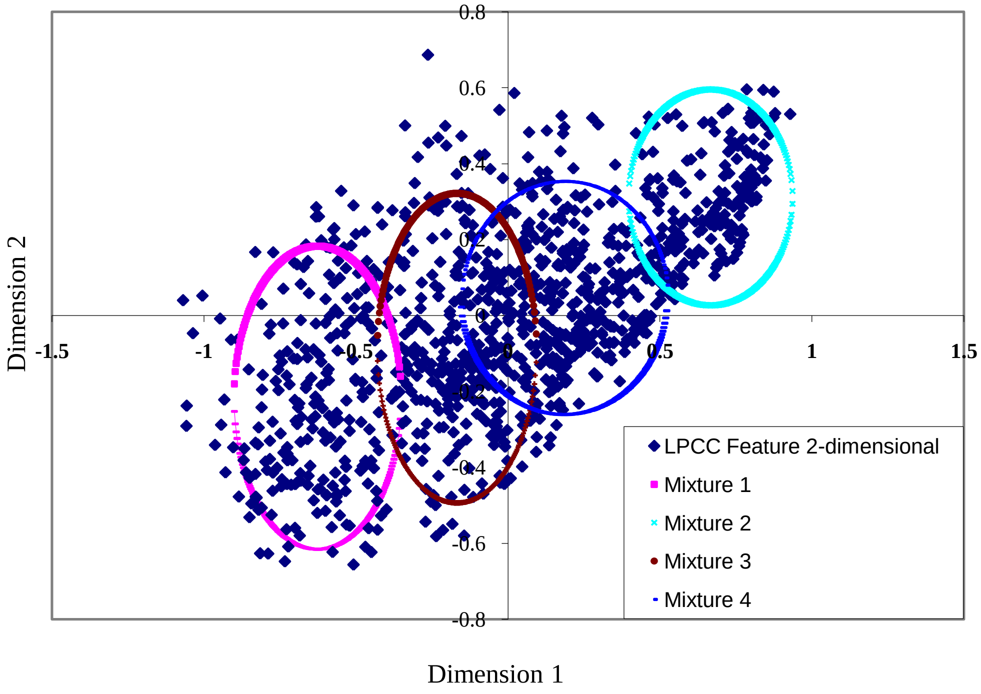


Figure 3.6: Cross-section of 4 Gaussian mixture densities based on two-dimensional LPCCs.

The GMM implemented in this work is based on diagonal nodal covariances. The reasons for this are three-fold. First, this is due to the use of cepstral feature parameters which are known to be highly uncorrelated. In other words, their covariances are negligibly small and the covariance matrix is diagonally dominant. Second, it has been found that the use of one covariance matrix per mixture provides better modelling capabilities, particularly for text-independent speaker recognition scenarios [118]. Finally, the use of diagonal matrices offers advantages in terms of smaller storage requirement, improved computational efficiency and simplicity.

There are two commonly used methods for estimating the parameters of the GMM. The first method is that of computing the model parameters using the iterative Expectation-Maximisation algorithm (EM). This is an unsupervised procedure which is based on the Maximum Likelihood (ML) principle (usually referred to as decoupled-GMM modelling). A detailed description of the Maximum Likelihood principle and the EM approach is given in Appendix A.

The other more popular method for training speaker-dependent GMMs is based on the Maximum a Posteriori (MAP) adaptation of a speaker independent model. This approach, which is based on the Bayesian framework, is usually referred to as adapted-GMM modelling or GMM-UBM. In this case, the main difference from the Maximum Likelihood training lies in the assumption of a prior distribution of the model which is usually derived from speaker independent distributions. This is obtained from the EM approach by using a very large population of speakers, commonly known as world model or universal background model (UBM) [12].

For the purpose of this research study, a modified version [12] of the original MAP approach is adopted. The approach which is hereafter referred to as *m*MAP is based on a single step adaptation process and has been shown to be more effective than the originally proposed approach in [119]. This can be described as follows.

Given a UBM, λ_{UBM} , and a set T training vectors, $\mathbf{O}$= {$\mathbf{o}_1$, $\mathbf{o}_2$,…., $\mathbf{o}_T$} (extracted from a speaker's speech segment), the probabilistic alignment of the training feature vectors in relation to the C mixtures of the UBM is first determined. This is obtained by computing the *a posteriori* probability for acoustic class, i, given the observation, $\boldsymbol{o}_t$

$$p(i|\boldsymbol{o}_t,\lambda) = \frac{w_i p_i(\boldsymbol{o}_t)}{p(\boldsymbol{o}_t|\lambda)} = \frac{w_i p_i(\boldsymbol{o}_t)}{\sum_{k=1}^{C} w_k\, p_k(\boldsymbol{o}_t)} \tag{3.23}$$

Next, the sufficient statistics for the weights, means and variances of each mixture *i* are computed as follows[12]

$$c_i = \sum_{t=i}^{T} p(i|\boldsymbol{o}_t, \lambda_{UBM}) \tag{3.24}$$

$$E_i(\mathbf{O}) = \frac{1}{c_i}\sum_{t=i}^{T} p(i|\boldsymbol{o}_t, \lambda_{UBM})\boldsymbol{o}_t \tag{3.25}$$

$$E_i(\mathbf{O}^2) = \frac{1}{c_i}\sum_{t=i}^{T} p(i|\boldsymbol{o}_t, \lambda_{UBM})\boldsymbol{o}_t{}^2 \tag{3.26}$$

where c_i, $E_i(\mathbf{O})$ and $E_i(\mathbf{O}^2)$ are the count, first and second moment of the training features respectively.

Based on the above statistics, the new estimates for each mixture of the adapted model are then obtained by combining them with the existing UBM parameters. This is given as [12]

$$\hat{w}_i = \left[\alpha_i^w c_i \frac{1}{T} + (1 - \alpha_i^w) w_i\right]\gamma \tag{3.27}$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i^\mu E_i(\mathbf{O}) + \left(1 - \alpha_i^\mu\right)\boldsymbol{\mu}_\mathrm{i} \tag{3.28}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \alpha_i^\sigma E_i(\mathbf{O}^2) + (1 - \alpha_i^\sigma)\,(\boldsymbol{\sigma}_i^2 - \boldsymbol{\mu}_\mathrm{i}^2) - \hat{\boldsymbol{\mu}}_i^2 \ , \tag{3.29}$$

where γ is a scaling factor that ensures all the mixture weights sum to unity. The coefficients $\alpha_i^p, p \in \{w, \mu, \sigma\}$ are the data adaptation coefficients for the ith mixture weight $w_{i,}$ mean $\boldsymbol{\mu}_\mathrm{i}$ and variance $\boldsymbol{\sigma}_i^2$. These control the degree of adaptation of the UBM adaptation and are given as [12]

$$\alpha_i^p = \frac{c_i}{c_i + R^p} \qquad , \tag{3.30}$$

where R^p is known as the relevance factor for the parameter p.

In general, a single adaptation coefficient $\alpha_i = \alpha_i^w = \alpha_i^\mu = \alpha_i^\sigma$ is used. In addition, it is also reported in [12] that the adaptation of only the mean statistics yields better performance for speaker recognition when compared to the full adaptation of all the GMM parameters (i.e. weights, means and covariances) .

Over the last decade, the adapted-GMM has become one of the dominant approaches for modelling a person's voice in speaker recognition applications [12, 26, 73, 120]. This is mainly because this method has been shown to give better performance that the decoupled modelling approach [12].

As mentioned above, the adapted-GMM approach involves the adaptation of a general model (or UBM), using each registered speaker's training material to obtain speaker specific GMMs. The UBM is usually trained using the Expectation-Maximisation (EM) approach on a large amount of development data. Hence, during the adaptation process, each speaker's model parameters are derived by updating the well-trained parameters in the world model according to the available training material. The adaptation process therefore results in a tighter coupling between the speaker's model and the UBM. This is because mixture parameters (representing broad acoustic classes) which are not observed in the training speech of a particular speaker are simply copied from the UBM. Moreover, the tighter coupling provided by the adapted-GMM approach enables a fast-scoring technique to be implemented during the test phase without any significant loss in accuracy [12]. This approach is based on two practical observations. First, it is observed that for each feature vector (from a given test utterance), only a few of the mixtures contribute significantly to the overall likelihood value. Second, it is observed that feature vectors which are close to a particular mixture in the UBM tend to also be close to the corresponding mixture in the speaker model. Thus, the fast-scoring approach combines these two observations in the following manner:

i. For each feature vector, the top Q scoring mixtures in the UBM are determined and the UBM likelihood is computed using only those mixtures.
ii. Then, the feature vector is scored against the corresponding Q mixtures in the speaker model to evaluate the speaker's likelihood.

Based on the above, if it is assumed that a UBM has C mixtures, the fast scoring approach would involve only $C + Q$ computations for each feature vector instead of $2C$ log-likelihood evaluation computations in the normal procedure. This is particularly important when the number of mixtures, C is large i.e. 1024 or 2048 [12].

a) *Classification Stage*

For speaker recognition, once the speaker GMMs are obtained, the next step is to make use of the model for authenticating speakers based on their test utterances. In the speaker verification context, the task is one of evaluating the probability of a hypothesised (claimed) speaker model, λ for a given observation, $\mathbf{O}$. This is given as $p(\lambda|\mathbf{O})$ and can be rewritten as follows by using the Baye's Theorem

$$p(\lambda|\mathbf{O}) = \frac{p(\mathbf{O}|\lambda)p(\lambda)}{p(\mathbf{O})} \quad , \tag{3.31}$$

where $p(\lambda)$ is the a priori probability of the target speaker model. This probability is considered equal for all models and can be neglected. $p(\mathbf{O})$ is the unconditional probability of the observation, $\mathbf{O}$, being produced by any speaker. This can be assumed to be a constant. Equation 3.31 can be simplified and transformed into the log domain to give the log-likelihood function

$$L(\mathbf{O}) = \log\left(p(\mathbf{O}|\lambda)\right. \tag{3.32}$$

The speaker verification decision is then made based on whether $L(\mathbf{O})$ is above or below a pre-defined threshold.

For the speaker identification scenario, the same principle is again used but this time to find the registered speaker model which produces the highest log likelihood against the given test segment. This is given as

$$S = \underset{1 \le n \le N}{\arg\max}\{\log\left(p(\mathbf{O}|\lambda_n)\right.\} \quad , \tag{3.33}$$

where N is the total number of registered speakers and S is the index of the most likely candidate in the set.

3.3.2 Support Vector Machine (SVM)

A SVM is a two-class discrimination technique which involves finding a hyperplane (boundary) for effective separation of the two classes considered. Although SVMs can perform binary separation in the input space for linearly separable cases, they usually operate in a higher dimensional space which is non-linearly related to the input space. In the classification stage, the SVM discriminant function [59] is used to evaluate the given test data vector in relation to the separating hyperplane. The SVM discriminant function can be expressed as

$$\begin{aligned} f(\mathbf{x}) &= \textstyle\sum_{i=1}^{N_{SV}} y_i \alpha_i \Phi(\mathbf{x}_\mathrm{i}).\Phi(\mathbf{x}) + \mathrm{b} \\ &= \textstyle\sum_{i=1}^{N_{SV}} y_i \alpha_i K(\mathbf{x}_\mathrm{i}, \mathbf{x}) + \mathrm{b} \quad , \end{aligned} \tag{3.34}$$

where $\mathbf{x}$ is the test data vector, and $\Phi(\cdot)$ is a mapping function that transforms the data vector from its input space to a higher dimensional space. $K(\mathbf{x}_\mathbf{i}, \mathbf{x})$ is a kernel function which defines the inner product $\Phi(\mathbf{x}_\mathbf{i}).\Phi(\mathbf{x})$ and therefore eliminates the need for explicitly evaluating $\Phi(\cdot)$. $\mathbf{x}_\mathrm{i}$ are the only training vectors which influence the definition of the said hyperplane. These are commonly known as support vectors, and are obtained from the training process [121]. y_i is the corresponding support vector's class label ($y_i \in$ {-1,1}) while N_{SV} is the number of support vectors. The values of α_i and the constant b are also obtained during the training stage. More details of the fundamental concept involved in the SVM approach are given in Appendix B.

For speaker verification based on SVM, it is crucial to be able to compare the given utterances regardless of their duration [29, 60]. To date, one popular approach which is based on SVM only has been proposed to represent each utterance using a fixed dimensional vector [60]. This is known as the Generalised Linear Discriminant Sequence (GLDS) kernel. In this approach, given a set of T feature vectors, $\mathbf{X}$= {$\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_T$}, where $\mathbf{x}_t$ is an F-dimensional feature vector, each feature vector is explicitly mapped into a higher dimensional feature space using a polynomial expansion, $\Phi_{\mathrm{GLDS}}(\mathbf{x}_t)$. For instance, a second order polynomial expansion of a three-dimensional vector $\boldsymbol{x} = (x_1, x_2, x_3)$ is given by $\Phi_{\mathrm{GLDS}}(x) = (1, x_1, x_2, x_3, x_1x_2, x_2x_3, x_1x_3, x_1^2, x_2^2, x_3^2)$.

For each speaker, the mean of the expanded features is computed, resulting in a fixed dimensional feature vector which is independent of the duration. This is given by

$$\mathbf{X}_{GLDS} = \frac{1}{T}\sum_{t=1}^{T}\Phi_{\mathrm{GLDS}}(\mathbf{x}_t) \ , \tag{3.35}$$

where the dimension of $\mathbf{X}_{GLDS}$ is dependent on the dimension of the feature vector and the order of the polynomial expansion used. However, while this method was amongst one of the first approaches to use SVM and has been shown to give good speaker recognition performance, in some cases, the averaging process may lead to loss of useful speaker information [29, 60].

a) GMM supervector approach

More recently, it has been shown in [14, 28, 29, 71] that, using the GMM supervector approach with SVM can help overcome the above limitation and yield the current state-of-the-art speaker verification performance. The idea behind the GMM supervector method is to allow each utterance, independent of its duration, to be represented by a concatenation of the means obtained from an adapted GMM model [122]. The GMM supervector obtained in this way can be expressed as

$$\Phi(\mathbf{X}) = \begin{bmatrix} \mu_{\mathbf{X}}^{1} \\ \cdots \\ \mu_{\mathbf{X}}^{F} \\ \cdots \\ \mu_{\mathbf{X}}^{CF} \end{bmatrix}, \tag{3.36}$$

where F is the dimension of the feature vectors extracted from the given utterance $\mathbf{X}$, $\mu_{\mathbf{X}}^{i}$ are the means of the GMM obtained through the adaptation of the UBM, and C is the number of mixtures in the UBM.

Some of the most commonly used kernels in the literature which are based on the GMM supervector approach are the Background data Scaling Linear (BSL) kernel [29], GMM supervector linear kernel [28, 71], the non-linear GMM-supervector-kernel [14] and the Maximum Likelihood Linear Regression (MLLR) kernel [123]. A brief overview of the said approaches is given below.

The Background data scaling linear kernel can be considered as one of the simplest approaches based on GMM supervector of means. In this approach, the input supervectors (in equation 3.36) are normalised such that they have unit variance in each dimension based on the statistics of a large number of background supervectors. The aim of the variance normalisation is to ensure that each dimension of the supervector contributes equally to the SVM training or testing process. This is given by [29]

$$K_{BSL}(\mathbf{X}, \mathbf{Y}) = \langle \Phi(\mathbf{X}).\boldsymbol{B}^{-1}\Phi(\mathbf{Y}) \rangle \qquad , \tag{3.37}$$

where $\Phi(\mathbf{X})$ and $\Phi(\mathbf{Y})$ represent the GMM supervector of mMAP adapted means from utterances $\mathbf{X}$ and $\mathbf{Y}$ respectively. $\boldsymbol{B}$ is the diagonal covariance matrix of the background supervectors.

The GMM supervector (GSV) kernel which has been proposed in [28, 71], is derived by bounding the Kullback-Leibler (KL) measure between two GMMs (in terms of their supervector of means) [28, 71]. This distance is given by

$$d\big(\Phi(\mathbf{X}),\Phi(\mathbf{Y})\big) = \sum_{i=1}^{C} w_i(\,\boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{X}} - \boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{Y}})\,\boldsymbol{\Sigma}_i{}^{-1}\big(\boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{X}} - \boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{Y}}\big)^t\,, \qquad (3.38)$$

where $\boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{X}}$ and $\boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{Y}}$ are the speaker-dependent adapted mean vector using mMAP adaptation of the i^{th}-mixture for utterances $\mathbf{X}$ and $\mathbf{Y}$ respectively. w_i and $\boldsymbol{\Sigma}_i$ are the weights and covariance for the corresponding mixture.

Based on the distance in equation 3.38, the kernel function can be formulated in terms of an inner product as follows [14, 28, 71]

$$K_{GSV}(\mathbf{X},\mathbf{Y}) = \sum_{\mathrm{i}=1}^{\mathrm{C}} w_i \boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{X}}\,\boldsymbol{\Sigma}_i{}^{-1}\big(\boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{Y}}\big)^t$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{C}} \langle(\sqrt{w}_i\boldsymbol{\Sigma}_i{}^{-\left(\frac{1}{2}\right)}\ \boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{X}}).(\sqrt{w}_i\boldsymbol{\Sigma}_i{}^{-\left(\frac{1}{2}\right)}\ \boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{Y}})\rangle \qquad (3.39)$$

where $\boldsymbol{\mu}_{\mathrm{i}}$ are the speaker-dependent adapted mean vector using mMAP adaptation of the i^{th}-mixture respectively. In other words, all the adapted mean vectors in equation (3.36) have to first be normalised by $\sqrt{w}_i\boldsymbol{\Sigma}_i{}^{-\left(\frac{1}{2}\right)}$ before concatenating them to form the supervector of means. As with the BSL kernel, this process can also be considered as a form of variance normalisation [110].

The non-linear GMM supervector kernel is also based on the KL divergence between GMMs. In this case, however, the kernel is obtained by taking the exponent of the negative of the distance function in (3.38) such that

$$K_{non-linear}(\mathbf{X},\mathbf{Y}) = \mathrm{e}^{-\,d(\Phi(\mathbf{X}),\Phi(\mathbf{Y}))} \qquad (3.40)$$

As such, the non-linear kernel represents the normalised exponential of the GSV kernel. Moreover, unlike the GMM supervector linear kernel, $K_{non-linear}(\mathbf{X},\mathbf{Y})$ does not imply an explicit expansion of the input vectors into the feature space . In this case, the resulting kernel closely resembles that of the Gaussian kernel [14].

Alternatively, the Maximum Likelihood Linear Regression (MLLR) kernel approach is based on adapting the means of the UBM using the MLLR approach instead of the *m*MAP adaptation. This involves computing an affine transform (i.e. a linear transformation followed by a translation) of the UBM means as follows

$$\boldsymbol{\mu}_{\mathrm{MLLR}}^{i} = \boldsymbol{A}\boldsymbol{\mu}_{\mathrm{UBM}}^{i} + \ b \ , \tag{3.41}$$

where $\boldsymbol{\mu}_{\mathrm{MLLR}}^{i}$ and $\boldsymbol{\mu}_{\mathrm{UBM}}^{i}$ are the speaker-dependent MLLR adapted mean vector and the UBM mean vector of i^{th}- mixture respectively. The parameters $\boldsymbol{A}$ and b define the affine transform and are estimated by maximising the likelihood of the training data with a modified EM algorithm [123]. Once the MLLR adapted mean vectors are obtained, they are concatenated as in (3.36) to obtain the supervectors.

b) SVM optimisation and classification

The SVM optimisation (training) process then involves finding the support vectors, $\mathbf{x}_{\mathrm{i}}$, the Lagrange multipliers, α_i and the value of the offset b in Equation (3.34). For this purpose, each client supervector is assigned a label of +1 while a set of supervectors from a background dataset representing a large number of impostors are given a label of -1. During testing phase, the exact procedure used in extracting supervectors as in the training stage is used (in this case, no labels are given to the supervectors). An inner product between the test supervector and the SVM model is then computed to obtain a classification score which represents the distance of the test vector to the SVM hyperplane.

The experimental implementation of this approach together with other complementary techniques which have been shown to give the current state-of-the-art speaker recognition performance are described in the next chapter.

3.4 Tackling mismatch noise conditions

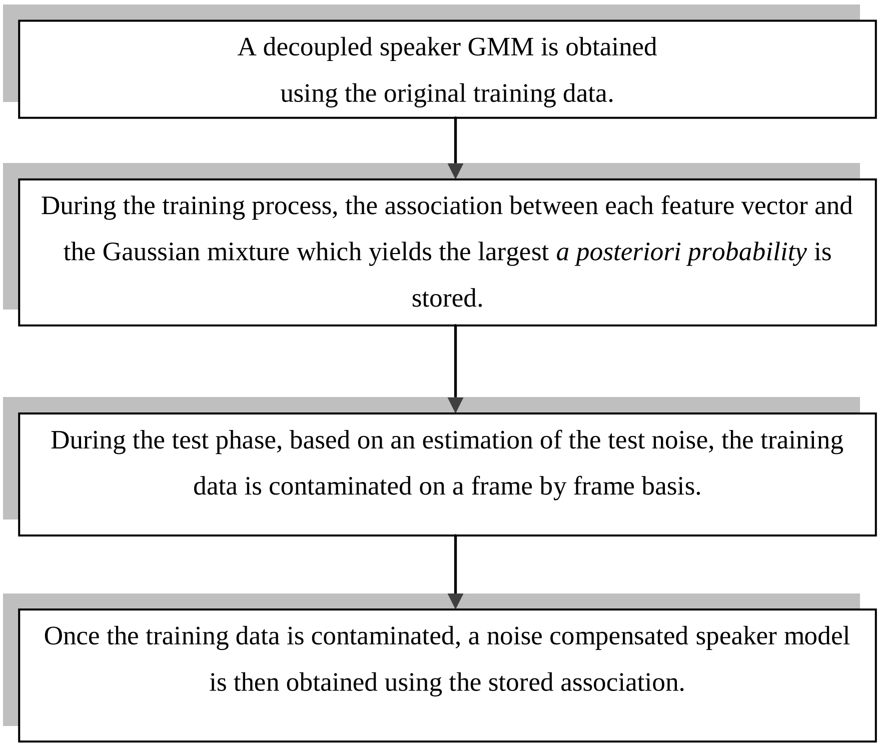
As discussed in the literature review, variations in speech remain the major impeding factor for speaker recognition systems. These variations occur due to various causes such as environmental noise, channel effects, or uncharacteristic sounds by the speakers. The net result is a mismatch between the corresponding test and reference material for the same speaker, which in turn reduces the accuracy speaker recognition applications. In this thesis, as mentioned in Chapter 1, the experimental investigations are focussed on the problem of speaker recognition when the speech samples are distorted by environmental noise. The literature review in the previous chapter has shown that a number of techniques have been proposed to tackle this problem. Such techniques can be classified into four main categories, depending on the level at which they operate, namely: speech-level, feature-level, model-level, and score-level. The work presented in this study is focused on approaches in the latter two categories. These include Parallel Model Combination (PMC) and score normalisation. A detailed description of the said approaches is given in the next sub-sections.

3.4.1 Parallel Model Combination

The PMC technique which has originally been proposed in [124] for the speech recognition task is based on the use of HMM with single Gaussian output probability. The objective of the approach is to use an estimate of the test noise during the recognition stage for building noise compensated models from the reference material (clean speech models). To achieve this, a model of the background noise is generated using the available noise samples. The clean speech models and noise model are then combined in the log-spectral domain to obtain the best possible estimate of the corrupted-speech models. The reason for operating in this domain is because it allows the effects of the additive noise on the speech feature vectors to be approximated when the original training utterances are not available. To date, various approximations have been proposed for estimating these effects and computing the corrupted model parameters [32-35].

On the other hand, when the original reference material is available, it has been shown that the simple yet effective and efficient data-driven approach in [92], can be very useful in the context of speaker verification. The main advantage of this technique is that the approximations which are usually required for combining the models (noise and reference model) are eliminated by using the original training data. This is particularly important in order to accurately model the effects of the additive noise on the speech parameters and therefore enables a robust computation of the noise compensated model parameters. In addition, the use of the original reference material allows the temporal context of each speech frame to be retained. As a result, delta parameters which have been shown to improve the performance of speaker recognition can be accurately and easily computed.

Figure 3.7 provides an overview of the various steps which are carried out in order to obtain the noise compensated models. This technique forms the basis of the work carried out in Chapter 5 for dealing with mismatched noise conditions in speaker recognition.



A decoupled speaker GMM is obtained using the original training data.

During the training process, the association between each feature vector and the Gaussian mixture which yields the largest *a posteriori probability* is stored.

During the test phase, based on an estimation of the test noise, the training data is contaminated on a frame by frame basis.

Once the training data is contaminated, a noise compensated speaker model is then obtained using the stored association.

Figure 3.7: Illustration of the different steps involved in the data-driven PMC approach for speaker recognition operating under noise contaminated conditions.

3.4.2 Score Normalisation Approaches

As discussed in the literature review, a widely used approach for tackling the problem of mismatched noise conditions in speaker verification is that of score normalisation [73, 96, 99]. The approach is based on obtaining a normalisation factor(s) using the match score(s) computed for the test utterance against a set of background (competing) models or a single universal background model [12, 73]. The aim of score normalisation approaches is to alleviate the impact of noise mismatch by reducing the overlapping of the score distributions between client and impostors. These techniques can be classified into two main categories.

a) Bayesian Solution

The first category is based on the Bayes' theorem which is given by

$$p(\lambda|\mathbf{O}) = \frac{p(\mathbf{O}|\lambda)p(\lambda)}{p(\mathbf{O})} \quad , \tag{3.42}$$

As noted in Section 3.3.1, $p(\mathbf{O})$, the unconditional probability of the observation set $\mathbf{O}$ being produced by any speaker is a constant. This term can therefore be discarded. However, in order realise the full benefit of the Bayesian solution, $p(\mathbf{O})$ should be approximated and included. This probability can also be interpreted as the conditional probability of the observation set $\mathbf{O}$, originating from a large speaker independent model of impostors, i.e. $(p(\mathbf{O}|\lambda_I)$. The log-likelihood ratio is then given as [73]:

$$L(\mathbf{O}) = \log(p(\mathbf{O}|\lambda_{T/ML}) - \log(p(\mathbf{O}|\lambda_I) \quad , \tag{3.43}$$

where $\lambda_{T/ML}$ is the model representing the target speaker model or the model which yields the highest maximum likelihood for the speaker verification and open-set speaker identification scenarios respectively. λ_{I} represents an impostor model (which does not exist in practice).

The main approaches for obtaining an appropriate approximation of this model are Universal Background Model Normalisation, Cohort Normalisation (CN) or Unconstrained Cohort Normalisation (UCN) [73].

(i) Universal Background Model Normalisation

This technique approximates the impostor model, λ_{I}, with a model generated using utterances from a large population of speakers, λ_{UBM} . This is known as a Universal Background Model (UBM) or world model and is given as [73]

$$L(\mathbf{O}) = \log(p(\mathbf{O}|\lambda_{T/ML}) - \log(p(\mathbf{O}|\lambda_{UBM}) \quad , \tag{3.44}$$

(ii) Cohort Normalisation (CN)

In this approach, each registered speaker model is associated with the most competitive speaker model cohort. The competitiveness of any two speaker models is in relation to their closeness in the speaker space. The cohort selection is done *a priori* (offline) and the log-likelihood ratio for a cohort of *K* speakers is computed as [73]

$$L(\mathbf{O}) = \log\left(p(\mathbf{O}|\lambda_{T/ML}\right) - \log\left(\frac{1}{K}\sum_{k=1}^{K}\log\left(p(\mathbf{O}|\lambda_{T/ML,k}\right)\right) \quad , \qquad (3.45)$$

where $\lambda_{c/ML,k}$ for {*k*=1,2….,*K*} are the cohort speaker models associated with $\lambda_{C/ML}$.

(iii) Unconstrained Cohort Normalisation (UCN)

The main difference between UCN and the two previous methods is that this approach does not require any additional process prior to the test phase. In other words, in UCN, the selection of the most competitive background speaker models is solely based on their closeness to the test segment. Here, the log-likelihood ratio is given by [73]

$$L(\mathbf{O}) = \log\left(p(\mathbf{O}|\lambda_{T/ML}\right) - \log\left(\frac{1}{K}\sum_{k=1}^{K}\log\left(p(\mathbf{O}|\lambda_{k}\right)\right) \quad , \qquad (3.46)$$

where λ_k for {*k*=1,2….,*K*} are the cohort speaker models which yield the next highest *K* likelihood scores to $\log\left(p(\mathbf{O}|\lambda_{T/ML}\right)$.

b) Standardisation of score distributions

The second category of score normalisation is based on the standardisation of the target or impostor score distributions. In practice, however, the scaling is usually performed on the impostor score distributions. This is because the estimation of reliable normalisation parameters (i.e. mean and variance) requires large amounts of data and, currently, the available databases only contain enough data from impostors. Two of the most commonly used normalisation approaches in this group are Test normalisation (T-Norm) and Zero normalisation (Z-norm) [73].

(i) Zero normalisation (Z-norm)

The aim of Z-norm is to compensate for mismatches in speaker models that are generated under different training conditions [31]. Such mismatches, which are often referred to as model specific biases, can be represented by the mean and standard deviation of impostor scores generated against the associated model. Thus, each registered speaker model is tested against a set of example impostor utterances (during a development stage) and the log-likelihood scores are used to obtain the normalisation parameters. During the test phase, Z-norm is applied as follows

$$L(\mathbf{O}) = \frac{\log\left(p\left(\lambda_{T/ML}\middle|\mathbf{O}\right) - \mu_z(\lambda_{T/ML})\right.}{\sigma_{\mathrm{z}}\left(\lambda_{T/ML}\right)} \quad , \tag{3.47}$$

where $\mu_z(.)$ and $\sigma_{\mathrm{z}}(.)$ are the mean and standard deviation of the impostor score distribution associated to the speaker model $\lambda_{T/ML}$. It can be noticed that the above equation involves a posteriori probability. This implies that Z-norm has to be used in conjunction with other score normalisation methods. More details of the implementation of Z-norm can be obtained in [31].

(ii) Test normalisation (T-norm)

In the T-norm approach, unlike the Z-Norm method, the computation of the mean and variance parameters is carried out dynamically during the test phase by using a cohort of impostor models. This, therefore, eliminates the risk of an acoustic mismatch between the test utterance and the normalisation parameters, which can arise with the Z-norm method. The computation of T-norm is given as

$$L(\mathbf{O}) = \frac{\log\left(p\left(\mathbf{O}\middle|\lambda_{T/ML}\right) - \mu_I(\mathbf{O})\right.}{\sigma_I(\mathbf{O})} \quad , \tag{3.48}$$

where $\mu_I(\mathbf{O})$ and $\sigma_I(\mathbf{O})$ are the mean and standard deviation obtained from the log-likelihood scores for a set of impostor speaker models during the test stage.

To date, UCN and T-norm have been shown to be the most effective for the speaker recognition task [95, 99, 125]. In general, it is also demonstrated that the performance of these two normalisation techniques is very similar. More recently, a new variation of T-Norm, known as Adaptive T-Norm (AT-Norm) has been proposed in the literature [126]. The approach involves assigning a specific (and smaller) set of background speaker models to each target speaker instead of using a general background cohort of speakers for all speakers. In this approach, the size of these speaker-specific sets is chosen to be a fraction of the entire background speaker population, but sufficiently large for computing the T-Norm parameters reliably. As expected, this approach has been shown to be more efficient that the conventional T-Norm method [126]. However, its effectiveness is usually dependent on the availability of adequately large and varied cohorts of background speaker models.

For this reason, it is decided that only the T-norm approach should be adopted for all score normalisation purposes related to this study. This approach also forms the basis of the work carried out in Chapter 5 for tackling the effects of noisy operating conditions on speaker recognition.

3.5 Chapter Summary

This chapter has presented details of the techniques in speaker verification which have been adopted for the purpose of this study. The descriptions have included the operations involved in extracting LPCC features from the speech signal and techniques for modelling speakers using GMM and SVM.

In the extraction of LPCC features, the importance of the various pre-processing stages is discussed. Pre-emphasis is shown to be useful for compensating the spectral roll-off in speech and improving the numerical stability in LP analysis. Subsequently, the operation of windowing is found to be useful in improving the spectral characteristics of the short-term speech signal. Finally, it is shown that a voice activity detection module is also crucial in ensuring that the speaker verification process focuses on speaker-dependent characteristics and not on silence segments.

It is then shown that the LPC model results in a smoothed spectral envelope of the short term speech spectrum being analysed. The theory of the cepstral analysis technique which aims to separate the convolved components of the vocal tract and the excitation from the speech waveform is then discussed. A section on methods for capturing the transitional spectra (delta ceptrum) of the speech signal which can be useful in discriminating between different speaker utterances is also included. Following this, the importance of using Cepstral Mean Normalisation (CMN) to reduce the effects of different communication channels on the speech signal is discussed.

The GMM is one of the most popular approaches to modelling speech cepstra. The speaker model can be obtained using the Expectation Maximisation (EM) approach to obtain decoupled-GMMs or MAP principles to obtain adapted-GMMs from a Universal Background Model.

The SVM is another popular approach for modelling speakers' utterances in speaker verification. The approach involves discriminating between two classes by finding a hyperplane for effective separation of the two classes considered. This makes it inherently suitable for the speaker verification task. However, one limitation of SVM is that the dimension of the input data has to be fixed regardless of the duration of the utterances. To tackle this problem, SVM based on GMM supervectors approach has been proposed and shown to give state-of-the-art speaker verification performance. A description of commonly used kernels which have been reported to give good performance using the said approach is then given.

This chapter has also presented details of two effective techniques for minimising mismatch noise conditions, namely Parallel Model Combination (PMC) and score normalisation based approaches. The objective of PMC is to use an estimation of the test noise during the recognition stage for building noise compensated models from the reference material (clean speech models). On the other hand, score normalisation approaches are based on obtaining a normalisation factor using the match score(s) computed for the test utterance against a set of background (competing) models or a single universal background model. The normalisation factor is then utilised to alleviate the impact of noise mismatch by reducing the overlapping of the score distributions between client and impostors.

CHAPTER 4

INVESTIGATIONS INTO STATE OF THE ART SPEAKER VERIFICATION

Chapter Overview

In this chapter, the most popular techniques for speaker verification (i.e. GMM-UBM and GMM-SVM) are investigated for their effectiveness. The chapter starts with a description of these techniques and details complementary methods which help to enhance the speaker verification performance. This is given in Section 4.1. Details of the experimental setup used for comparing the relative effectiveness of the said speaker verification approaches are given in Section 4.2. A description of the experiments, investigating the relative effectiveness of the GMM-UBM and GMM-SVM approaches, are then presented in Section 4.3. This part of the study includes an analysis of the performance of the considered speaker verification methods under both matched and mismatched data conditions.

4.1 Classification Methods

As discussed in Chapter 3, the most popular techniques for the speaker verification task are based on Gaussian Mixture Models (GMM) or Support Vector Machines (SVM) methodologies. Over recent years, the effectiveness of the above approaches has been considerably enhanced by the introduction of complementary techniques for dealing with variation in operating conditions [14, 15, 29, 127-130]. These include such methods as Nuisance Attribute Projection (NAP) [71, 131, 132] and Model-normalisation (M-Norm) [14, 18, 133]. A short description of the above mentioned methods is provided in the following sub-sections.

4.1.1 GMM-UBM

The GMM-UBM approach for speaker verification can be considered as a four stage process. First, a gender-independent Universal Background Model (UBM) is generated. This is a Gaussian Mixture Model (GMM) built based on the Expectation-Maximisation (EM) algorithm and using utterances from a very large population of speakers [12, 118]. The speaker specific models are then obtained through the adaptation of the means from the UBM using the speakers' training speech and the *m*MAP approach [12, 31]. In the test phase, a fast scoring procedure is used in order to reduce the amount of computation [12].This involves determining the top few (e.g. 5) scoring mixtures in the UBM for each feature vector and then computing the likelihood of the target speaker model using only the scores for its corresponding mixtures. The scoring process is then repeated for all the feature vectors in the test utterance to obtain the average log likelihood score for each of the UBMs and the target speaker model. Finally, UBM-based normalisation is performed by subtracting the log likelihood score of the UBM from that of the target speaker model. This is firstly to minimise the effects of unseen data, and secondly to deal with the data quality mismatch [12, 73].

4.1.2 GMM-SVM

The GMM-SVM approach for speaker verification can also be considered as a process based on a set of consecutive stages. The first step is identical to the GMM-UBM approach where a gender-independent Universal Background Model (UBM) is generated using the EM algorithm. Training utterances from the clients and a large number of impostors are then used to obtain adapted speaker models based on the *m*MAP adaptation of the means from the UBM. Once these are obtained, client and impostor supervectors are extracted by concatenating the means obtained from their corresponding adapted GMM models. This is then followed by SVM training in order to obtain the client model (in terms of the support vectors $\mathbf{x_i}$, α_i values and the constant b). For this purpose, each client training supervector is assigned a label of +1 while the impostor supervectors are assigned a label of -1. During the classification stage, based on the test utterance, the procedure used for extracting the test supervector is exactly the same as that in the training stage (in the testing phase, no labels are given to the supervector). Finally, the classification score is obtained by evaluating the distance of the test supervector in relation to the SVM model. This is given by [121]:

$$f(\mathbf{x}) = \sum_{i=1}^{N_{SV}} y_i \alpha_i \Phi(\mathbf{x}_\mathrm{i}).\Phi(\mathbf{x}) + \mathrm{b} \tag{4.1}$$

$$= \sum_{i=1}^{N_{SV}} y_i \alpha_i K(\mathbf{x}_\mathrm{i}, \mathbf{x}) + \mathrm{b}\ , \tag{4.2}$$

where $\mathbf{x}$ is the test data vector, and $\Phi(\cdot)$ is a mapping function that transforms the data vector from its input space to a higher dimensional space. $K(\mathbf{x_i}, \mathbf{x})$ is a kernel function which defines the inner product $\Phi(\mathbf{x_i}).\Phi(\mathbf{x})$ and therefore eliminates the need for explicitly evaluating $\Phi(\cdot)$. $\mathbf{x}_\mathrm{i}$ are the only training vectors which influence the definition of the said hyperplane. These are commonly known as support vectors, and are obtained from the training process [121]. y_i is the corresponding support vector's class label ($y_i \in$ {-1,1}) while N_{SV} is the number of support vectors. The values of α_i and the constant b are also obtained during the training stage.

a) Nuisance Attribute Projection (NAP)

The main objective of NAP [71, 131, 132], which is used in the SVM framework, is to project points from the feature (supervector) space to another subspace which is more robust to channel and session degrading factors. This is achieved by finding a projection matrix **P**, based on a background corpus which consists of many different speaker recordings (sessions) without explicit labelling. This is given as

$$\widehat{\mathbf{m}}_{\mathbf{X}} = \mathbf{P}\,\mathbf{m}_{\mathbf{X}} = (\mathbf{I} - \boldsymbol{S}\boldsymbol{S}^{t})\,\mathbf{m}_{\mathbf{X}} \;, \tag{4.3}$$

where $\mathbf{m}_{\mathbf{X}}$ is the input supervector obtained during the training or testing stage, **I** is the identity matrix and $\widehat{\mathbf{m}}_{\mathbf{X}}$ is the NAP supervector. **S** is a rectangular matrix whose columns L , represent orthonormal eigenvectors that identify the subspace where the variations between different sessions are the largest and t denotes the transpose operation. In this work, a value of L=40 is chosen. This value which represents the 40 eigenvectors with the highest eigenvalues has been reported to give good speaker recognition results [71, 130]. An efficient and relatively easy approach to finding the matrix **S** is described in [130].

b) Model-normalisation (M-norm)

The Model normalisation (M-norm) technique [18, 133] has been shown to complement the performance of the GMM-SVM approach [14]. The objective of M-norm, as shown in Figure 4.1, is to normalise the input supervectors at the model level, such that the distance between the M-normalised supervectors and the supervector extracted from the UBM is a constant (e.g. 1). This normalisation process can be interpreted as the elimination of the variations in distances relative to the UBM, which exist between different speaker models. The approach is motivated from the hypothesis that these differences, which can affect the overall effectiveness of the speaker verification system, arise due to the speaker non-discriminative information present in the utterance(s) used to build the model. The removal of this degrading factor, therefore, allows the verification process to focus primarily on the speaker discriminative characteristics represented by the direction (with respect to the UBM) which the model takes in the model space [18].

This normalisation is given as

$$\tilde{\mathbf{m}}_{\mathbf{X}} = \frac{1}{\mathrm{D_e}(\mathbf{X},\mathrm{UBM})}\,\mathbf{m}_{\mathbf{X}} + \left(1 - \frac{1}{\mathrm{D_e}(\mathbf{X},\mathrm{UBM})}\right)\mathbf{m}_{\mathrm{UBM}}\ , \qquad (4.4)$$

where $\mathrm{D_e}(\mathbf{X},\mathrm{UBM})$ is the Euclidean distance between the GMM representing utterance $\mathbf{X}$ and the UBM, $\mathbf{m}_{\mathrm{UBM}}$ is a supervector of means extracted from the UBM and $\tilde{\mathbf{m}}_{\mathbf{X}}$ is the M-normalised supervector.

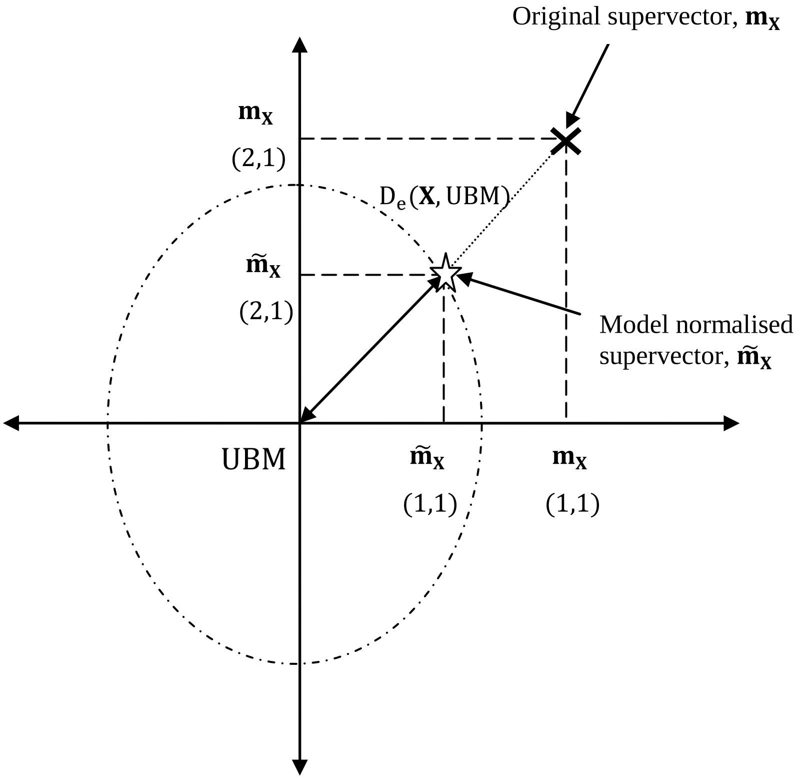

Original supervector, $\mathbf{m}_{\mathbf{X}}$
$\mathbf{m}_{\mathbf{X}}$
(2,1)
$\mathrm{D_e}(\mathbf{X},\mathrm{UBM})$
$\tilde{\mathbf{m}}_{\mathbf{X}}$
(2,1)
Model normalised supervector, $\tilde{\mathbf{m}}_{\mathbf{X}}$
UBM
$\tilde{\mathbf{m}}_{\mathbf{X}}$
(1,1)
$\mathbf{m}_{\mathbf{X}}$
(1,1)

Figure 4.1: Illustration of the Model normalisation process in a two-dimensional space [18].

4.2 Experimental Investigations

In this section, a number of experiments are conducted to investigate the effectiveness of the above mentioned speaker verification approaches. The aim of the first set of investigations is to implement benchmark methods which have been reported to give the current state-of-the-art speaker verification performance. The next set of experiments is then carried out to evaluate the performance and characteristics of such approaches under different experimental conditions.

4.2.1 Speech Data

The experiments in this study are conducted using the speech data obtained in telephonic audio conditions and clean audio conditions. For the telephonic conditions, a subset of the NIST-SRE 2003 [103] database is used. This involves 142 registered speakers, a UBM trained by pooling two gender-dependent UBM (each trained using about 4 hours of speech from speakers other than the ones used for client training, true trials or out-of-set impostor trials), 1293 true trials and 1408 impostor trials [31].

For clean audio conditions, speech data from the TIMIT database [104] is considered. This set includes 100 registered speakers and 80 unknown speakers, each with 10 utterances. The individual utterances are about 3 seconds long. The training material for each speaker model is based on concatenating 5 utterances. This setup results in 500 client scores and 129,500 impostor scores. The speech material used for building the UBM consists of 10 utterances from each of 200 speakers other than the ones registered or used as unknown speakers. It should be noted that the speaker set used for UBM and the sets of registered and unknown speakers are all gender-balanced.

4.2.2 Feature Extraction

For the purpose of the work described in this study, the t^{th} frame of the input speech data is represented as $\mathbf{c}_t \equiv \{c_t(1), c_t(2),\ldots, c_t(K), \Delta c_t(1), \Delta c_t(2),\ldots, \Delta c_t(K)\}$, where $c(k)$ is the k^{th} mean subtracted, linear predictive coding-derived cepstral (LPCC) parameter and $\Delta c(k)$ is the k^{th} delta LPCC parameter. The extraction of LPCC parameters is based on pre-emphasising the input speech data using a first order digital filter, performing Voice Activity Detection and then segmenting it into 20 ms frames at intervals of 10 ms using a Hamming window. As discussed in Chapter 3, the value of K is dependent on the sampling frequency of the speech data and is chosen as (F+4), where F is the sampling frequency. The values for K used for the considered speech databases are given in Table 4.1.

Dataset	Sampling Frequency	Dimension of feature vector
NIST SRE 2003	8 kHz	12 LPCC+ 12 Δ
TIMIT	16kHz	20 LPCC+ 20 Δ

Table 4.1: Dimensions of the feature vector for the two different datasets.

4.2.3 GMM-UBM Baseline

The baseline system used in this study is based on Gaussian mixture models (GMM). Each speaker model is adapted from a 128 mixture, gender-independent UBM using *m*MAP adaptation. The Gaussian mixture densities are parameterised with mean vectors and diagonal covariance matrices. As described in Section 4.1.1, during the test phase, a fast scoring procedure is carried out to obtain the log-likelihood score of the test utterance with respect to the target model. The match score is then subjected to UBM-based normalisation. As mentioned in Chapter 3 (Section 3.3.1), over the last decade, this approach has become one of the dominant approaches for modelling a person's voice in speaker verification applications [12, 26, 73, 120]. For this reason, the baseline GMM-UBM approach is adopted as one of the state-of-the-art speaker verification system for the purposes of the experimental investigations described in Section 4.3.

4.2.4 GMM-SVM speaker verification

The structure of the GMM-SVM system used in this study is illustrated in Figure 4.2. The GMMs are obtained from training, testing and background utterances using the same procedure as that in the GMM-UBM system. The GMM supervectors are then extracted, projected out using NAP to remove session variability, and then normalised using M-norm. Next, using the statistics obtained from the background dataset, the supervectors are scaled to unit variance. As mentioned in the previous chapter, this approach which is referred to as the background data scaling kernel (BSL), is carried out to allow each dimension of the supervector to contribute equally to SVM training and subsequent testing [29]. This is followed by SVM training to obtain the client models. In the test stage, the computation of classification scores is based on equation (4.5).

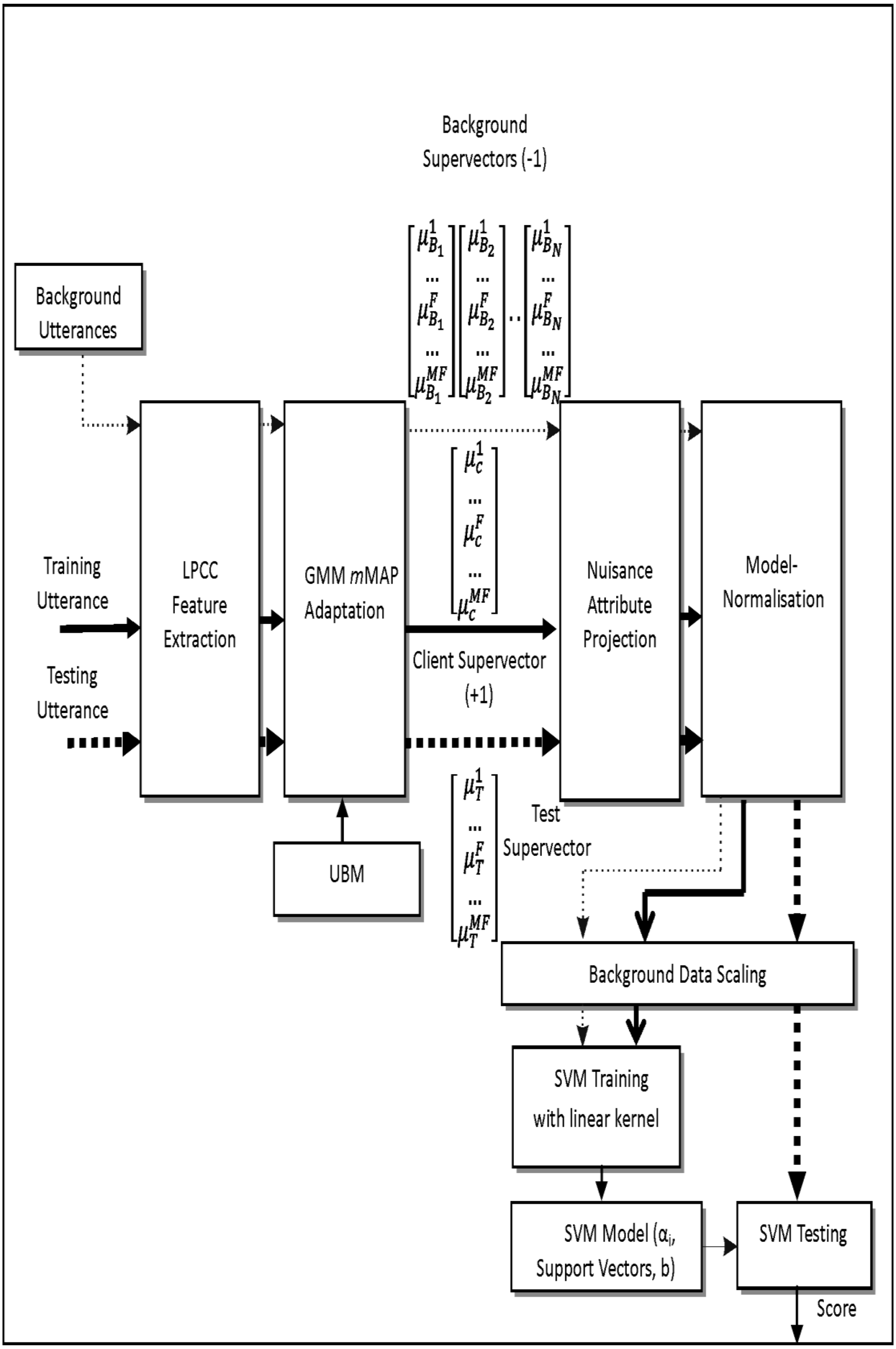

Background Supervectors (-1)
$\mu_{B_1}^{1}$ ... $\mu_{B_1}^{F}$ ... $\mu_{B_1}^{MF}$
$\mu_{B_2}^{1}$ ... $\mu_{B_2}^{F}$ ... $\mu_{B_2}^{MF}$
$\mu_{B_N}^{1}$ ... $\mu_{B_N}^{F}$ ... $\mu_{B_N}^{MF}$
Background Utterances
Training Utterance
Testing Utterance
LPCC Feature Extraction
GMM mMAP Adaptation
μ_{c}^{1} ... μ_{c}^{F} ... μ_{c}^{MF}
Client Supervector (+1)
Nuisance Attribute Projection
Model-Normalisation
UBM
μ_{T}^{1} ... μ_{T}^{F} ... μ_{T}^{MF}
Test Supervector
Background Data Scaling
SVM Training with linear kernel
SVM Model (αi, Support Vectors, b)
SVM Testing
Score

Figure 4.2: Illustration of the GMM-SVM Speaker verification system

It should be pointed out that the Background data Scaling Linear kernel is adopted in this study based on some preliminary investigations where it has been shown to give very similar performance to the GSV kernel, non-linear kernel or MLLR kernel. This could be attributed to the relatively small[6] number of background supervectors (negative examples) which are utilised in the context of the present study when compared to other published studies [14, 15, 127, 134-136]. To be precise, while most other studies utilise a complete NIST database for this purpose, in this study, the background supervectors are obtained using the same data as that used for training the UBM training [29, 137]. It should also be noted that a 128 mixture UBM is being used to limit the size of the supervector and allow faster training and testing of the SVM models [27, 29].

Table 4.3 shows the result obtained using the GMM-SVM method together with that for GMM-UBM. The comparison of the performance of these two approaches is further illustrated using the DET plots in Figure 4.3. It can be seen from the results, that the GMM-SVM approach reduces the speaker verification error rate by over 27% when compared to the baseline GMM-UBM approach. This is in agreement with the results reported in [28, 71, 130]. In addition, it is observed that applying T-Norm on top of M-Norm in the GMM-SVM approach does not provide any significant reduction in EER. Based on the outcomes of the experimental investigations, the GMM-SVM approach is therefore adopted as the other state-of-the-art speaker verification system for the purposes of the experimental investigation in the next section.

SV-EER (%)			
GMM-UBM		**GMM-SVM**	
UBM Normalisation	T-Norm	Model Normalisation	T-Norm
10.47 ±0.85	9.68±0.82	7.51 ± 0.73	7.50 ± 0.72

Table 4.2: Relative effectiveness of the GMM-SVM approach based on the NIST SRE 2003.

[6] This is due to lack of available and appropriate data for this purpose in this study

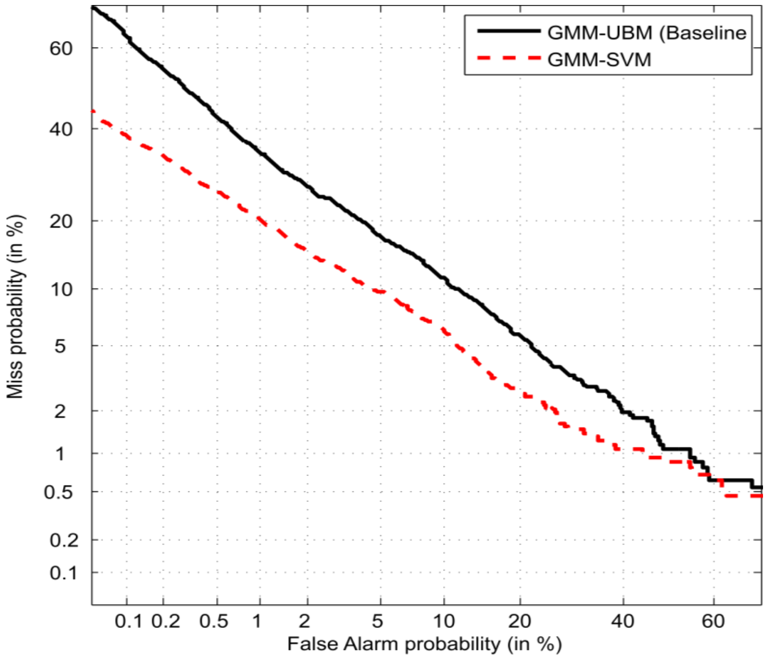


Figure 4.3: Relative verification effectiveness offered by the GMM-SVM approach based on the NIST SRE 2003.

4.3 Relative effectiveness of GMM-UBM and GMM-SVM

To date, most of the investigations with the current state-of-the-art speaker verification techniques have been carried out using the NIST SRE databases. This means the investigations have been limited in terms of the difference between the levels of noise contamination in the training and testing data. This is a condition which cannot be considered realistic in many real-world applications. For instance, the mobile nature of many speaker verification applications can result in noisy test data conditions which are not experienced in the training stage. These can potentially lead to severe degradation of the system performance. Up till now, the literature appears to lack extensive evaluations of the aforementioned techniques under unseen noisy conditions, which is believed to be crucial in establishing their effectiveness in more stringent and realistic scenarios.

This section presents an evaluation of the GMM-UBM and GMM-SVM techniques for matched and mismatched levels of noise contamination during the training and testing stages. It should be noted that the said approaches implemented for this part of the study are the same as the ones described in the previous sections.

4.3.1 Matched Noise Conditions

The first set of experiments evaluates the speaker verification performance of GMM-UBM and GMM-SVM using the TIMIT database when the quality of the speech data is the same during the training and testing phases. For this purpose, the speech data is contaminated with different levels of Gaussian white noise. This provides a range of speech SNRs (15dB, 10dB, 5dB) in addition to uncontaminated speech for the purpose of investigations.

Although it is known that conventional score normalisation techniques such as Test-Normalisation (T-Norm) [73] can offer good improvements with the GMM-UBM approach, its potential benefits have not yet been thoroughly investigated in the GMM-SVM context. This approach is therefore deployed in this study by using the cohort of speakers available within the set of registered users during the test phase.

Table 4.4 presents the experimental results for this part of the study, in terms of Speaker Verification Equal Error Rate (SV-EER) with a 95% confidence interval. It is observed that in clean conditions, the performance of GMM-UBM appears to be better than that of GMM-SVM. As expected, it is seen that there is a drop in accuracy for both approaches with decreasing Signal to Noise Ratio (SNR), although GMM-UBM yields better verification rates for a contamination level of 10dB. It is also observed that the use of score normalisation provides further improvements for both classification methods. This is particularly evident for the 10dB scenario, where the error rate is almost halved with the inclusion of T-Norm in the GMM-SVM approach. The use of M-Norm for GMM-SVM, which involves scaling the GMM means in the supervectors with respect to the UBM in order combat variations, appears to have limited effects in this situation. In this setup, it can be argued that such a phenomenon arises, because all the models are adapted from a clean gender-balanced UBM regardless of the noise degradation of their speech feature vectors.

SV-EER (%)									
GMM-UBM					***GMM-SVM***				
Test/Training Data					**Test/Training Data**				
Score normalisation	*Clean*	*SNR: 15dB*	*SNR: 10dB*	*SNR: 5dB*	*Score normalisation*	*Clean*	*SNR: 15dB*	*SNR: 10dB*	*SNR: 5dB*
Clean UBM	2.00 ±0.62	5.60 ±1.02	11.80 ±1.44	28.52 ±2.02	**Without additional normalisation**	2.59 ±0.71	6.92 ±1.13	15.87 ±1.63	29.60 ±2.04
T-norm	1.60 ±0.56	3.40 ±0.81	7.21 ±1.16	16.20 ±1.65	**T-norm**	1.60 ±0.50	4.42 ±0.91	8.23 ±1.23	19.24 ±1.76

Table 4.3: Speaker verification results for GMM-UBM and GMM-SVM in matched data conditions.

4.3.2 Mismatched Noise Conditions

The purpose of the next set of experiments is to determine the effectiveness of GMM-UBM and GMM-SVM in the absence of information about the noise conditions during the test trials in relation to that in the training phase. In order to create such a condition, clean training data is used during the modelling process while degraded data is used in the test phase. Although different scenarios such as degraded training data/clean testing data or degraded training data/degraded testing data with mixed contamination levels can also be considered, it is believed that the setup chosen should provide a reasonably accurate indication of the problem of unseen data conditions. As before, speech data from the TIMIT database is used and Gaussian white noise is added to degrade the test data, achieving SNRs of 15 dB, 10 dB and 5 dB respectively. In addition, three examples of real-world noise, namely car noise, office noise, and factory noise, obtained from the NOISEX 92 [138] and Piper [34] databases are also used in the experimental investigations. For each noise type, the test data is contaminated using a randomly selected segment (with the same duration as the test utterance) of the original noise file to achieve SNRs of 15dB, 10dB and 5dB.

The experimental results given in Table 4.5 show that the verification EERs for GMM-UBM are higher for mismatched conditions with Gaussian white noise when compared to those for matched noisy conditions (Table 4.4). Interestingly, it is seen that such a trend does not apply in the case of GMM-SVM which yields comparable results to the matched conditions for SNRs of 15 dB and 10 dB although worse results are obtained for an SNR of 5 dB. The performance of the two classification methods in this scenario is seen to be very similar for SNRs of 15 dB and 10 dB, while GMM-SVM performs slightly better than GMM-UBM under the worst condition considered (i.e. 5dB). In addition, it is observed that the use of T-Norm in this setup, unlike the previous scenario, does not have a significant effect on the performance of the two classification methods.

SV-EER (%)							
GMM-UBM				***GMM-SVM***			
	Test Data				**Test Data**		
Score normalisation	*SNR: 15dB*	*SNR: 10dB*	*SNR: 5dB*	*Score normalisation*	*SNR: 15dB*	*SNR: 10dB*	*SNR: 5dB*
Clean UBM	6.80 ±1.12	16.60 ±1.66	37.40 ±2.16	**Without additional normalisation**	6.24 ± 1.08	15.20 ±1.61	33.20 ±2.10
T-norm	5.20 ±0.99	15.53 ±1.62	36.20 ±2.15	**T-norm**	5.40 ±1.01	14.57 ±1.58	33.40 ±2.11

Table 4.4: EERs in speaker verification experiments with GMM-SVM and GMM-UBM under mismatched data conditions using Gaussian white noise.

Table 4.6 presents the verification experiments involving mismatched conditions with a range of contaminated speech using real world noise. Although the degradation in performance for GMM-SVM and GMM-UBM with real-world noise is not as severe as that for Gaussian white noise, a considerable increase in SV-EER is still observed with decreasing SNRs. It is also observed that, in general, the difference between the effectiveness of the two methods is not significant for any type of real-world noise considered. Additionally, it is noted that again, the usefulness of T-Norm in reducing error rates is rather limited.

SV-EER (%)								
	GMM-UBM				***GMM-SVM***			
	Test Data				**Test Data**			
Noise	*Score normalisation*	*SNR: 15dB*	*SNR: 10dB*	*SNR: 5dB*	*Score normalisation*	*SNR: 15dB*	*SNR: 10dB*	*SNR: 5dB*
Car	**Clean UBM**	4.99 ±0.97	7.80 ±1.19	16.74 ±1.67	**Without additional normalisation**	5.20 ±0.99	8.09 ±1.22	16.60 ±1.66
	T-norm	4.00 ±0.87	7.20 ±1.16	15.64 ±1.62	**T-norm**	5.00 ±0.97	7.40 ±1.17	15.20 ±1.61
Office	**Clean UBM**	4.84 ±0.96	7.80 ±1.20	18.60 ±1.74	**Without additional normalisation**	5.43 ±1.04	8.79 ±1.27	18.60 ±1.74
	T-norm	3.84 ±0.86	7.59 ±1.18	18.20 ±1.73	**T-norm**	5.20 ±0.99	8.20 ±1.23	18.16 ±1.72
Factory	**Clean UBM**	4.60 ±0.94	6.68 ±1.11	17.60 ±1.70	**Without additional normalisation**	5.45 ±1.01	8.12 ±1.22	18.52 ±1.71
	T-norm	4.60 ±0.94	6.68 ±1.11	17.60 ±1.70	**T-norm**	5.00 ±0.97	7.34 ±1.17	17.24 ±1.69

Table 4.5: EERs in speaker verification for GMM-SVM and GMM-UBM under mismatched data conditions using real world noise.

4.4 Chapter Summary

In this chapter, the effectiveness of the current state of the art speaker verification approaches has been experimentally analysed. The first part of the experiments has provided investigations into the relative performance of the most widely used approaches for speaker verification using the NIST SRE 2003 database. It is shown that the SVM with GMM supervector approach coupled with the Nuisance Attribute Projection (NAP) and Model-normalisation (M-Norm) provides substantial improvements over the baseline GMM-UBM system. In the second part of the investigations, the relative effectiveness of the GMM-UBM and GMM-SVM approaches has been analysed under matched and mismatched data conditions. In this study, the main limitations of the two classification approaches have been outlined. It is observed that when the test data is degraded with Gaussian noise or real-world noise, in general, the difference between the effectiveness of the two methods is not significant under either matched or mismatched data conditions. It is also noted that while T-Norm can be very beneficial in further improving the accuracy of both classification methods under matched data conditions, its usefulness in reducing error rate under mismatch conditions is rather limited.

CHAPTER 5

IMPROVING THE SPEAKER RECOGNITION ACCURACY UNDER MISMATCH CONDITIONS

Chapter Overview

It is observed in the previous chapter that the problem of mismatched data conditions can severely affect the performance of state-of-the-art speaker verification techniques. In this chapter, a modified realisation of the parallel model combination (PMC) method is introduced and a new form of test normalisation (T-norm), termed condition adjusted T-norm, is proposed to tackle this problem. An account of the motivation behind the modified PMC GMM-UBM approach, together with a description of its characteristics, is given in Section 5.1. This is followed by a set of experimental investigations to evaluate its effectiveness in relation to the full PMC GMM-UBM approach. Section 5.2 introduces the concept of condition-adjusted T-Norm, investigates its relative effectiveness under different mismatched data conditions and presents an analysis of the results. In section 5.3, a bilateral PMC GMM-UBM approach is proposed and its relative effectiveness investigated for speaker verification operating under conditions where the training and testing utterances are both contaminated with noise. Section 5.4 introduces the use of the modified PMC GMM-UBM with CT-Norm approach into the context of OSTI-SI. Based on the outcomes of the experimental investigations, it is demonstrated that the said approach can be of considerable value for both speaker verification and speaker identification.

5.1 Modified PMC Approach

As discussed in Chapter 2, several speech-level [72, 77, 80] and feature-level approaches [76, 84-86] have been proposed in the literature for tackling the effects of variations between the training and test data caused by additive noise. These approaches usually focus on enhancing the quality of the test material before the testing process. In other words, they assume that the training material is free from any form of degradation. In many practical applications, however, the training and testing utterances can both be degraded. Since the characteristics of degradation in these utterances can be considerably different, the actual problem is one of minimising the data mismatch conditions and/or the effects of these. To address this problem, the use of a data-driven parallel model combination (PMC) has been proposed in [92]. The technique involves estimating the degradations in the testing and training material and using these to minimise the data mismatch conditions (by appropriately contaminating the reference model and test utterance in each trial). The investigations in [92], which have been based on the use of decoupled GMMs, provide a clear indication of the potential benefits of PMC. In the case of GMM-UBM, the direct use of PMC involves a complete reference model generation process in the test phase. Such a process includes rebuilding a UBM (with degraded speech material) as well as the adaptation of the new UBM using the degraded version of the training utterances for the target speaker. Repeating this whole process (in particular, rebuilding a new UBM) for each test trial can unduly increase the computational load of the GMM-UBM approach. Thus, in order to enhance the computational efficiency in the test phase, the use of a modified PMC procedure is proposed. As seen in Figure 5.1, during the test phase, an estimation of the test noise is used to contaminate the target speaker’s training material. A noise compensated target speaker model is then obtained through the *m*MAP adaptation [12] of a UBM trained *a priori* (offline) using clean speech (based on the corresponding contaminated training material). Finally, the noise degraded test utterance is matched against the noise compensated target speaker model and the clean UBM to obtain a likelihood ratio score (i.e. UBM normalisation) which is then used to decide whether to accept or reject the claimant.

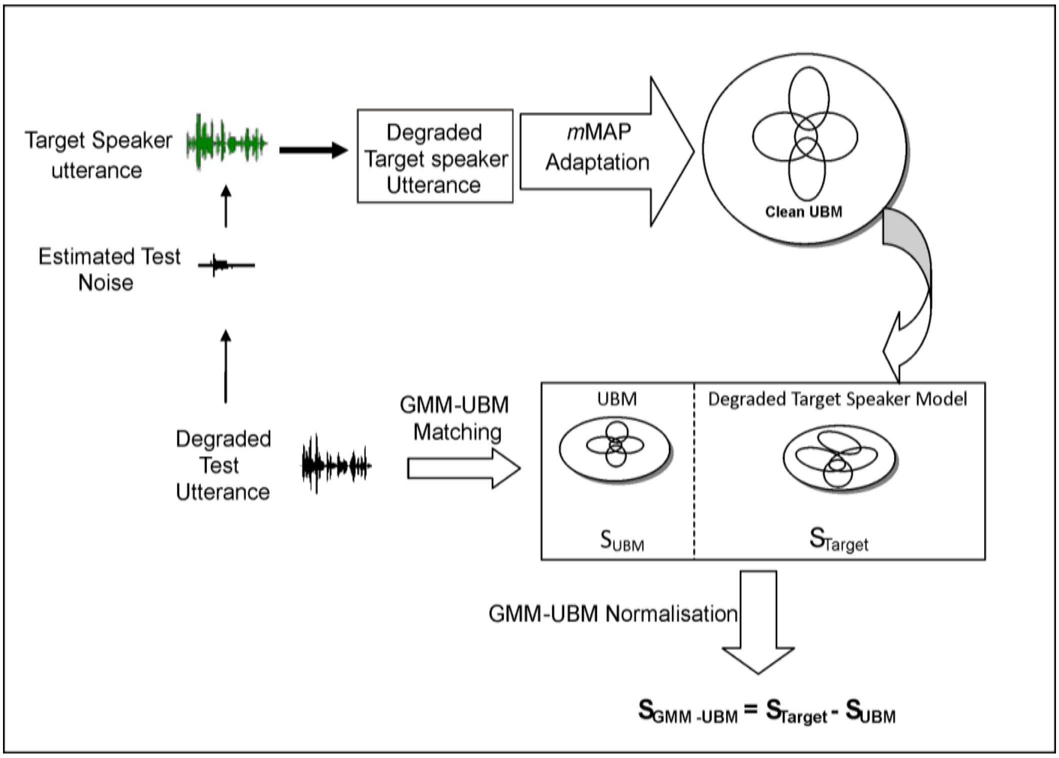


Figure 5.1: Illustration of the proposed procedure for obtaining compensated client models using PMC.

5.1.1 Experimental Investigations and results

In order to determine the effectiveness of the proposed approach relative to that of the direct use of PMC with GMM-UBM, a set of pilot experiments is carried out using car noise. For the sake of comparison, the speech dataset and speaker representation used for the purpose of the experiments in this study are identical to those in Chapter 4. A brief summary is provided in Table 5.1.

Database :TIMIT [9]	Speech Feature Vectors: 20th order LPCC + Delta
Number of registered speaker :100 Number of unknown speakers:80	UBM Characteristics: Gender independent trained using 200 speakers
Number of client scores: 500 scores Number of impostor scores: 129,500	GMM-UBM based on modified MAP adaptation [2]

Table 5.1: Summary of the experimental setup

The procedure deployed for contaminating test utterances is the same as that described in Chapter 4. For the purpose of PMC, in each test trial, the first 200 ms of noise used for degrading the test utterance is considered as an estimate of the test utterance contamination. The results obtained for these two methods are presented in Table 5.2 and Table 5.3. As before, all the results for this part of the experimental investigations are presented in terms of Speaker Verification Equal Error Rate (SV-EER) with a 95% confidence interval.

It is observed by comparing the results in tables 5.2 and 5.3 with those obtained in Table 4.6, that whilst the direct use of PMC with GMM-UBM can significantly enhance the verification accuracy under the noise-mismatch condition considered, the results for the modified approach are not as impressive. This is further illustrated in Figure 5.2. The relative superior performance of the direct PMC GMM-UBM is due to building UBM using speech degraded based on an estimation of the test utterance contamination. In real applications, however, such rebuilding of UBM in each test trial may not be practical because of the additional computational cost involved.

SV-EER (%)				
Modified PMC GMM-UBM				
		Test Data		
Noise	*Score normalisation*	*SNR: 15dB*	*SNR: 10dB*	*SNR: 5dB*
Car	**Clean UBM**	5.94±1.05	7.74±1.19	10.66±1.30

Table 5.2: Verification results for the proposed PMC GMM-UBM method in mismatched data conditions using car noise.

SV-EER (%)				
Full PMC GMM-UBM				
		Test Data		
Noise	*Score normalisation*	*SNR: 15dB*	*SNR: 10dB*	*SNR: 5dB*
Car	**Appropriately degraded UBM**	2.60±0.71	2.83±0.74	5.20±0.99

Table 5.3: Verification results for the direct PMC GMM-UBM method in mismatched data conditions using car noise.

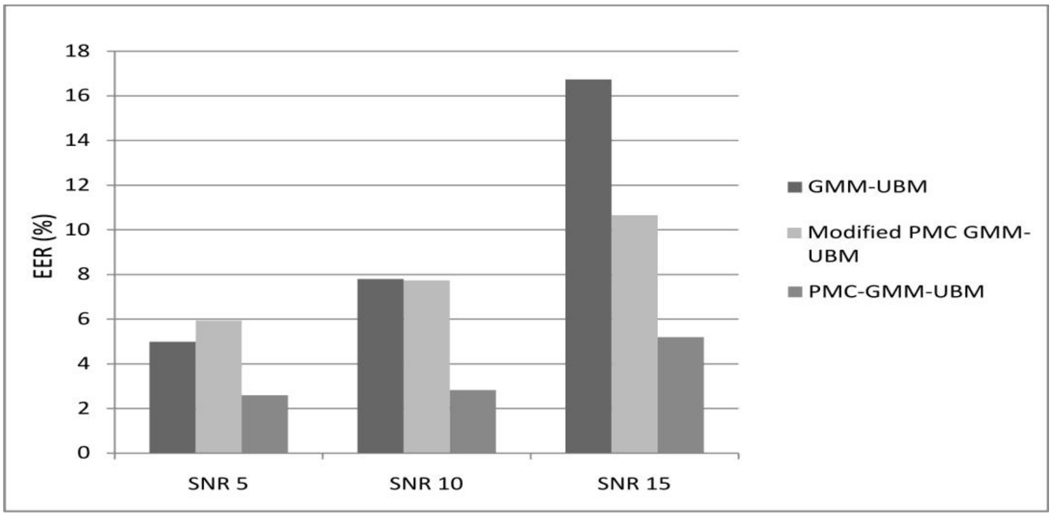


Figure 5.2: Relative effectivess of Modified PMC GMM-UBM approach for car noise

5.2 CT-Norm for speaker verification

It is seen in the previous section that although the modified PMC GMM-UBM approach offers enhanced computational efficiency, it is not as effective as the full PMC GMM-UBM method in dealing with the effects of mismatch noise conditions. This can be attributed to the mismatch between the clean UBM and the noise compensated target model which in turn, does not provide an effective means of score normalisation. Similarly, it has been observed in Chapter 4, that whilst T-Norm can be very beneficial in improving the verification accuracy under relatively matched noisy conditions, its usefulness in reducing error rate under mismatch conditions is rather limited. Thus, in order to tackle this problem while retaining the computational efficiency of the modified PMC GMM-UBM method, a condition adjusted T-Norm (CT-Norm) approach is proposed. As shown in Figure 5.3, this approach involves adjusting the noise contamination of the target speaker utterance as well as background speaker utterances in accordance with the estimated test utterance degradation. Noise adjusted target and background speaker models are then obtained through the *m*MAP adaptation of a fixed clean UBM. During the matching phase, the degraded test utterance is scored against the condition adjusted speaker models (i.e. target and background). Following this, the required normalisation parameters (i.e. mean and variance) are computed, using the likelihood scores of the background speaker models, and Test-normalisation (T-Norm) is applied.

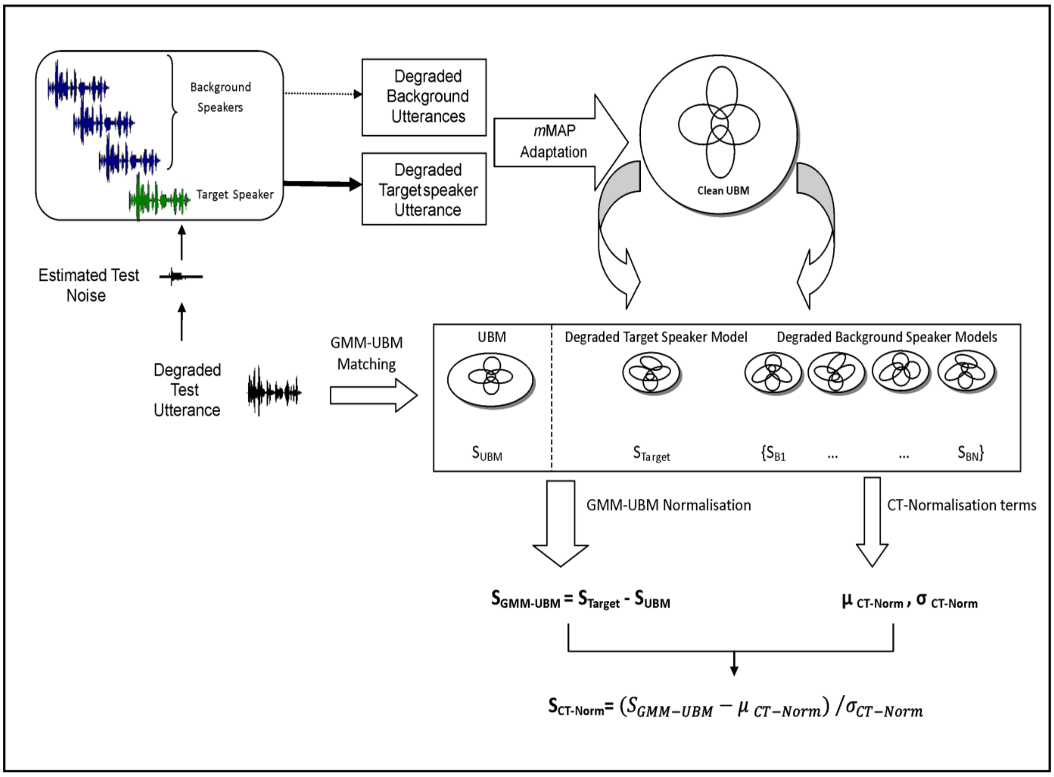

Background Speakers
Target Speaker
Degraded Background Utterances
Degraded Targetspeaker Utterance
mMAP Adaptation
Clean UBM
Estimated Test Noise
Degraded Test Utterance
GMM-UBM Matching
UBM
Degraded Target Speaker Model
Degraded Background Speaker Models
S_{UBM}
S_{Target}
$\{S_{B1}$ ... ... $S_{BN}\}$
GMM-UBM Normalisation
CT-Normalisation terms
$S_{GMM-UBM} = S_{Target} - S_{UBM}$
$\mu_{CT-Norm}, \sigma_{CT-Norm}$
$S_{CT-Norm} = (S_{GMM-UBM} - \mu_{CT-Norm}) / \sigma_{CT-Norm}$

Figure 5.3: Illustration of the proposed procedure for improving verification accuracy in mismatched data conditions.

To examine the effectiveness of CT-norm, a set of experimental investigations is conducted with the modified PMC GMM-UBM, and using the three types of real world noise considered. The procedures used for the noise-based degradation of test utterances, and estimating the resulting contamination in the test phase are the same as those discussed in Chapter 4 and Section 5.1 respectively. It should be noted that the implementation of CT-Norm is based on the training utterances from the cohort of speakers available within the set of registered users (i.e. 99 speakers on each occasion).

Table 5.4 presents the results of this study. These results provide a clear indication of the effectiveness of CT-norm in reducing the verification error rates under different noise mismatch conditions. It can be seen by comparing the results in Table 4.6 with those obtained in Table 5.4, that the improvements achieved are particularly significant for the worst data conditions (i.e. 10dB and 5dB) where the minimum relative improvements in the case of factory noise are in excess of 61% and 69% respectively. This is further illustrated in Figure 5.4. It is also noted that the results for car noise are comparable or better than those obtained with the direct

PMC GMM-UBM (Table 5.2). The relative effectiveness improvements offered by the CT-norm approach under mismatch conditions are further illustrated through the DET plots in Figure 5.5 (car noise), Figure 5.6 (office noise) and Figure 5.7 (factory noise). In all cases, the SNR for the test data is 10 dB. These Figures clearly show the advantages offered by CT-Norm over the standard T-Norm method.

SV- EER (%)				
Modified PMC GMM-UBM				
		Test Data		
Noise	*Score normalisation*	*SNR:15dB*	*SNR:10dB*	*SNR:5dB*
Car	**Clean UBM**	5.94±1.05	7.74±1.19	10.66±1.30
	CT-norm	2.00±0.62	2.60±0.71	3.55±0.82
Office	**Clean UBM**	5.00±1.85	8.67±1.25	19.40±1.76
	CT-norm	2.60±0.71	3.53±0.82	7.20±1.15
Factory	**Clean UBM**	6.04±1.06	8.60±1.25	19.08±1.75
	CT-norm	1.85±0.60	2.20±0.65	4.43±0.92

Table 5.4: Effectiveness offered by CT-norm in speaker verification based on the modified PMC GMM-UBM approach.

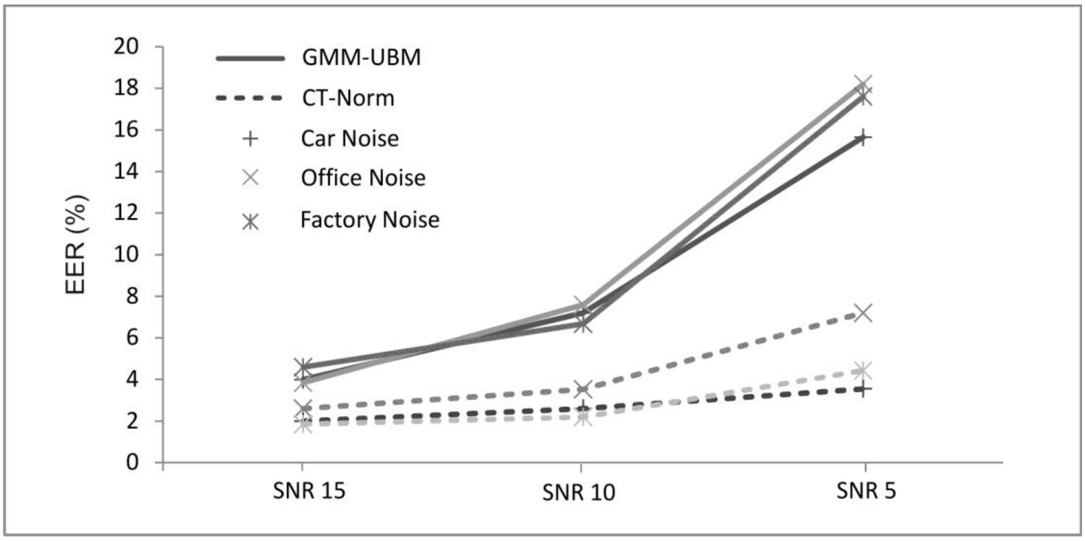


Figure 5.4: Effectiveness of the CT-Norm approach compared to the standard GMM-UBM.

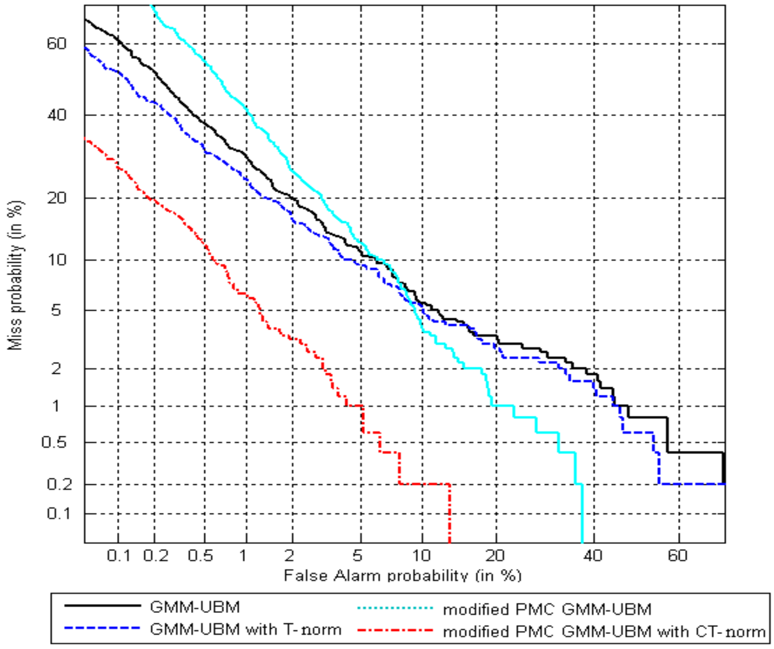


Figure 5.5: Relative verification effectiveness offered by the use of CT-norm with the modified PMC GMM-UBM approach in mismatched data conditions using car noise.

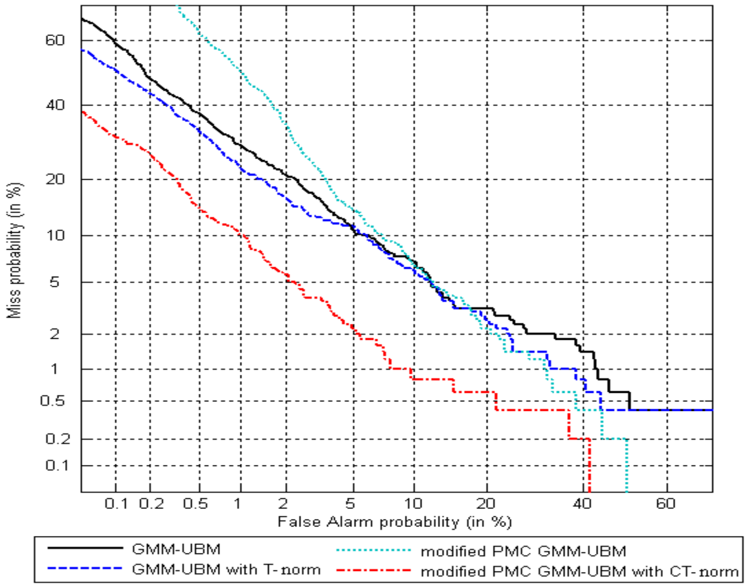


Figure 5.6: Relative verification effectiveness offered by the use of CT-norm with the modified PMC GMM-UBM approach in mismatched data conditions using office noise.

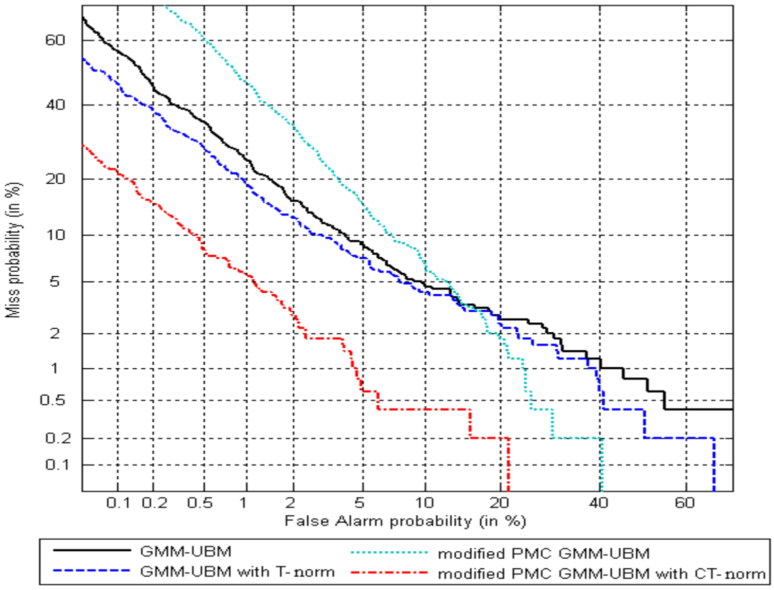


Figure 5.7: Relative verification effectiveness offered by the use of CT-norm with the modified PMC GMM-UBM approach in mismatched data conditions using factory noise.

In the case of GMM-SVM, the use of PMC will require a complete training procedure during each test trial, involving noise-adjusted models for the target and background speakers. Because of the particular characteristics of the SVM procedure involved, this can result in a significant increase in computational load in the test phase. This is because in this case, during each test trial, noise compensated target and background speaker models have to first be built (using the PMC GMM-UBM approach) before their corresponding supervector of means can be extracted for SVM training. Despite this, and for completeness, a set of verification experiments with modified PMC GMM-SVM is conducted using car noise. The investigations are carried out with and without using CT-norm. The results of this study (Table 5.5) again show considerable improvements in verification accuracy when CT-norm is deployed. However, it is also observed that, in this case, the EERs are not as low as those obtained using the modified PMC GMM-UBM with CT-norm (Table 5.4).

SV-EER (%)				
Modified PMC GMM-SVM				
			Test Data	
Noise	*Score normalisation*	*SNR: 15dB*	*SNR: 10dB*	*SNR:5dB*
Car	**No additional normalisation**	5.08±0.98	7.12±1.15	11.21±1.41
	CT-norm	3.00±0.76	4.60±0.94	6.60±1.11

Table 5.5: EERs in verification experiments for PMC GMM-SVM with and without using CT-norm.

5.3 Bilateral Parallel Model Combination

In the previous section, the experimental investigations carried out with CT-Norm have been based on the assumption that the reference material of all the speakers are recorded under controlled conditions and kept free from any noise degradation. This is a favourable assumption which means the characteristics of the reference model for each speaker (client/background) are not influenced by the particular type of noise present at the time of enrolment. However, in practice, imposing such a stringent condition during the enrolment process is not always feasible. As a result, most speaker verification applications operate on a more realistic assumption; that is, the noise contamination of the training utterances used for speaker modelling is considered to be reasonably limited.

The aim of the experiments presented in this section is to investigate the effectiveness of the CT-Norm approach when both the training and testing utterances are contaminated with environmental noise. Under this setup, two approaches for Parallel Model Combination (PMC) are investigated. The first method is based on the modified PMC GMM-UBM proposed in the previous section. The second method, on the other hand, involves a two-stage noise contamination process. The first stage involves contaminating the training utterances for each speaker (client and background) using an estimate of the test noise. Noise-adjusted speaker models are then built by appropriately adapting a fixed (original) UBM. This is identical to the modified PMC GMM-UBM approach. In the second stage, the test utterance is also contaminated using an estimate of the noise present in the training utterance. The complete approach is hereafter referred to as Bilateral PMC GMM-UBM.

The experimental setup for this part of the study is based on a highly unfavourable scenario. This involves contaminating the training material with three examples of real-world noise (i.e. car noise, office noise, and factory noise), obtained from the NOISEX 92 [138] and Piper [34] databases to achieve SNRs of 15dB. The test material is then contaminated using a different type of noise to the one used for contaminating the training utterances in each case in order to achieve a SNR of

5dB. For example if car noise is used to contaminate the training material, office or factory noises are then used to contaminate the test material. The results for this part of the investigations are presented in Table 5.6.

		SV-EER (%)		
Training Data Noise	**Test Data Noise**	**Normalisation**	**modified PMC GMM-UBM**	**Bilateral PMC GMM-UBM**
Car SNR 15dB	Office SNR 5dB	CT-Norm	6.89 ±1.13	6.60 ±1.11
	Factory SNR 5dB	CT-Norm	6.20 ±1.07	5.90 ±1.04
Office SNR 15dB	Car SNR 5B	CT-Norm	3.60 ±0.83	3.60 ±0.83
	Factory SNR 5dB	CT-Norm	6.20 ±1.07	6.20 ±1.07
Factory SNR 15dB	Car SNR 5B	CT-Norm	3.94 ±0.87	3.80 ±0.85
	Office SNR 5dB	CT-Norm	7.00 ±1.14	6.80 ±1.13

Table 5.6: Relative effectiveness of the Bilateral PMC GMM-UBM approach when both the training and testing utterances are contaminated with real-world noise**.**

It can be observed from Table 5.6 that, in general, the relative effectiveness of the Bilateral PMC GMM-UBM approach with CT-Norm is very similar to those offered by the modified PMC GMM-UBM method with CT-Norm. For instance, the best relative improvement obtained (out of the six different scenarios considered) with the Bilateral PMC GMM-UBM approach is only about 3%. This is obtained when the training material is degraded with factory noise and the test material is contaminated with car noise. Evidently, despite the added computational complexity associated with Bilateral PMC GMM-UBM there no significant advantages. Therefore, for the purpose of consistency, the experimental investigations in the remainder of this thesis are based on the use of clean training utterances.

5.4 Performance of OSTI-SI under mismatched noise conditions

As described in Chapter 1, the problem of automatic speaker identification can be defined as one of determining the speaker of a given test utterance, from a population of registered speakers [9]. If the process includes the option of declaring that the test utterance does not belong to any of the registered speakers, it is termed open-set speaker identification. Otherwise, it is a closed-set identification process. In principle, the process of open-set speaker identification consists of two successive stages of identification and verification. In other words, first, it is required to identify the speaker model in the set which best matches the given test utterance. Then, it must be verified whether the test utterance has actually been spoken by the speaker associated with the best-matched model, or by some unknown speaker outside the registered set. When there are no constraints on the text content of test utterances, the process is referred to as open-set, text-independent speaker identification (OSTI-SI) [9]. This is the most challenging class of speaker recognition with applications in various areas including document indexation, surveillance, and authorisation control in smart environments.

As with the speaker verification scenario, a factor adversely affecting the accuracy of OSTI-SI in practice is that of variations in speech characteristics [9, 99]. Such variations result is a mismatch between the corresponding test and reference material for the same speaker, which in turn reduces the accuracy of OSTI-SI.

Similar to the speaker verification scenario, a widely used approach for tackling the problem of mismatched noise conditions in speaker identification is that of score normalisation [73, 96, 99]. However, as seen in the previous chapter, in general, the effectiveness of score normalisation reduces considerably when the data mismatch, resulting from noise contamination in the test material, becomes significant [139]. As indicated in Section 5.2, the use of CT-Norm (condition adjusted T-Norm) can significantly reduce the adverse effects of data mismatch on the accuracy of speaker verification. However, as mentioned earlier, the problem in the second stage of OSTI-SI is more challenging than that of the standard speaker verification [9, 99, 140]. This is due to the fact that the requirement in the second stage of OSTI-SI is to discriminate each out-of-set speaker from its best matched speaker in the registered set. Therefore, it may not be possible to fully predict the effectiveness of CT-Norm in this case, based on the results obtained for SV [139]. Moreover, the benefits of using the computationally efficient PMC GMM-UBM approach for speaker identification also need investigating. The aim of this part of the study is therefore to complement the experiments in the previous sections by investigating the effectiveness of the efficient PMC GMM-UBM approach and CT-Norm in the context of open-set speaker identification.

It is important to point out that, according to the study in Section 5.2, despite its enhanced efficiency, the use of the modified PMC with the GMM-SVM approach can result in an undesirably high level of computational cost. This, as discussed in Section 5.2, is mainly due to the specific characteristics of this SVM-based approach which could make the incorporation of the modified PMC unsuitable for most practical applications. For this reason, the GMM-SVM [28] classification method is not considered in this part of the study.

Figure 5.8 illustrates the use of the modified PMC approach with GMM-UBM for OSTI-SI. As shown in this Figure, an estimate of the test utterance degradation is used to contaminate the training utterances of the registered speakers. The noise-adjusted registered speaker models are then built by appropriately adapting the fixed (original) UBM using an *m*MAP estimation [12, 99]. Once the new models are obtained, the test utterance is matched against all the registered speaker models and the model that yields the largest score is retained. This process is based on the fast scoring procedure using the top five scoring UBM mixtures identified for each test feature vector [12]. As indicated in Figure 5.8, the score for the speaker model selected as above is then subjected to normalisation using T-Norm.

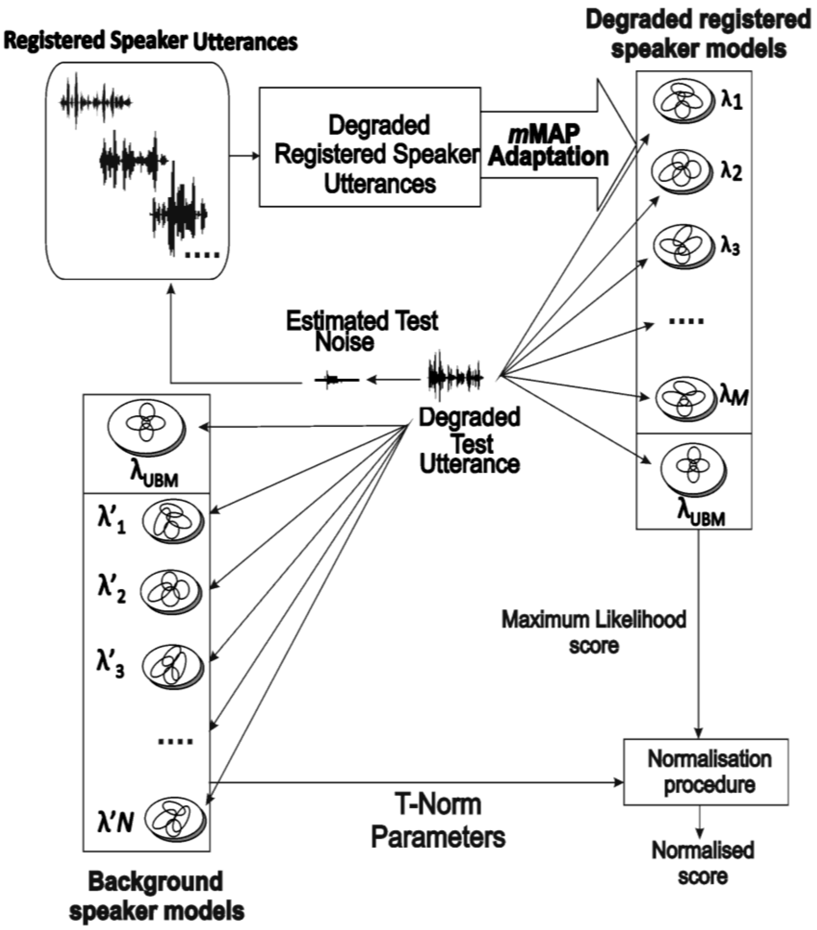


Figure 5.8 : OSTI-SI based on the modified PMC GMM-UBM approach.

Figure 5.9 illustrates the incorporation of CT-Norm (instead of T-Norm) in the OSTI-SI framework presented in Figure 5.8. As observed, the method in Figure 5.9 involves an additional procedure for adjusting the noise contamination of background speaker utterances (and hence their models), in accordance with the estimated test utterance degradation. The determination of the normalisation parameters is then based on these contaminated background speaker models.

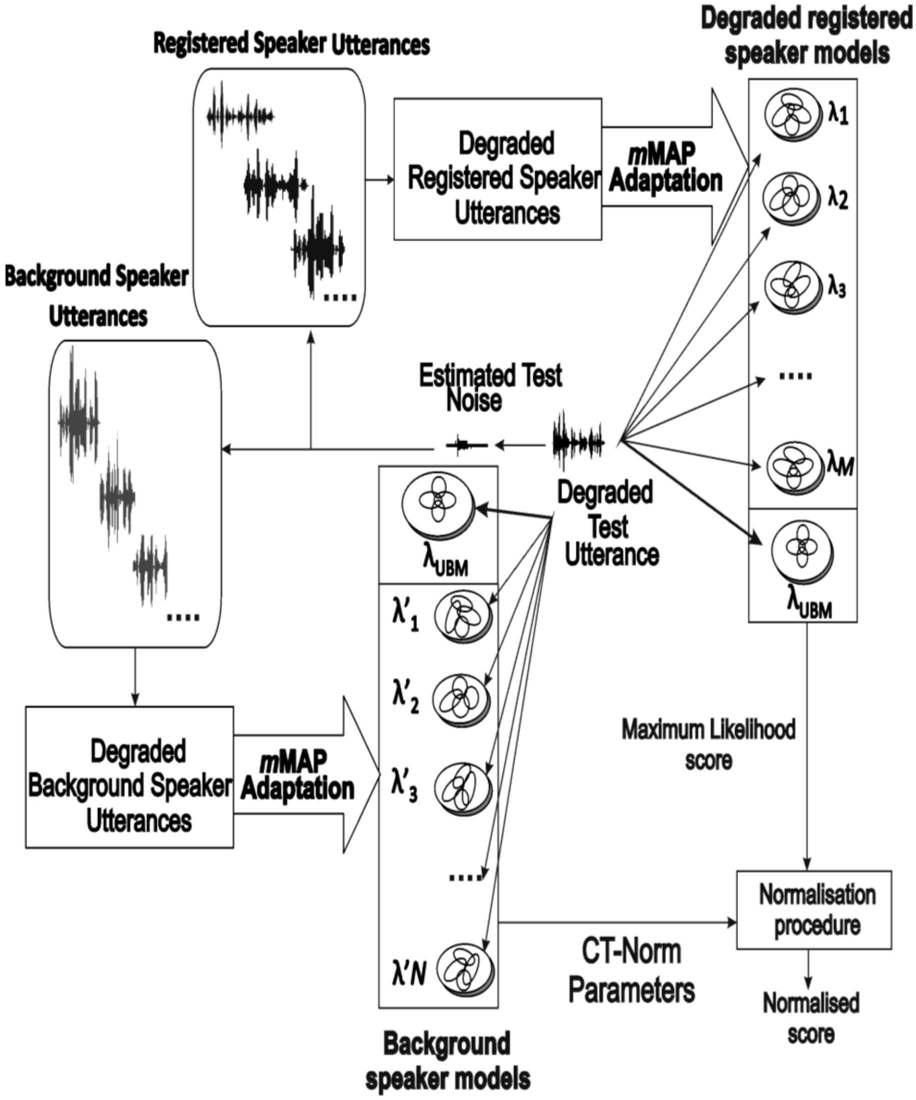

Registered Speaker Utterances
Degraded registered speaker models
Degraded Registered Speaker Utterances
mMAP Adaptation
λ1
λ2
λ3
....
λM
λUBM
Background Speaker Utterances
Estimated Test Noise
Degraded Test Utterance
λUBM
λ'1
λ'2
λ'3
....
λ'N
Degraded Background Speaker Utterances
mMAP Adaptation
Maximum Likelihood score
Normalisation procedure
CT-Norm Parameters
Normalised score
Background speaker models

Figure 5.9: OSTI-SI based on the modified PMC GMM-UBM approach with CT-Norm.

5.4.1 Experimental investigations

The speech dataset used for the purpose of the experimental investigations is extracted from the TIMIT database. 100 registered speakers and 80 unknown speakers are used, each having 10 utterances. Utterances from 200 speakers, other than the ones registered or considered as unknown speakers, are used for training a UBM. As before, it should be noted that the speaker set used for UBM and the sets of registered and unknown speakers are all gender-balanced. In order to facilitate the experimental investigations, in each test trial, the implementation of CT-Norm (or T-Norm where appropriate) is based on the use of the training utterances from the cohort of speakers available within the set of registered users (i.e. 99 speakers on each occasion).

The aim of the first set of experiments is to determine the effectiveness of GMM-UBM for OSTI-SI in the absence of information about the relative noise conditions in the test and training phases. For this purpose, clean training data is used in the modelling process while degraded data is used in the test phase. As before, three examples of real-world noise (i.e. car noise, office noise, and factory noise), obtained from the NOISEX 92 [138] and Piper [34] databases, are used to degrade the test data; achieving SNRs of 15dB, 10dB and 5dB. It should be noted that the experimental setup is identical to the one used for evaluating the proposed approach in the speaker verification context.

5.4.2 Results and discussions

Table 5.7 presents the results in terms of identification error rate (IER) and open set identification equal error rate (OSI-EER) with a 95% confidence interval. It is observed that for all the real world noise types considered, there is a substantial increase in error rates (OSI-EERs and IERs) with decreasing SNR. This is particularly significant for the IERs where a difference in performance of over 50% is observed for data SNRs of 10dB and 5dB. To further illustrate the effects of mismatch conditions on the accuracy of OSTI-SI, the results in Table 5.7 should be compared with those in Table 5.8 which are obtained under clean matched data conditions. These results clearly outline the negative impacts on both the OSI-EERs and IERs, which occur from varying levels of noise degradation between the training and testing data. It is also noted that, similar to the results obtained for speaker verification, the benefits of T-Norm are very limited in the case of considerable mismatched data conditions.

OSI-EER (%)				
GMM-UBM				
		Test Data		
Noise	*Score normalisation*	*SNR:* *15dB*	*SNR:* *10dB*	*SNR:* *5dB*
Car	**Clean UBM**	20.50±1.94	26.50±2.29	31.13±3.29
	T-norm	17.25±1.82	24.38±2.23	30.38±3.24
IER (%)		14.20	26.00	59.60
Office	**Clean UBM**	19.38±1.91	23.88±2.25	31.75±3.57
	T-norm	17.50±1.84	22.75±2.21	31.13±3.24
IER (%)		15.20	28.00	66.00
Factory	**Clean UBM**	20.37±1.94	23.75±2.16	37.50±3.79
	T-norm	18.25±1.85	22.37±2.12	33.50±3.68
IER (%)		13.40	22.60	67.200

Table 5.7: Accuracy of OSTI-SI under mismatch conditions.

OSI-EER (%)	
GMM-UBM	
UBM	13.50 ±0.62
T-Norm	8.00±0.56
IER (%)	5.20

Table 5.8: Performance of OSTI-SI under clean match conditions.

5.4.3 Performance of the condition adjusted normalisation approach

To examine the relative effectiveness of the modified PMC-GMM-UBM with CT-Norm approach for OSTI-SI, a set of experimental investigations is conducted using the setup described in Section 5.4.1. As before, the same set of real world noise is used to degrade the test data and a 200 ms segment of noise is used as the estimation of test utterance contamination. The results for this part of the experimental study are presented in Table 5.9.

OSI-EER (%)				
Modified PMC GMM-UBM				
		Test Data		
Noise	*Score normalisation*	*SNR: 15dB*	*SNR: 10dB*	*SNR: 5dB*
Car	**Clean UBM**	23.88 ±1.96	26.75 ±2.04	31.25 ±2.23
	CT-norm	12.62 ±1.53	17.87 ±1.76	22.00 ±1.99
IER (%)		5.40	5.80	13.60
Office	**Clean UBM**	20.13 ±1.86	26.38 ±2.11	37.50 ±2.63
	CT-norm	13.00 ±1.56	18.25 ±1.85	23.00 ±2.29
IER (%)		6.60	12.60	32.20
Factory	**Clean UBM**	22.50 ±1.91	27.63 ±2.06	36.50 ±2.32
	CT-norm	13.63 ±1.58	15.75 ±1.68	23.00 ±2.03
IER (%)		4.80	5.80	14.00

Table 5.9 : Performance the condition adjusted T-Norm approach.

There are a number of interesting observations which can be made from Table 5.9. Firstly, as expected, it is noted that the use of the modified PMC GMM-UBM on its own does not have any considerable benefits on the accuracy in the second stage of OSTI-SI. It is, however, seen that the said process is considerably beneficial to the accuracy in the first stage, leading to significant improvements in IER for all types of noise considered. For instance, when the test data quality is reduced to 5 dB using factory noise, the improvement achieved in IER (relative to that in Table 5.7) is in excess of 79%. In addition, it is observed that CT-Norm is considerably more effective than T-norm in reducing OSI-EER. Considering all types of noise and degradation levels in this study, the average improvement achieved in OSI-EER relative to the best results in Table 5.9 is about 25%. The relative improvements offered by the CT-Norm approach under mismatch conditions are further illustrated through the DET plots in Figure 5.10. In all cases, the SNR for the test data is 5dB.

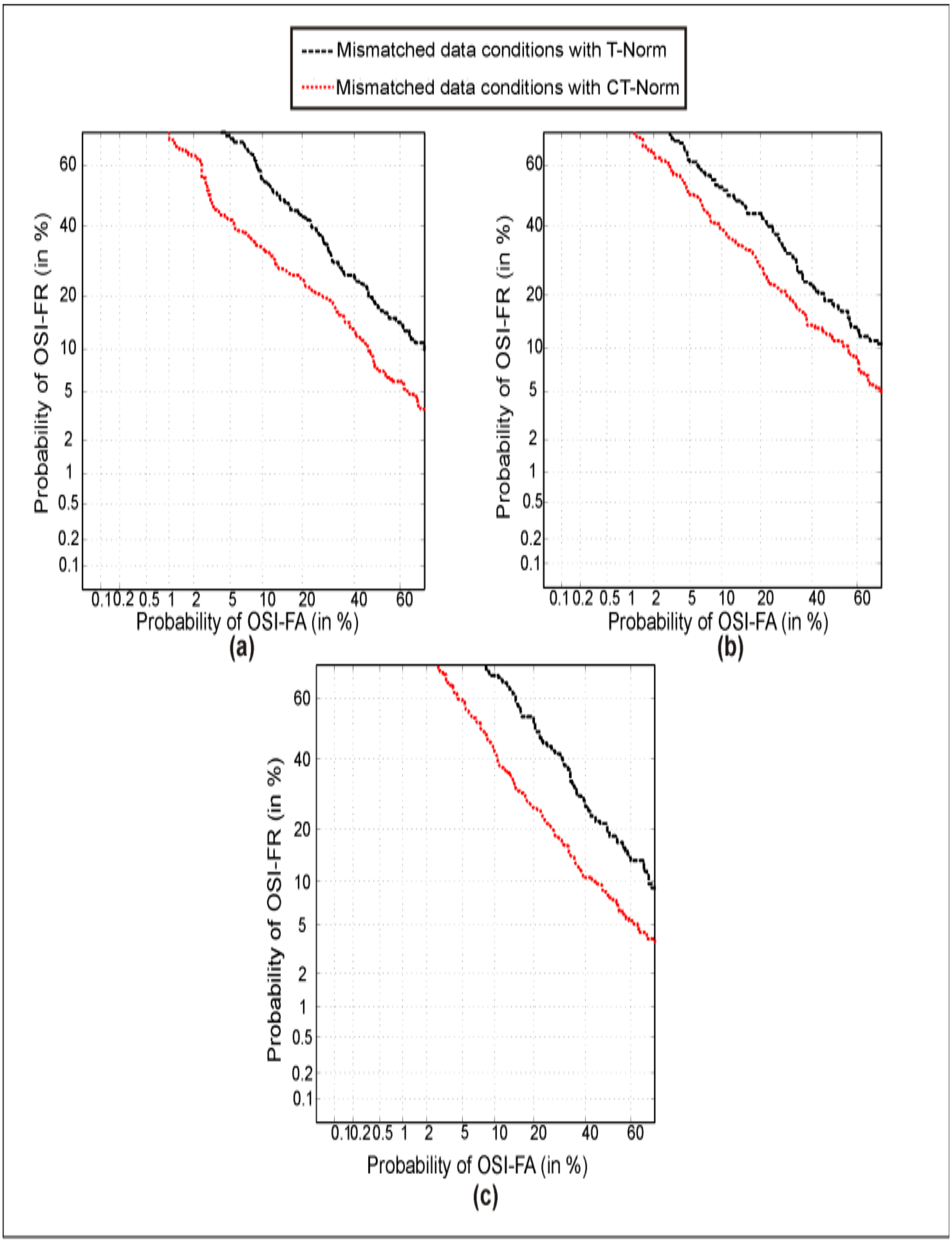


Figure 5.10 : Relative verification effectiveness offered by the use of CT-norm in mismatched data conditions using (a) car noise (b) office noise (c) factory noise.

It is also important to compare the results in Table 5.9 with the corresponding results obtained under the same experimental conditions for speaker verification (Table 5.4). As shown in Figure 5.11, such a comparison clearly shows that the adverse effects of mismatch data conditions are more significant in the second stage of OSTI-SI than in standard SV. It also appears that the proposed method is more effective in standard speaker verification than in the second stage of OSTI-SI. These further highlight the additional challenges in the second stage of OSTI-SI.

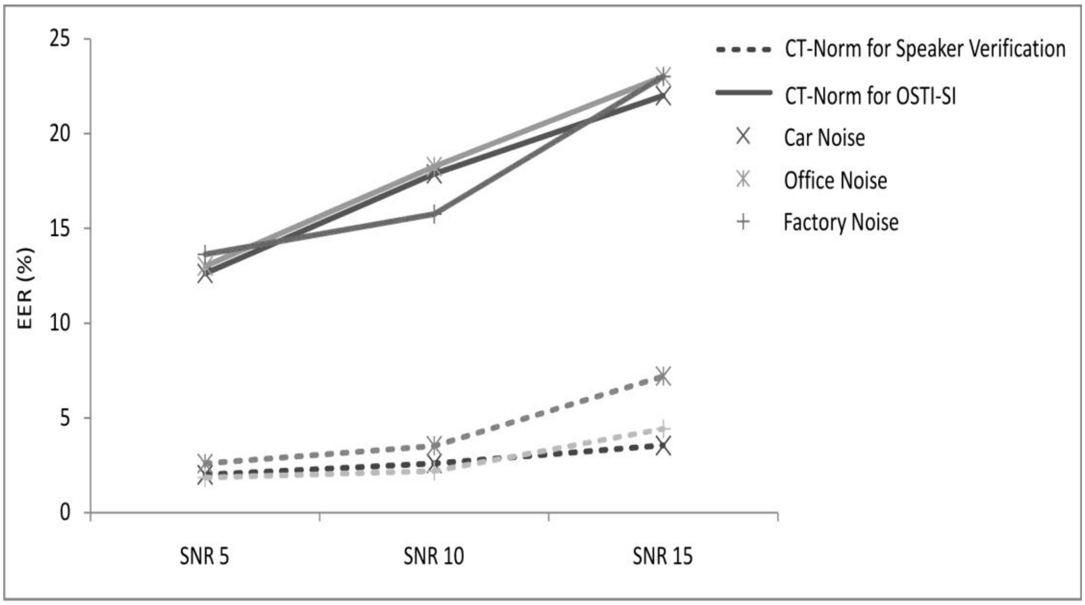


Figure 5.11: Relative effectives of the CT-Norm approach for speaker verification and OSTI-SI.

5.5 Chapter Summary

In this chapter, a modified data-driven parallel model combination (PMC) approach is proposed for tacking the effects of mismatched data conditions (caused by environmental noise) on speaker verification. Based on the experimental results, it is found that the modified PMC, which offers the advantage of computational efficiency when compared to the direct use of PMC with GMM-UBM, cannot be as effective as the latter. The attempt to further improve the verification accuracy of the modified PMC GMM-UBM under such conditions has led to the introduction of CT-norm (condition adjusted T-norm). It is shown experimentally that this normalisation method can considerably enhance the verification accuracy in mismatched noise conditions. Based on investigation carried out using car noise, it is demonstrated that the combination of CT-norm with modified PMC GMM-UBM provides a higher accuracy than that obtainable with the direct PMC GMM-UBM. Moreover, it is shown that the performance of GMM-SVM can also be improved considerably using modified PMC together with CT-norm. However, the added computational cost in this case suggests that such a combined approach is currently unsuitable for most practical applications.

As part of the study, a Bilateral PMC GMM-UBM approach for speaker verification operating in conditions where the training and testing utterances are both contaminated is also proposed and investigated. Based on the outcomes of the investigations, it is shown that, there are no significant advantages to be obtained by using the said approach when compared to the modified PMC GMM-UBM.

For the purpose of completeness, an investigation into the relative effectiveness the modified PMC GMM-UBM with CT-Norm approach for OSTI-SI has also been presented. It has been shown that the performance of OSTI-SI is severely affected when the level of degradation in the test material is different from that in the training utterances. The outcomes of the experimental investigations have clearly demonstrated that in these adverse scenarios, deploying the modified PMC GMM-UBM approach can significantly improve the accuracy of the first stage of the OSTI-SI process (up to 79% for severely degraded data conditions). It is also shown that that the use of CT-Norm with the said approach is of considerable benefit to the verification stage. In this case, the average accuracy improvement relative to conventional GMM-UBM is found to be around 25%.

CHAPTER 6

MULTI-SNR CT-NORM FOR SPEAKER RECOGNITION

Chapter Overview

This chapter presents a new approach to condition-adjusted test-normalisation (CT-Norm) for speaker verification under significant mismatched noise conditions. The experimental investigations are conducted using GMM-UBM and examples of real-world noise. Based on the outcomes, it is demonstrated that the proposed approach effectively outperforms CT-Norm in extreme cases of noisy test data. This is attributed to the greater ability of the proposed method to reduce the mismatch between the training and testing material, and also to the fact that the approach lends itself more effectively to the fast-scoring principles in the GMM-UBM paradigm. Section 6.1 describes the motivations for this study. The proposed method and its characteristics are detailed in Section 6.2. The experimental investigations and an analysis of the results are then presented in Section 6.3.

In Section 6.4, the use of the proposed approach for OSTI-SI is considered, and a new method termed Multi-SNR Fast CT-Norm is introduced to retain its computational efficiency in this case. The experimental investigations are detailed in Section 6.5.

6.1 Motivations for proposed approach

In Chapter 5, it is demonstrated that T-Norm becomes highly effective when the target and the background speaker models are adapted to the noise condition in the test data. Figure 6.1 summarises the approach proposed for this purpose (the technique is referred to as *C*ondition adjusted *T-Norm*: *CT-Norm*). As indicated in the Figure, the method involves first contaminating the speech material for the target and background speakers according to an estimate of noise in the given test utterance. The resulting (contaminated) speech utterances are subsequently used to adapt a clean UBM in order to generate the required speaker models (i.e. target and background). It has been pointed out that although it is more appropriate to use a noise-adjusted UBM, creating this in the test phase is not viable due to the associated increase in computational cost . However, according to the study in the previous chapter, the adverse effects of using a clean UBM become more noticeable with the increased severity of noise contamination in the test utterance.

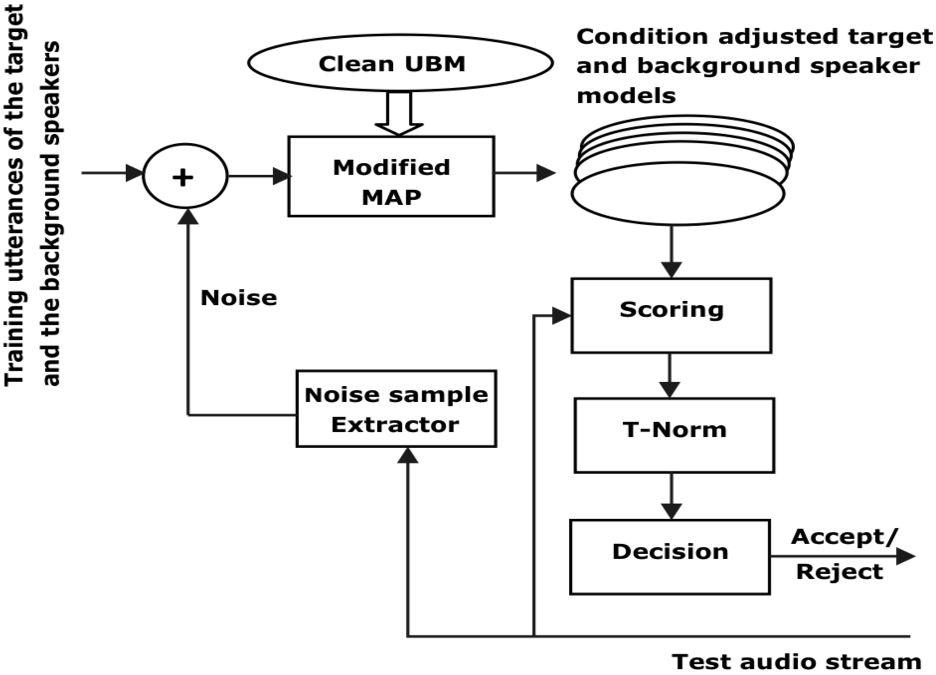

Training utterances of the target and the background speakers
Clean UBM
Condition adjusted target and background speaker models
+
Modified MAP
Noise
Scoring
Noise sample Extractor
T-Norm
Decision
Accept/ Reject
Test audio stream

Figure 6.1: CT-Norm approach.

6.2 Multi-SNR CT-Norm

In order to tackle the problem highlighted in the previous section, a departure from the original approach to CT-Norm is proposed with the view to achieve improvements in the verification accuracy (especially, for severely contaminated test utterances), whilst the computational efficiency in the test phase is largely retained. The idea involves replacing the single clean UBM used in the original method with a set of degraded UBMs. Each such UBM is built by first contaminating the given training utterances using white noise to achieve a specific level of SNR (signal-to-noise-ratio). In the verification phase, first an estimate of the noise in the test utterance is used to contaminate the whole of training material for the target and background speakers. Then, the test utterance is scored against each of the available degraded UBMs. Finally, the UBM which yields the highest likelihood is selected for obtaining adapted target and background speaker models using the degraded reference material resulted in the first step. This reinforces the closeness of the degradation condition in the target and background models to that in the test utterance. This method, which is referred to as Multi-SNR CT-Norm in the remainder of this chapter, is illustrated in Figure 6.2.

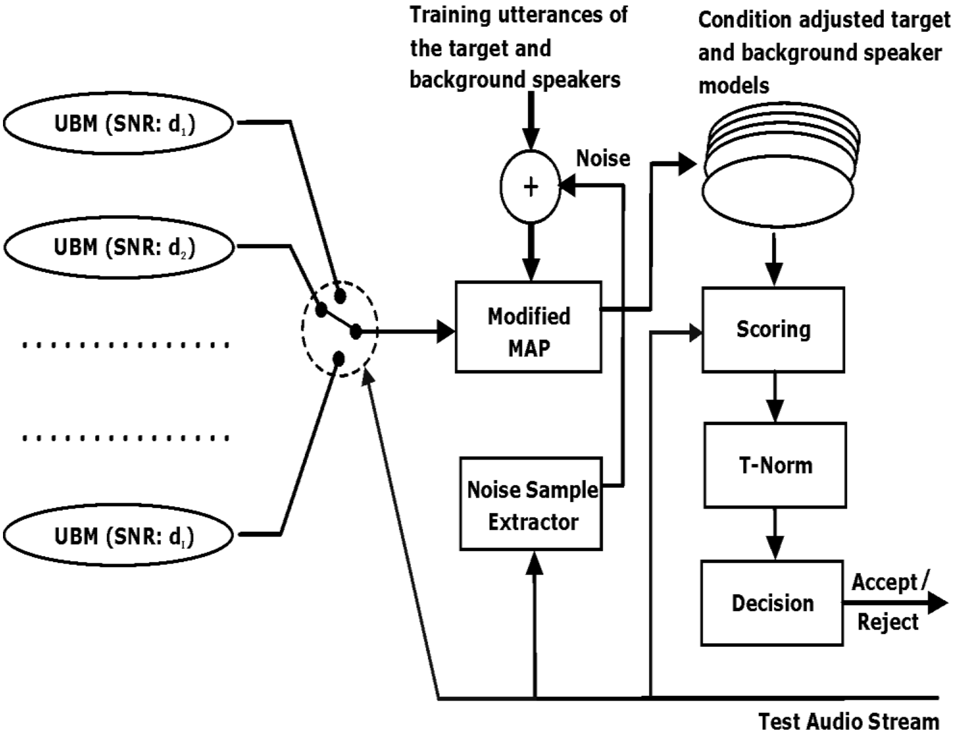


Figure 6.2: Multi-SNR CT-Norm approach

It should be noted that a Multi-SNR GMM approach has previously been proposed in the literature [93]. In that study, each registered speaker is represented by multiple decoupled GMMs, each built (a priori) using training utterances which are contaminated with $1/f^{\alpha}$ noise to achieve different SNRs (note: the noise is defined in terms of its spectrum $1/f^{\alpha}$, where f is the frequency and α is an adjustable parameter which controls the noise spectrum). In the test phase, the degraded target model which best matches the test utterance is chosen for the purpose of verification.

The technique proposed in this chapter operates in the GMM-UBM paradigm, and its novelty is that it attempts to reduce the mismatch between the training condition of the UBM and the test condition. This results in a twofold advantage. First, it facilitates an improved adjustment of the target and background speaker models (which are obtained by UBM adaptation) to the noise condition in the test utterance. Secondly, it matches the *fast-scoring* principles in the GMM-UBM paradigm [12] more closely than the original CT-Norm method, and thereby offers enhanced verification score accuracy, particularly, in the case of more severely contaminated test utterances.

Suppose that the multi-SNR UBMs are represented as $\left\{\boldsymbol{\lambda}_i = \left\{w_{i,m}, \boldsymbol{\mu}_{i,m}, \mathbf{C}_{i,m}\right\}_{m=1,\dots\dots,M}\right\}_{i=1,\dots\dots,I}$ where $w_{i,m}$, $\boldsymbol{\mu}_{i,m}$ and $\mathbf{C}_{i,m}$ are the weight, mean and covariance associated with the m^{th} mixture of the i^{th}-degraded UBM, M is the total number of mixtures in each UBM and I is the total number of UBMs. Additionally, suppose that condition adjusted speaker models are denoted as $\hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_1, \dots, \hat{\boldsymbol{\lambda}}_N$, where $\hat{\boldsymbol{\lambda}}_0$ is the target speaker model and the rest are the background speaker models. In the Multi-SNR *CT-Norm* approach, the verification score is obtained as

$$L_{\text{SV}} = \left\{p\left(\mathbf{Y} \middle| \hat{\boldsymbol{\lambda}}_0\right) - \mu_b(\mathbf{Y})\right\}/\sigma_b(\mathbf{Y}), \tag{6.1}$$

where $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T\}$ is the test vector sequence, and $\mu_b(\mathbf{Y})$ and $\sigma_b(\mathbf{Y})$ are the mean and standard deviation of $p\left(\mathbf{Y} \middle| \hat{\boldsymbol{\lambda}}_1\right), p\left(\mathbf{Y} \middle| \hat{\boldsymbol{\lambda}}_2\right), \dots, p\left(\mathbf{Y} \middle| \hat{\boldsymbol{\lambda}}_N\right)$. Here, $p\left(\mathbf{Y} \middle| \hat{\boldsymbol{\lambda}}_n\right)$ for $n = 0,1\,,2, \dots, N$ are computed in the following manner: (the procedure below is an adaptation of the fast scoring technique proposed in [12].

$$p\left(\mathbf{Y}\middle|\,\widehat{\boldsymbol{\lambda}}_n\right) = \frac{1}{T}\sum\nolimits_{t=1}^{T} logl_{n,t}, \tag{6.2}$$

where

$$l_{n,t} = \sum_{k=1}^{K} w_{\theta,\psi(t,\theta,k)}\mathcal{N}\left(\widehat{\boldsymbol{\mu}}_{n,\theta,\psi(t,\theta,k)}, \mathbf{C}_{\theta,\psi(t,\theta,k)}, \mathbf{y}_t\right), \tag{6.3}$$

$$\theta = \underset{i=1,\ldots,I}{\operatorname{argmax}} \sum_{t=1}^{T} \log \sum_{m=1}^{M} w_{i,m}\mathcal{N}\left(\boldsymbol{\mu}_{i,m}, \mathbf{C}_{i,m}, \mathbf{y}_t\right), \tag{6.4}$$

$$\{\psi(t,\theta,k): 1 \leq k \leq K\} = \ \arg \underset{m=1,\ldots M}{K\max} \left\{w_{\theta,m}\mathcal{N}\left(\boldsymbol{\mu}_{\theta,m}, \mathbf{C}_{\theta,m}, \mathbf{y}_t\right)\right\},\ K \ll M, \tag{6.5}$$

and

$$\widehat{\boldsymbol{\mu}}_{\,n,\theta,m} = \frac{\left\{\sum_{t=1}^{T'(n)} P_{\theta,m}\left(\widehat{\mathbf{x}}_{n,t}\right)\widehat{\mathbf{x}}_{n,t}\ \right\} + R\ \boldsymbol{\mu}_{\theta,m}}{\sum_{t=1}^{T'(n)} P_{\theta,m}\left(\widehat{\mathbf{x}}_{n,t}\right) + R}. \tag{6.6}$$

In equations (6.3) – (6.5), $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C}, \bullet)$ represents a multivariate Gaussian probability density function with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$. In (6.6), $\mathbf{X}_n = \{\widehat{\mathbf{x}}_{n,1}, \widehat{\mathbf{x}}_{n,2}, \ldots, \widehat{\mathbf{x}}_{n,T'(n)}\}$ are the set of contaminated training vectors associated with the n^{th} speaker, R is the relevance factor for the mean statistics [12], and $P_{\theta,m}(\widehat{\mathbf{x}}_{n,t})$ is the probability of $\widehat{\mathbf{x}}_{n,t}$ belonging to the m^{th}-mixture of the chosen UBM (i.e. θ^{th}-UBM).

This probability is estimated in the following manner [12]

$$P_{\theta,m}(\hat{\mathbf{x}}_{n,t}) = \frac{w_{\theta,m}\,\mathcal{N}(\boldsymbol{\mu}_{\theta,m}\,, \mathbf{C}_{\theta,m}, \hat{\mathbf{x}}_{n,t})}{\sum_{m=1}^{M} w_{\theta,m}\,\mathcal{N}(\boldsymbol{\mu}_{\theta,m}, \mathbf{C}_{\theta,m}, \hat{\mathbf{x}}_{n,t})}. \tag{6.7}$$

It is important to note that $\hat{\boldsymbol{\mu}}_{\cdot,\cdot,\cdot}$ (Equation (6.6)) need to be evaluated only when required, and all the computed values of $\hat{\boldsymbol{\mu}}_{\cdot,\cdot,\cdot}$ can be cached for reuse. In other words, there is no need to perform a full target/background speaker model adaptation in the test trial. Instead, the requirement for mean adaptation can be identified and then fulfilled as part of the scoring process to save computation. To be specific, the procedure deployed in the scoring process can be expressed as follows. For each test vector, first determine the top *K* mixture densities through equation (6.5). Then, for each such density, check the cache for the availability of the adapted mean, or the lack of it. In the case of the latter, adapt the corresponding UBM mean using equation (6.6) and place the result in the cache. Using the adapted means available in the cache for the remainder of the scoring process (in the same test trial) can significantly reduce the computational cost. The exact extent of the computational saving achieved in this way varies from trial to trial, as it depends on the acoustic content of the test utterance.

The Multi-SNR *CT-Norm* technique is also well suited for both distributed computing and multi-core processor environments as the intense parts of the calculations can be divided into concurrent tasks. Based on these observations, a computationally efficient realisation of the proposed approach, as shown in Fig.6.3, can be considered.

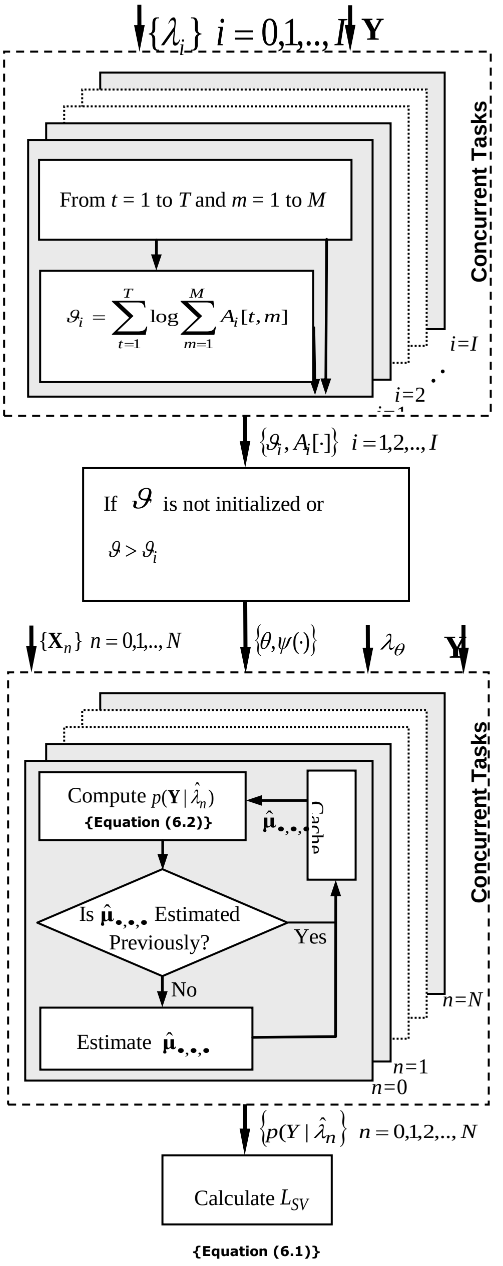

$\{\lambda_i\}\ i = 0,1,..,I$
$\mathbf{Y}$
From t = 1 to T and m = 1 to M
$\vartheta_i = \sum_{t=1}^{T} \log \sum_{m=1}^{M} A_i[t,m]$
Concurrent Tasks
$i=I$
$i=2$
$i=1$
$\{\vartheta_i, A_i[\cdot]\}\ i = 1,2,..,I$
If ϑ is not initialized or
$\vartheta > \vartheta_i$
$\{\mathbf{X}_n\}\ n = 0,1,..,N$
$\{\theta, \psi(\cdot)\}$
λ_θ
$\mathbf{Y}$
Compute $p(\mathbf{Y} \mid \hat{\lambda}_n)$
{Equation (6.2)}
$\hat{\boldsymbol{\mu}}_{\bullet,\bullet,\bullet}$
Cache
Is $\hat{\boldsymbol{\mu}}_{\bullet,\bullet,\bullet}$ Estimated Previously?
Yes
No
Estimate $\hat{\boldsymbol{\mu}}_{\bullet,\bullet,\bullet}$
Concurrent Tasks
$n=N$
$n=1$
$n=0$
$\{p(Y \mid \hat{\lambda}_n\}\ n = 0,1,2,..,N$
Calculate L_{SV}
{Equation (6.1)}

Figure 6.3: Implementation of Multi-SNR CT-Norm approach.

6.3 Experimental Investigations

6.3.1 Speech Data, speaker representation

The speech dataset used for the purpose of the experimental investigations is extracted from the TIMIT database [104]. The set includes 100 registered speakers and 80 unknown speakers, each with 10 utterances. The individual utterances are about 3 seconds long. The training material for each speaker model is based on concatenating five utterances. This setup results in 500 client scores and 129 500 impostor scores. The speech material used for building the UBM consists of ten utterances from each of 200 speakers other than the ones registered or used as unknown speakers. As in the experimental setup described in the previous chapters, it should be noted that the speaker set used for UBM and the sets of registered and unknown speakers are all gender-balanced.

The implementation of CT-Norm (or T-Norm where appropriate) is based on the training utterances from the cohort of speakers available within the set of registered users (i.e. 99 speakers on each occasion).

6.3.2 Experimental Results and Discussions

The previous investigations into CT-Norm [139] have involved both GMM-UBM and GMM-SVM methods. For the purpose of consistency, that study has been based on the use of a UBM of size 128 mixtures for both classifiers. However, it has already been established that the use of a higher order UBM with the GMM-UBM technique can, in general, lead to a higher accuracy in speaker recognition [12]. Knowing that the use of a high-order UBM with CT-Norm increases the computational load considerably, it is necessary to determine the level of accuracy benefit offered by such a UBM in this case. For this purpose, two sets of experiments with CT-Norm are conducted under mismatched noise conditions. The first set involves a UBM of size 128 mixtures whereas the UBM used in the second set of experiments is of 1024 mixtures. Three examples of real-world noise (i.e. car noise, office noise, and factory noise) obtained from the NOISEX 92 [138] and Piper [34] databases are used. As before, for each noise type, the test data is contaminated using a randomly selected segment (with the same duration as the test utterance) of the original noise file to achieve SNRs of 15 dB, 10 dB and 5 dB. The training data and UBMs are based on clean speech. The experimental results for this investigation (and other experiments in this study) are presented in terms of Equal Error Rate (EER) with a 95% confidence interval.

The outcomes of this comparative study are presented in tables 6.1 and 6.2. It is noted that, in general, there are no advantages to be gained in terms of accuracy by using a UBM of 1024 mixtures. Given the computational efficiency offered by using a smaller UBM, it is therefore decided to adopt a UBM of 128 mixtures for the purpose of investigations in this study.

EER (%)				
Clean UBM				
		Test Data		
Noise	*Score normalisation*	*SNR :15dB*	*SNR :10dB*	*SNR: 5dB*
Car	**CT-Norm**	2.00 ± 0.62	2.60 ± 0.71	3.55 ± 0.82
Office	**CT-Norm**	2.60 ± 0.71	3.53 ± 0.82	7.20 ± 1.15
Factory	**CT-Norm**	1.85 ± 0.60	2.20 ± 0.65	4.43 ± 0.92

Table 6.1: EERs in speaker verification (under various mismatched noise conditions) conducted using modified PMC-GMM-UBM with and without CT-Norm and a UBM of order 128.

EER (%)				
Clean UBM				
		Test Data		
Noise	*Score normalisation*	*SNR :15dB*	*SNR :10dB*	*SNR: 5dB*
Car	**CT-Norm**	2.25 ± 0.66	2.66 ± 0.71	3.60 ± 0.83
Office	**CT-Norm**	2.40 ± 0.68	3.80 ± 0.85	6.63 ± 1.11
Factory	**CT-Norm**	2.60 ± 0.71	2.43 ± 0.68	3.61 ± 0.83

Table 6.2: EERs in speaker verification (under various mismatched noise conditions) conducted using modified PMC-GMM-UBM with and without CT-Norm and a UBM of order 1024.

The aim of the next set of experiments in this study is to compare the relative effectiveness of Multi-SNR CT-Norm and CT-Norm. In this setup, the degraded UBMs which are employed in the proposed approach are built by contaminating the allocated training data with Gaussian white noise to achieve SNRs of 15dB, 10dB, 5dB and 0dB. It is important to point out the 5 dB SNR interval used for degraded UBMs is chosen based on the outcome of a preliminary set of investigations showing that a smaller interval (e.g. 1 dB) significantly increases the computational load in the test phase, without offering any relative improvements in accuracy. The estimation of test utterance degradation in the verification phase is based on the use of the first 200 ms of the contaminating noise. This is then used for contaminating the speech material for the target and background speakers. Furthermore, it should be pointed out that in this part of the study, for each noise type, the test data is contaminated to achieve SNRs of 15dB, 13dB, 10dB, 8dB, 5dB and 3dB. The reason for using three additional SNR levels is to evaluate the effectiveness of the proposed method when the contamination level of the test utterance does not exactly match that of one of the stored degraded UBMs, as well as when it does. The experimental results for this investigation are presented in tables 6.3 and 6.4.

EER (%)							
Clean UBM							
		Test Data					
Noise	*Score normalisation*	*SNR : 15dB*	*SNR: 13dB*	*SNR : 10dB*	*SNR: 8dB*	*SNR: 5dB*	*SNR: 3dB*
Car	**CT-Norm**	2.00 ±0.62	2.50 ±0.69	2.60 ±0.71	3.40 ±0.79	3.55 ±0.82	4.20 ±0.89
Office	**CT-Norm**	2.60 ±0.71	2.65 ±0.71	3.53 ±0.82	3.72 ±0.84	7.20 ±1.15	12.20 ±1.46
Factory	**CT-Norm**	1.85 ±0.60	2.10 ±0.64	2.20 ±0.65	4.02 ±0.88	4.43 ±0.92	12.45 ±1.47

Table 6.3: Effectiveness of CT-Norm for speaker verification.

EER (%)							
Multi SNR UBMs							
		Test Data					
Noise	*Score normalisation*	*SNR : 15dB*	*SNR: 13dB*	*SNR : 10dB*	*SNR: 8dB*	*SNR: 5dB*	*SNR: 3dB*
Car	**CT-Norm**	1.98 ±0.62	2.20 ±0.65	2.41 ±0.69	2.60 ±0.71	2.80 ±0.73	3.40 ±0.81
Office	**CT-Norm**	1.60 ±0.56	2.30 ±0.67	2.26 ±0.66	3.10 ±0.77	3.40 ±0.81	6.23 ±1.08
Factory	**CT-Norm**	2.00 ±0.63	2.00 ±0.63	2.29 ±0.67	2.60 ±0.71	3.00 ±0.76	6.13 ±1.07

Table 6.4: Effectiveness of the Multi-SNR CT-Norm approach for speaker verification.

The above results clearly indicate that, whilst the Multi-SNR CT-Norm offers better overall performance, its superiority becomes highly significant for more severely degraded test utterances (i.e. 3dB and 5dB). In these cases, the average relative improvements offered by the proposed method are in excess of 24%, 103 % and 75% for car, office and factory noise respectively. Evidently, the relative improvements achieved are more considerable when the additive noise is of a less stationary nature (e.g. office noise).

The next set of experiments compares the effectiveness of the proposed Multi-SNR with CT-Norm with an appropriate realisation of the Multi-SNR GMM method [93]. It should be noted that, as indicated earlier, the original version of Multi-SNR GMM method is based on the use of decoupled GMMs. Therefore, for the purpose of comparison, a modified version of the said technique is implemented in the GMM-UBM context. This is referred to as Multi-SNR GMM-UBM in the rest of this chapter. In this approach, for the purpose of consistency, the adaptation process is based on the same set of degraded UBMs used in the Multi-SNR CT-Norm. The difference, however, is that in the Multi-SNR GMM-UBM approach, the target and background speaker models are adapted offline using training data contaminated with white noise. Moreover, in the case of this method, since the speaker models are not adjusted to the condition of the test utterance, the experiments conducted here are based on the use of conventional T-Norm (i.e. using the set of background speaker models which are subjected to the same level of degradation as the selected degraded target model). The experimental results for this part of the study are presented in Table 6.5.

EER (%)							
Multi SNR GMM-UBM							
		Test Data					
Noise	*Score normalisation*	*SNR: 15dB*	*SNR: 13dB*	*SNR: 10dB*	*SNR: 8dB*	*SNR: 5dB*	*SNR: 3dB*
Car	**T-Norm**	3.19 ±0.79	3.40 ±0.81	3.60 ±0.83	3.83 ±0.85	5.00 ±0.97	8.00 ±1.21
Office	**T-Norm**	2.20 ±0.66	2.60 ±0.68	2.90 ±0.74	3.32 ±0.80	4.73 ±0.95	10.00 ±1.34
Factory	**T-Norm**	3.00 ±0.81	2.60 ±0.72	3.40 ±0.81	4.20 ±0.89	8.27 ±1.3	16.18 ±1.64

Table 6.5: Effectiveness of the modified Multi-SNR GMM-UBM approach for speaker verification.

Comparing the results in tables 6.4 and 6.5, it can immediately be noticed that, under all the different data conditions considered, Multi-SNR CT-Norm is more effective than Multi-SNR GMM-UBM. These results clearly help to establish the enhanced capabilities of the former approach when dealing with unknown noise in the test stage. Additionally, it is again observed that the superior performance of Multi-SNR CT-Norm is highly significant for more severely contaminated test utterances (i.e. 5dB & 3dB). The relative effectiveness improvements offered by the Multi-SNR CT-norm approach under mismatch conditions are further illustrated through the DET plots in Figure 6.4 (car noise), Figure 6.5 (office noise) and Figure 6.6 (factory noise). In all cases, the SNR for the test data is 3 dB. According to these DET plots, the proposed approach not only helps in reducing the overall EER but also decreases the relative miss probability and false alarm probability across all operating points.

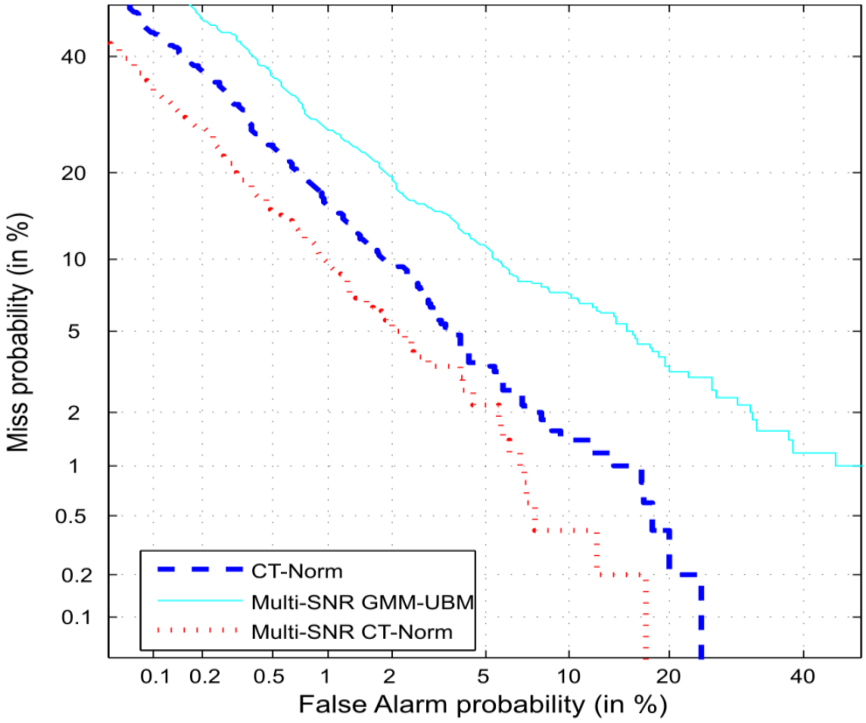


Figure 6.4: Relative verification effectiveness offered by the use of Multi-SNR CT-Norm under mismatched data conditions using car noise.

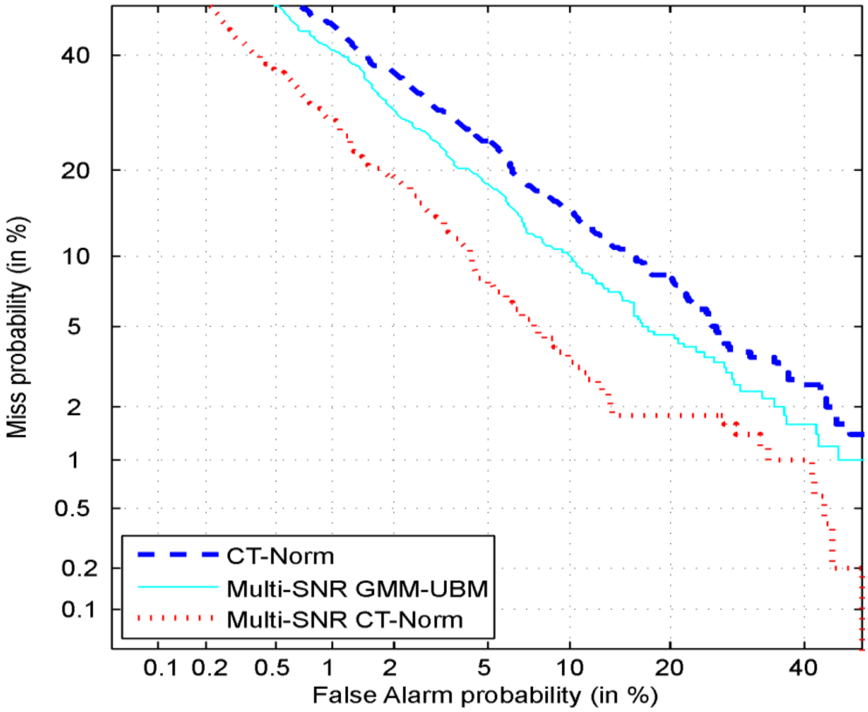


Figure 6.5: Relative verification effectiveness offered by the use of Multi-SNR CT-Norm under mismatched data conditions using office noise.

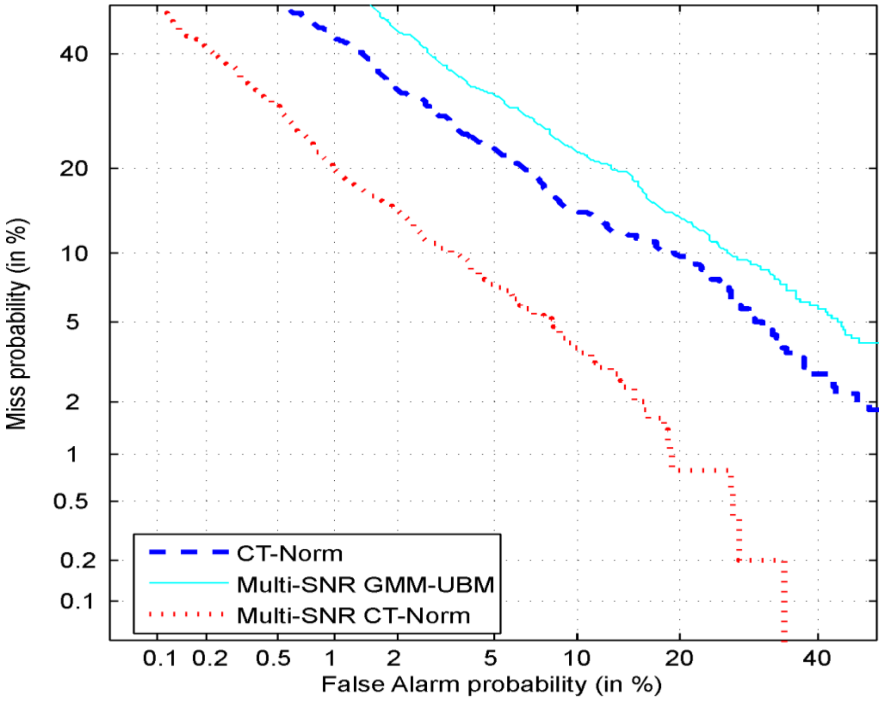


Figure 6.6: Relative verification effectiveness offered by the use of Multi-SNR CT-Norm under mismatched data conditions using factory noise.

6.4 Multi SNR CT-Norm for OSTI-SI

As indicated in Chapter 5, the nature of the problem in the second stage of open-set, text-independent speaker identification (OSTI-SI) makes it more challenging than that of conventional speaker verification. This problem can be re-expressed as a special (but unlikely) scenario in speaker verification in which each impostor targets the speaker model in the system for which (s)he can achieve the highest score [9]. As such, it may not be possible to foresee the effectiveness of the Multi-SNR CT-Norm approach proposed in Section 6.2 based on the results obtained for speaker verification.

Figure 6.7 illustrates the use of Multi-SNR CT-Norm approach for OSTI-SI. As shown in this Figure, in the verification phase, first, the test utterance degradation is used to contaminate the training utterances of all the registered speakers. Then, the test utterance is scored against each of the available degraded UBMs. The UBM which yields the highest likelihood is selected for obtaining noise-adjusted registered speaker models based on the *m*MAP adaptation of the means. As in the speaker verification context, the adaptation process is given by

$$\hat{\boldsymbol{\mu}}_{n,\theta,m} = \frac{\left\{\sum_{t=1}^{T'(n)} P_{\theta,m}\left(\hat{\mathbf{x}}_{n,t}\right) \hat{\mathbf{x}}_{n,t}\right\} + R\,\boldsymbol{\mu}_{\theta,m}}{\sum_{t=1}^{T'(n)} P_{\theta,m}\left(\hat{\mathbf{x}}_{n,t}\right) + R}. \tag{6.8}$$

$$P_{\theta,m}\left(\hat{\mathbf{x}}_{n,t}\right) = \frac{w_{\theta,m}\,\mathcal{N}(\boldsymbol{\mu}_{\theta,m}\,, \mathbf{C}_{\theta,m}, \hat{\mathbf{x}}_{n,t})}{\sum_{m=1}^{M} w_{\theta,m}\,\mathcal{N}(\boldsymbol{\mu}_{\theta,m}, \mathbf{C}_{\theta,m}, \hat{\mathbf{x}}_{n,t})}. \tag{6.9}$$

where all the symbols have the same meaning as in Section 6.2.

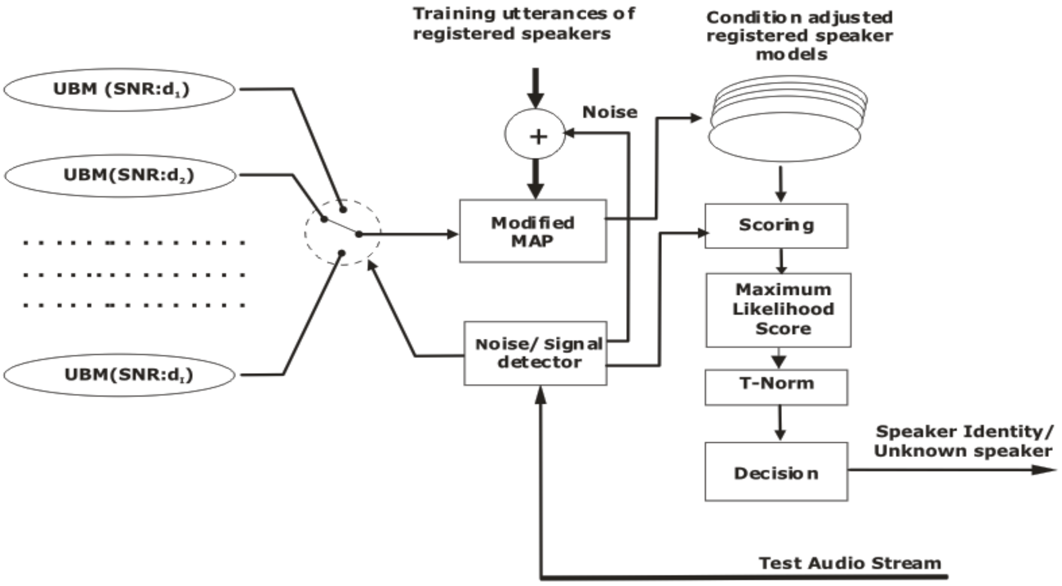


Figure 6.7: Multi-SNR CT-Norm for OSTI-SI

Based on equations (6.8) and (6.9), it is clear that on-the-fly model adaptation, especially in the case of a large number of registered speakers, can become computationally intensive. This is because the probabilistic alignment of each contaminated feature vector with respect to the individual mixtures in the degraded UBM will have to be computed before the new mean statistics can be obtained.

It should be pointed out that this can also become a problem in the speaker verification context, when the number of background speakers used for CT-Norm is increased. However, it has already been shown in the literature that this can be dealt with efficiently and effectively by assigning a specific (and smaller) set of background speaker models to each target speaker based on the *Adaptive T-Norm* (*AT-Norm*) method [126]. In this approach, the size of these speaker-specific sets is chosen to be a fraction of the entire background speaker population, but sufficiently large for computing the T-Norm parameters reliably.

On the other hand, in the open-set speaker identification scenario, the problem is somewhat different because the number of registered speakers cannot be reduced to retain the computational efficiency. To address this problem, a new approach is proposed here for reducing the number of computations in the test phase. This is based on the assumption that the probabilistic alignments of artificially degraded feature vectors and those of their corresponding real-world noise contaminated ones (at a specific SNR) are not significantly different.

Under the above assumption, it may be possible to significantly reduce the computational load involved in building noise adjusted registered models through the UBM adaptation process. This is carried out as follows. First, the training utterances used for building registered speaker models and the UBM are contaminated with Gaussian white noise to achieve a set of equally spaced Signal to Noise Ratios (SNRs) at 15dB, 10dB, 5dB and 0dB. Under each SNR condition, the probabilistic alignments (Equation 6.7) of the degraded feature vectors of each registered speaker (with respect to the mixtures of the correspondingly degraded UBM) are computed and stored offline. In other words, in this scenario, each registered speaker model is associated with four sets of probabilistic alignments.

In the test phase, a short segment (e.g. 200 ms) of the noise contaminating the test utterance is used to degrade the clean reference material from all the registered speakers. The test utterance is then scored against each of the available degraded UBMs. The UBM which yields the highest likelihood score as well as the set of stored probabilistic alignments (for each speaker) which corresponds to the SNR of the above selected UBM are then selected for obtaining noise-adjusted registered speaker models. In this case, however, the probabilistic alignments are not computed online but simply imported from the stored set of alignments to compute the adapted mean vectors (Equation 6.8). As a result, the computational efficiency during the test phase in relation to the Multi SNR CT-Norm approach is enhanced substantially. It is important to point out that, without this approach, it would be required to compute Equation (6.9) about ($T_{\text{ave}} \times (B+1) \times C$) times in each test trial, where T_{ave} is the average number of feature vectors for each given utterance, B is the number of background speaker utterances, and C is the number of mixtures in the UBM. Once the noise adjusted models are obtained, the procedure is identical to the conventional Multi-SNR CT-Norm approach for OSTI-SI described above, i.e. the test utterance is matched against all the registered speaker models and the model that yields the maximum likelihood score is retained. Finally, the score for the speaker model selected as above is subjected to normalisation using T-Norm. The process is summarised in the following algorithm.

Algorithm 1

```
Training stage
for x= 0,5,10,15 do
    Contaminate registered speakers' clean training utterances and clean
    training utterances for training UBM with Gaussian white noise to achieve
    Signal to Noise Ratio (SNR) of x dB.
    Train contaminated UBMs using the Expectation Maximisation (EM) algorithm
    and store.
end
for i=1 to nb_registered speakers do
    Compute and store the probabilistic alignments for each set of
    degraded registered speaker utterances (eq.6.9).
end

Test stage
  Match test utterance against each of the available degraded UBMs
  Select UBM which yields the highest likelihood score
  Contaminate registered speakers' clean training utterances using estimated test
  noise
for all registered_speakers
    Estimate adapted mean vectors, $\hat{\boldsymbol{\mu}}_m$ (eq. 6.8) using the set of stored probabilistic
    alignments which corresponds to the SNR of the above selected UBM
    Compute log-likelihood scores and retain speaker model which yields the
    maximum likelihood score
    end
  Compute CT-Norm on the selected score as in [10]
```

It can be seen in the above algorithm that the proposed computationally efficient approach involves the calculation, storage and use of all the probabilistic alignments for each feature vector with respect to the mixtures in the UBM. In practice, however, this may not be a necessity. Based on the study in [12], it can be argued that in the *m*MAP-based model adaptation, each feature vector of the given utterance exhibit strong alignments only to a small subset of the mixtures in the UBM. This point is further illustrated by the example in Figure 6.8 which shows the probabilistic alignment values (arranged in descending order) for a given feature vector, with respect to the mixtures in a UBM of 128th order. As observed in this figure, the main alignments of the feature vector are only with about 5-6 mixtures in the UBM, which then contribute strongly to the model adaptation.

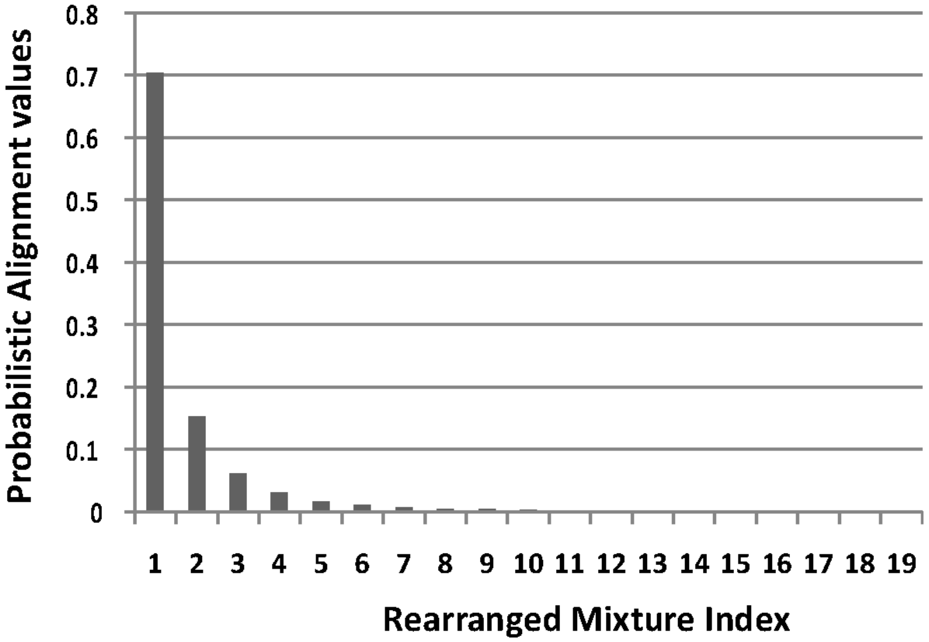
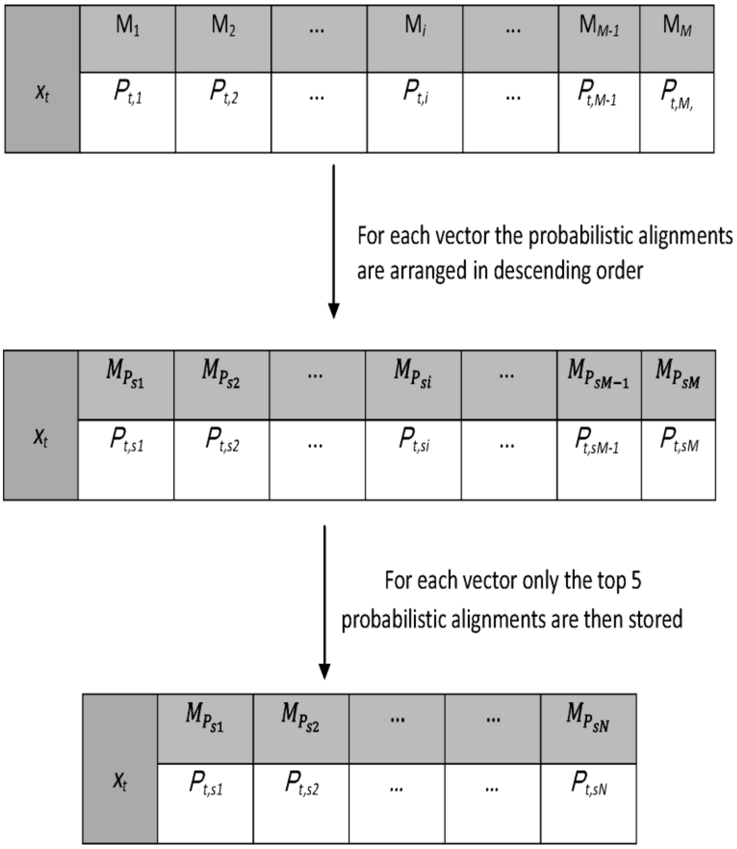


Figure 6.8: Example of the probabilistic alignment values of a feature vector with respect to a 128th order UBM, rearranged in descending order. Only the top 20 (out of 128) mixtures are shown here.

The above is believed to provide a useful basis for further modifying the proposed approach in order to retain the computational efficiency of Multi-SNR CT-Norm. As with the previous approach, for each chosen SNR, the probabilistic alignments of the degraded feature vectors (of each registered speaker) with respect to their corresponding degraded UBM are computed in the training phase. In this case, however, for each feature vector, only the top N alignment values together with the corresponding mixture indices are stored. This is shown in Figure 6.9. It should be noted that in this study, a value of N = 5 which is in agreement with the study in [12], is found to give the optimum performance.

During the verification phase, the procedure is similar to the one described earlier. The only difference is that instead of using the full set of alignments for each feature vector to compute the adapted means, only the top 5 alignments are utilised. As such, the proposed approach, which is hereafter referred to as Multi-SNR *Fast* CT-Norm can considerably enhance the computational efficiency in relation to the Multi-SNR CT-Norm approach during the test phase. It should be noted that term '*Fast*' in this case does not refer to the GMM-UBM fast scoring procedure [12] but to the use of only the top 5 probabilistic alignments in the model adaptation stage.



x_t	M_1	M_2	...	M_i	...	M_{M-1}	M_M
	$P_{t,1}$	$P_{t,2}$	...	$P_{t,i}$	...	$P_{t,M-1}$	$P_{t,M,}$

For each vector the probabilistic alignments are arranged in descending order

x_t	$M_{P_{s1}}$	$M_{P_{s2}}$	...	$M_{P_{si}}$	...	$M_{P_{sM-1}}$	$M_{P_{sM}}$
	$P_{t,s1}$	$P_{t,s2}$	...	$P_{t,si}$	...	$P_{t,sM-1}$	$P_{t,sM}$

For each vector only the top 5 probabilistic alignments are then stored

x_t	$M_{P_{s1}}$	$M_{P_{s2}}$	...	...	$M_{P_{sN}}$
	$P_{t,s1}$	$P_{t,s2}$	...	...	$P_{t,sN}$

Note : $P_{t,s(i)}$ *is the ith sorted alignment probability of the tth feature vector with respect to the corresponding mixture index* $M_{P_{si}}$.

Figure 6.9: Probabilistic alignment selection in the Multi-SNR Fast CT-Norm approach

6.5 Experimental Investigations

6.5.1 Experimental setup

For the sake of comparison and consistency, the speech dataset and speaker representation used for the purpose of the experiments here are identical to those in used in Chapter 5 in the context of OSTI-SI. To be precise, the speech dataset used is extracted from the TIMIT database [104]. 100 registered speakers and 80 unknown speakers are used, each having 10 utterances. Utterances from 200 speakers, other than the ones registered or considered as unknown speakers, are used for training a UBM. In order to facilitate the experimental investigations, in each test trial, the implementation of CT-Norm (or T-Norm where appropriate) is based on the use of the training utterances from the cohort of speakers available within the set of registered users (i.e. 99 speakers on each occasion). As before, three examples of real-world noise (i.e. car noise, office noise, and factory noise), obtained from the NOISEX 92 [138] and Piper [34] databases, are used to degrade the test data; achieving SNRs of 15dB, 13dB, 10dB, 8dB, 5dB and 3dB.

6.5.2 Experimental Results and Discussions

The aim of the first set of experiments is to compare the relative effectiveness of Multi-SNR CT-Norm and CT-Norm in the context of OSTI-SI. The second set of experiments then compares the effectiveness of Multi-SNR CT-Norm with that of the proposed Multi-SNR *Fast* CT-Norm approach. The experimental results for both sets of investigations are presented in terms of Identification Error Rate (IER) and Open-Set Identification Equal Error Rate (OSI-EER) in tables 6.6 and 6.7 respectively.

IER (%)				
Noise	**SNR (dB)**	**CT-Norm**	**Multi-SNR CT-Norm**	**Multi-SNR *Fast* CT-Norm**
Car	15	5.40	5.40	5.80
	13	6.80	5.40	5.80
	10	5.80	6.40	6.20
	8	7.40	7.20	8.20
	5	13.60	11.20	11.80
	3	17.00	13.00	18.40
Office	15	6.60	4.80	5.00
	13	8.00	5.40	5.60
	10	12.60	8.20	8.40
	8	17.40	10.80	11.80
	5	32.20	18.60	18.00
	3	50.80	32.40	41.00
Factory	15	4.80	5.00	6.20
	13	5.80	5.60	6.20
	10	5.80	5.80	7.40
	8	11.40	7.60	10.00
	5	14.00	9.80	18.20
	3	45.20	20.00	39.20

Table 6.6: Relative effectiveness of CT-Norm, Multi-SNR CT-Norm and Multi-SNR *Fast* CT-Norm in terms of IER (%)

It can be observed from the results in Table 6.6 that, in general, the effectiveness of the Multi-SNR CT-Norm approach in terms of IER is better than that obtained for the CT-Norm method. Such an improvement in accuracy is seen to become more considerable in cases where the test utterances are severely contaminated (i.e. 3dB & 5dB) with noise types which are less stationary in nature (e.g. office & factory noise). In these cases, the average relative improvements in IER offered by the Multi-SNR CT approach are in excess of 38% and 39% respectively.

Furthermore, by comparing the results for Multi-SNR CT-Norm and Multi-SNR *Fast* CT-Norm, it is noticed that the performance of the two approaches does not appear to be significantly different when the SNR of the test data is between 15dB-10dB. However, it is seen that when the test data is further degraded (e.g. SNR = 3-8 dB), there is a substantial drop in the accuracy of the latter approach for all the types of noise considered.

OSI-EER (%)				
Noise	**SNR (dB)**	**CT-Norm**	**Multi-SNR CT-Norm**	**Multi-SNR *Fast* CT-Norm**
Car	15	12.62±1.53	12.20±1.49	12.25±1.49
	13	14.62±1.64	12.25±1.51	12.37±1.51
	10	17.87±1.76	14.37±1.62	13.37±1.57
	8	20.25±1.87	14.87±1.65	14.62±1.65
	5	22.00±1.99	17.63±1.81	20.12±1.92
	3	23.37±1.99	18.25±1.85	20.50±1.92
Office	15	13.00±1.56	12.75±1.53	12.13±1.50
	13	13.87±1.56	14.75±1.63	12.63±1.53
	10	18.25±1.85	15.63±1.69	13.62±1.59
	8	19.12±1.93	17.28±1.79	16.25±1.79
	5	23.00±2.29	21.12±2.02	20.75±2.02
	3	26.37±2.78	24.37±2.33	25.75±2.54
Factory	15	13.63±1.58	10.37±1.41	10.75±1.43
	13	13.13±1.58	11.13±1.44	11.50±1.47
	10	15.75±1.68	13.37±1.57	13.37±1.57
	8	17.00±1.78	14.25±1.63	14.25±1.63
	5	23.00±2.29	19.50±1.86	20.00±1.98
	3	26.63±2.61	24.00±2.14	26.63±2.59

Table 6.7: Relative effectiveness of CT-Norm, Multi-SNR CT-Norm and Multi-SNR *Fast* CT-Norm in terms of OSI-EER(%).

It is observed from Table 6.7 that, in terms of OSI-EER, the overall effectiveness of the Multi-SNR CT-Norm approach is again better than that of the original CT-Norm. As before, it is also seen that the superior performance of the said approach becomes considerable when the test data is significantly contaminated with real-world noise. Similar to the investigations with CT-Norm for OSTI-SI, it also appears that the Multi-SNR CT-Norm is more effective in standard speaker verification (Table 6.4) than in the second stage of OSTI-SI. This again highlights the additional challenges in the second stage of OSTI-SI.

Interestingly, taking into account the confidence intervals, it is observed that the difference in the results obtained using full *m*MAP adaptation (i.e. Multi-SNR CT-Norm) and those obtained using the approximated *m*MAP adaptation (i.e. Multi-SNR *Fast* CT-Norm) is not considerable. Put another way, whilst the use of Multi-SNR *Fast* CT-Norm considerably reduces the number of computations during the test phase and the storage requirements when compared to storing the complete set of probabilistic alignments for each speaker, the variation in the level of accuracy is almost negligible. In fact, the enhancement in the computational efficiency offered by said approach (in the adaptation process) is in excess of 95% in relation to the both CT-Norm and Multi-SNR CT-Norm where a full *m*MAP adaptation is carried for each speaker. This level of enhancement is for the case of using a 128-mixture UBM. In fact, the percentage of enhancement in computation efficiency, $V = [(C - N) / C] \times 100$), where C is the number of mixtures in the UBM and N is the number of stored probabilistic mixtures increases linearly with the size of UBM. Hence, depending on the application in which Multi-SNR *Fast* CT-Norm is deployed, it can be argued that the method provides a reasonable trade-off between computational efficiency, storage requirements and accuracy.

6.6 Chapter Summary

An approach to enhancing the effectiveness of CT-Norm for speaker verification under severe noise-mismatched conditions has been investigated. The method, which is termed Multi-SNR CT-Norm, aims to provide a closer adjustment of the target and background speaker models to the noise condition in the test utterance, than that obtainable with the standard CT-Norm method. This is achieved by means of multi-SNR UBMs which also offer the additional advantage of supporting the fast-scoring procedure in the GMM-UBM paradigm. Based on experimental investigations, it has been shown that through the use of the Multi-SNR CT-Norm method, the verification accuracy can be significantly improved for severe noise-mismatched conditions. Additionally, it has been observed that Multi-SNR CT-

Norm offers considerable improvement in the verification accuracy when the noise in the test data is of a more non-stationary nature. For the purpose of completeness the performance of the proposed method is also compared with that of a relevant realisation of the Multi-SNR GMM approach. The results have clearly confirmed that Multi-SNR CT-Norm is more effective than the latter approach for all types and levels of noise considered.

The Multi-SNR CT-Norm approach is then investigated in the open-set, text-independent speaker identification (OSTI-SI) scenario. An analysis of the implementation requirements of the said approach in this context has revealed that, for a large number of registered speakers, it can become computationally intensive. This is mainly attributed to the need in performing a full *m*MAP adaptation in order to obtain noise-compensated models for each registered speaker. To tackle this problem, a Multi-SNR *Fast* CT-Norm approach is proposed. The technique is based on the assumption that the probabilistic alignments of artificially degraded feature vectors and those of their corresponding real-world noise contaminated ones (at a specific SNR) are not significantly different. In addition, for each speaker, only the top probabilistic alignments are stored and utilised during the test stage for model adaptation purposes. This is because each feature vector (from a given training utterance) usually exhibits strong alignments only to a small subset of the mixtures in the UBM. Based on the outcomes of the experimental investigations, it is showed that the overall performance of Multi-SNR CT-Norm is better than that of the original CT-Norm approach. Interestingly, it is observed that whilst the use of Multi-SNR *Fast* CT-Norm reduces the computational cost and storage requirements considerably, the variation in performance when compared to Multi-SNR CT-Norm (in terms of OSI-EER) is almost negligible.

CHAPTER 7

SUMMARY, CONCLUSIONS AND FUTURE WORK

The aim of this research has been to develop effective approaches for voice biometrics (speaker recognition) under mismatched noise conditions. To this end, the research study has been focussed on minimising the noise mismatch between reference speaker models and the given test utterance when using the state-of-the-art speaker recognition approaches. This work has been carried out in the context of both text-independent speaker verification and open-set text-independent speaker identification (OSTI-SI). The summary and overall conclusions of this research together with some suggestions for future work are presented in sections 7.1 and 7.2 respectively.

7.1 Summary and conclusions

For over two decades, the field of automatic speaker recognition has been receiving a great deal of attention from the research community. This is mainly attributed to the need for robust operation under real-world conditions. One of the important facets of the extensive research in this field is that related to the robustness against background noise. The literature review, detailed in Chapter 2, has shown that a highly effective and widely adopted approach for this purpose is that of score normalisation. To date, however, most of the investigations with score normalisation techniques have been carried out using the relevant NIST databases [73, 103, 125]. This means the investigations have been limited in terms of the difference between the levels of noise contamination in the training and testing data; a condition which cannot be considered realistic in many real-world applications.

In order to further study the performance of score normalisation, a set of experimental investigations has been conducted as detailed in Chapter 4. This has involved evaluating the effectiveness of test-normalisation (T-Norm), which is a highly effective and widely deployed score normalisation method, with the state-of-the-art speaker verification techniques (i.e. GMM-UBM and GMM-SVM). The experiments have been conducted under both matched and mismatched noise conditions. Based on the results obtained, it has been observed that T-Norm can provide considerable improvements to the verification accuracy of both approaches under matched data conditions. In general, however, its effectiveness has been found to reduce drastically when the data mismatch, resulting from noise contamination in the test utterance, becomes significant.

In order to tackle this problem, a modified realisation of the parallel model combination (PMC) method for GMM-UBM has been introduced in Chapter 5. This is considered one of the major contributions to knowledge resulting from the study undertaken. As detailed in that chapter, in the case of GMM-UBM, the direct use of PMC involves the computationally expensive (and inefficient) process of rebuilding a UBM with degraded speech material during each test trial. The modified PMC approach involves the use of a fixed UBM built offline using clean speech to enhance the computational efficiency. The problem with this approach, however, is that it reduces the effectiveness of the UBM normalisation technique. This can be attributed to the mismatch which is introduced between the clean UBM and the noise compensated target model which in turn, does not provide an effective means of score normalisation. Thus, in order to maximise the effectiveness of the modified PMC GMM-UBM approach, the concept of *C*ondition adjusted *T-norm* (*CT-norm*) is proposed in Chapter 5. The method involves contaminating the speech material for background speakers (as well as the target speaker) according to an estimate of noise in the given test utterance. The resulting (contaminated) speech utterances are subsequently used to adapt a clean UBM in order to generate the required speaker models. During the matching phase, the degraded test utterance is scored against the condition adjusted speaker models (i.e. target and background). Following this, the normalisation parameters (i.e. mean and variance) are computed (based on the likelihood scores for the background speaker models) and used to perform Test-normalisation (T-Norm).

This approach has been shown to outperform the standard T-norm method under various noise-mismatched conditions. Based on the experimental results, it is observed that the relative improvement achieved for GMM–UBM (under the most severe mismatch condition considered) is in excess of 70%. Moreover, it has been found that, while the accuracy performance of GMM–SVM can also considerably benefit from the use of these techniques, the additional computational cost involved in this case can severely limits the use of such a combined approach in practice.

An investigation into the relative effectiveness of the modified PMC GMM-UBM with *CT-Norm* approach for OSTI-SI has also been presented in Chapter 5. It is argued that the problem in the second stage of OSTI-SI is more challenging than that of the standard speaker verification. This is due to the fact that the requirement in the second stage of OSTI-SI is to discriminate each out-of-set speaker from its best matched speaker in the registered set. Hence, making it unrealistic to fully predict the effectiveness of the proposed approach in the context of OSTI-SI, based on the results obtained for SV. The outcomes of the experimental investigations have clearly demonstrated that under mismatched noise scenarios, deploying the modified PMC GMM-UBM approach can significantly improve the accuracy of the first stage of the OSTI-SI process (up to 79% for severely degraded data conditions). It is also shown that that the use of CT-Norm with the said approach is of considerable benefit to the verification stage. In this case, the average accuracy improvement relative to conventional GMM-UBM is found to be around 25%.

Another key original aspect of this research work is the introduction of an approach to enhancing the effectiveness of CT-Norm for speaker verification under severe noise-mismatched conditions (Chapter 6). This is motivated by the outcomes of the study in Chapter 5 indicating that the adverse effects of using a clean UBM become more significant when the severity of noise contamination in the test utterance increases. To tackle this problem, the proposed method (termed Multi-SNR CT-Norm) aims to provide a closer adjustment of the target and background speaker models to the noise condition in the test utterance, than that obtainable with the standard CT-Norm method. This is achieved by means of a multi-SNR UBM approach which also offers the additional advantage of supporting the fast-scoring procedure in the GMM-UBM paradigm. Based on experimental investigations, it has been shown that through the use of the Multi-SNR CT-Norm method, the verification accuracy can be significantly improved for severe noise-mismatched conditions. Additionally, it has been observed that Multi-SNR CT-Norm offers considerable improvement in the verification accuracy when the noise in the test data is of a more non-stationary nature. For the purpose of completeness, the performance of the proposed method is also compared with that of a relevant realisation of the Multi-SNR GMM approach. The results obtained have clearly confirmed that Multi-SNR CT-Norm is more effective than the latter approach for all types and levels of noises considered.

The proposed Multi-SNR CT-Norm approach is also investigated in the open-set, text-independent speaker identification (OSTI-SI) scenario. An analysis of the implementation requirements of the said approach in this context has revealed that, for a large number of registered speakers, it can become computationally intensive. This is mainly attributed to the need for performing a full *m*MAP adaptation in order to obtain noise-compensated models for each registered speaker. To tackle this problem, a Multi-SNR *Fast* CT-Norm approach is proposed. The technique is based on the assumption that the probabilistic alignments of artificially degraded feature vectors and those of their corresponding real-world noise contaminated ones (at a specific SNR) are not significantly different. Under this assumption, it is shown that the probabilistic alignments of each degraded feature vector (with respect to the mixtures of the correspondingly degraded UBM) can be computed offline and simply imported during each test trial to compute the adapted mean vectors. In addition, for each speaker, only the top probabilistic alignments are utilised for model adaptation purposes. This is because each feature vector (from a given training utterance) usually exhibits strong alignments only to a small subset of the mixtures in the UBM. Based on the outcomes of the experimental investigations, it is demonstrated that the overall performance of Multi-SNR CT-Norm is better than that of the original CT-Norm approach. Interestingly, it is also observed that whilst the use of Multi-SNR *Fast* CT-Norm reduces the computational cost and storage requirements considerably, the variation in performance when compared to Multi-SNR CT-Norm (in terms of OSI-EER) is almost negligible.

7.2 Suggestions for future work

This section briefly discusses some avenues for future research which could be used to extend the work presented in this thesis.

The investigations presented in this thesis can be extended to complement other techniques which have been proposed in the speaker recognition literature to minimise the effects of mismatch data conditions caused by communication channel effects [15, 71, 128, 131, 132, 134, 135]. One such approach, which is used within the SVM framework, is that of Nuisance Attribute Projection (NAP) [71, 131, 132]. This approach has been experimentally investigated in this thesis and found to give good improvements in accuracy when the data is also affected by channel mismatch (i.e. on the NIST SRE database). More recently, a Joint Factor Analysis (JFA) approach has been shown to give promising results when dealing with the adverse effects channel mismatch [15, 128, 135, 141]. This is mainly attributed to the ability of JFA to explicitly model inter-session variability (i.e. channel/intra-speaker variations between the enrolment and test stages). To date, one important limitation of such approaches is that they rely on the availability of large labelled[7] development databases that characterise all the different communication channels which are expected during the test trials in order to be effective [110]. However, given that the JFA methodology is still in the development stages and is constantly being refined to improve its effectiveness, it can be expected that this limitation will also be addressed to enable the deployment of the said approach in practical situations. Further investigations will therefore need to be carried out to develop effective and efficient approaches of combining the methods proposed in this thesis with such emerging methods in order to enhance the overall speaker recognition accuracy under adverse operating conditions.

The investigations carried out in this study have been based on the assumption that relatively short utterances are obtained during each test trial (i.e. about 3 seconds). This is considered to be a reasonable assumption which also adds to the challenge of many speaker recognition applications in practice. Nevertheless, in many real-

[7] Each speaker's development utterances (recorded under different channel conditions) are clearly labelled such that they can be grouped together in the database.

world speaker recognition applications, the user is not constrained to provide short test utterances. Although it has been shown that approaches proposed in this thesis can deal with non-stationary type of noises, in such cases, it may not be appropriate to perform the noise compensation process using only a short estimation of the test noise (obtained in a non-voice segment at the beginning of the utterance). This is because in many operating environments, the characteristics of the noise contamination can vary drastically. Further research is therefore required to develop robust noise estimation approaches, which have the ability of identifying segments within the test utterance where changes occur in the characteristics of the background noise. In such a scenario, the noise contamination process may need to be carried out dynamically such that, whenever the characteristics of the background noise appear to have changed, new condition adjusted speaker models can be trained using the noise estimates. During the matching process, each noisy segments from the test utterance can be matched against the corresponding noise compensated speaker models (of the same noise characteristics) to generate more robust likelihood scores. The final speaker recognition decision can then be based on the score-level fusion of the said likelihood scores. There are various fusion methods such as logistic regression and support vector machines [3, 142, 143] that can be investigated in this context. This is however, just one possible approach which will need to be investigated thoroughly together with other potential techniques before an effective solution can be found.

The work reported in this thesis has been focussed on minimising the effects of mismatched noise conditions on speaker recognition by enhancing the effectiveness of score normalisation approaches. The proposed approaches have been thoroughly investigated on the TIMIT database with artificially added real-world noise. As discussed in Chapter 2, the main reason for using the TIMIT database is that it remains amongst the only widely used and readily available speech corpora which comprises speech utterances recorded under clean conditions. As such, real-world phenomena such as the Lombard effect, [144] which is the involuntary increase in the intensity of one’s voice when speaking in loud noise, have not been considered. Future research should therefore be concentrated on investigating the effectiveness of approaches proposed in this thesis under these conditions. In order to isolate and quantify the adverse effects of such phenomena, it is strongly believed that newer and larger databases for investigating noise contamination under real-world conditions should be collected and made available to the speech research community.

APPENDIX A
MAXIMUM LIKELIHOOD TRAINING

The Maximum Likelihood (ML) based GMM model training process is considered as the process of clustering the speaker's training feature vectors into *C* clusters (or mixtures) within the feature space. It should be pointed out that this process is called 'unsupervised' (also known as the incomplete data problem [40]) due to the fact that the acoustic class of each feature vector is not available *a priori*.

Given a set of *T* training vectors, $\mathbf{O}=\{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$, the aim of the ML estimation is to find the model parameters w_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i^2$, which maximise the likelihood function of the GMM. This is given as

$$p(\mathbf{O}|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\lambda) \qquad (\mathrm{A}.1)$$

As mentioned earlier, diagonal nodal covariance matrices are used during the training process. This implies that only the variances parameters are utilised[8]. Hence, here, $\boldsymbol{\sigma}_i^2$ is the variance vector for the c^{th} Gaussian component. Maximising the above function involves differentiating it with respect to the parameter set $\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2\}$, for *i*=1,…..,*C* and equating to zero as follows [40]

$$\frac{\partial\{p(\mathbf{O}|\lambda)\}}{\partial\lambda} = 0 \qquad (\mathrm{A}.2)$$

The problem, however, is that obtaining a closed-form solution for evaluating the above expression is difficult to obtain. To overcome this issue, an iterative process based on the Expectation-Maximisation (EM) algorithm is deployed [11]. This process guarantees a monotonic increase in the likelihood function. In other words after each iteration, the probability of the estimated model in relation to the distribution of the training feature vectors is expected to increase. The EM algorithm is a two-step process. This consists of the expectation step (or E-step)

[8] The diagonal elements of a covariance matrix are the variances of the vector.

where a new estimate of the parameters is computed based on the initial (or current) parameter estimates and the given training data. In this step, since the acoustic class correspondence of the feature vectors is unknown, it is estimated using the *a posteriori* probability for acoustic class, i, given the observation, $\boldsymbol{o}_t$

$$p(i|\boldsymbol{o}_t,\lambda) = \frac{w_i p_i(\boldsymbol{o}_t)}{p(\boldsymbol{o}_t|\lambda)} = \frac{w_i p_i(\boldsymbol{o}_t)}{\sum_{k=1}^{M} w_k\, p_k(\boldsymbol{o}_t)} \qquad (\mathrm{A}.3)$$

Based on the above *a posteriori* probability, the GMM parameters for each mixture component can be estimated as follows [118]

$$\widehat{w}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\boldsymbol{o}_t,\lambda) \qquad (\mathrm{A}.4)$$

$$\widehat{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T} p(i|\boldsymbol{o}_t,\lambda)\boldsymbol{o}_t}{\sum_{t=1}^{T} p(i|\boldsymbol{o}_t,\lambda)} \qquad (\mathrm{A}.5)$$

$$\widehat{\boldsymbol{\sigma}}_i^2 = \frac{\sum_{t=1}^{T} p(i|\boldsymbol{o}_t,\lambda)\boldsymbol{o}_t^2 - \widehat{\boldsymbol{\mu}}_i^2}{\sum_{t=1}^{T} p(i|\boldsymbol{o}_t,\lambda)} \qquad , \qquad (\mathrm{A}.6)$$

where the notation ^ represents an estimated parameter.

In the M-step, the model parameters are updated with those computed during the E-step, and the iteration is repeated until the likelihood function converges (i.e. $p(\mathbf{O}|\lambda^{k+1}) \approx p(\mathbf{O}|\lambda^{k})$.

During the implementation of the ML algorithm, there are three important factors that need to be taken into consideration. These include the approach chosen to obtain the initial estimates of the GMM, the number of mixtures in the GMM and variance limiting:

- There are several ways in which the initial estimate of the GMM parameters can be obtained. This can be done randomly or by some form of clustering of the training data such as the VQ [118], LBG [39] or distortion driven cluster splitting (DDCS) [31]. The latter approach has been reported to provide faster convergence speed and higher model likelihood during the ML procedure and is therefore adopted for the purpose of this work.

- The determination of the optimum number of mixtures is very important for the performance of the speaker recognition system. Choosing a GMM with a limited number of mixtures can produce speaker models which do not accurately model the inter-speaker characteristics from the training data. Conversely, choosing too many mixtures (especially when there is limited training material) can result in over fitting of the training data .Hence, the speaker GMM loses the ability to generalise to unseen data. In general, it is found that the best trade-off is that the model order M should not exceed $\sim T/100$ where T is the number of training vectors [145]. Obviously, this rule of thumb only applies when there is a sufficiently large number of training vectors. When the number of training vectors is limited, it appears from the study in [118], that the minimum number of mixtures to adequately model speaker voices appears should be 16.

- It is also observed that in some cases, the variances of the mixture densities which are estimated during the ML procedure can become very small in magnitude (negligible). This can result in a singularity of the likelihood function. To avoid this problem, variance limiting can be imposed during the training process [118]. In this work, a value of 0.01 is used. This value has been shown to be dependent on the type of speech cepstral features used [11].

APPENDIX B

SUPPORT VECTOR MACHINES: FUNDAMENTAL CONCEPTS

The discussion about Support Vector Machines (SVMs) in this appendix starts with a trivial linearly separable example. It is then shown in the subsequent sections how the same principle can be extended for more complex non-linearly non-separable problems. It should be noted that this discussion closely follows the tutorial in [59] and the introduction to SVMs in [58].

B.1 Linearly separable case

Figure B.1 illustrates a linearly separable example with training instances from two classes in a two-dimensional space. In this example, the dark line represents a separating hyperplane[9] which divides the space into two distinct classes. This is commonly given as

$$\begin{aligned} f(\mathbf{x}) &= \langle \mathbf{w}.\mathbf{x} \rangle + b \\ &= \sum_{i=1}^{\ell} w_i \mathbf{x}_i + b \quad , \end{aligned} \tag{B.1}$$

where $\mathbf{x} = (\mathrm{x}_1 ,\ldots\ldots, \mathrm{x}_\ell)$ are the training data points and ℓ is the number of training instances. $\langle \cdot . \cdot \rangle$ denotes an inner product. The vector **w** defines a region perpendicular (normal) to the hyperplane while varying the value of b moves the hyperplane parallel to itself. These two quantities are usually referred to as the weight and the bias respectively. Hence, $f(\mathbf{x}) = 0$, when the separating hyperplane correctly separates the two classes without errors.

[9] In this two-dimensional scenario the hyperplane is simply a line

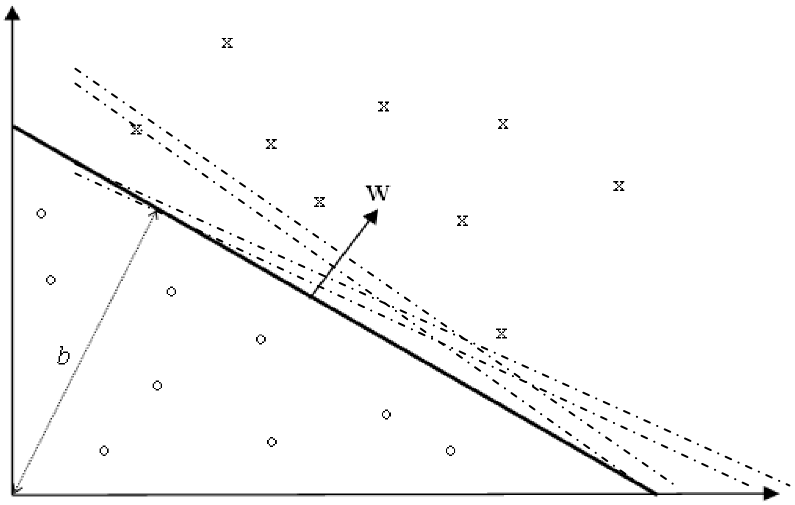


Figure B.1: Illustration of a separating hyperplane for a two-dimensional training set

However, it can immediately be seen from Figure B.1 that there are an infinite number of separating hyperplanes which can be found (few shown in dotted lines), all of which have zero error. An intuitive choice for the best decision boundary is therefore the hyperplane with the maximum margin. In other words, it is the hyperplane which is exactly half way between the two classes. In order to maximise the margin, two parallel lines to the hyperplane, which also separate the two classes without errors, should be considered. This is shown in Figure B.2. The idea behind this approach is to keep the lines parallel to each other while allowing them to rotate and move as far apart as possible without (either line) making an error. The chosen boundary is then the line that splits the margin into half.

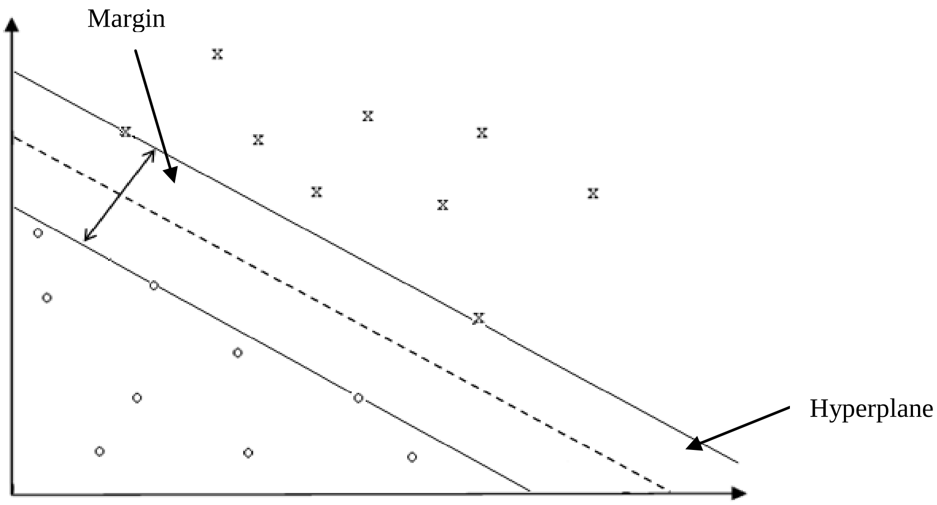


Figure B.2: Margin maximisation

As shown in Figure B.3, it can be assumed that any training instances which lie on the margin boundary (H1) will be +1 while those lying on the margin boundary (H2) will be -1. This is given by the following inequality

$$\langle \mathbf{w}.\mathbf{x}_{\mathrm{i}} \rangle + \mathrm{b} \geq +1 \text{ when } y_i = +1 \qquad (\mathrm{B}.2)$$

$$\langle \mathbf{w}.\mathbf{x}_{\mathrm{i}} \rangle + \mathrm{b} \leq -1 \text{ when } y_i = -1 \qquad , \qquad (\mathrm{B}.3)$$

where $\mathrm{y_i}$ is a label which corresponds to the class of $\mathrm{x_i}$ and $y_i \in \{-1,+1\}$

Equations (B.2) and (B.3) can be combined to give the following inequality constraint

$$\mathrm{y_i}(\langle \mathbf{x}_{\mathrm{i}}.\mathbf{w} \rangle + b) \geq 1 \qquad \text{for } i = 1, \ldots\ldots, \ell \qquad (\mathrm{B}.4)$$

In order to find the perpendicular distance H1 and H2, the distance between H and H1 is first computed. This is given by[10]

$$\frac{|\mathbf{w}.\mathbf{x} + \mathrm{b}|}{\sqrt{\langle \mathbf{w}.\mathbf{w} \rangle}} = \frac{1}{\sqrt{\langle \mathbf{w}.\mathbf{w} \rangle}} = \frac{1}{||\mathbf{w}||} \ , \qquad (\mathrm{B}.5)$$

where ||.|| represents the Euclidean norm.

The margin width (distance between H1 and H2) is then given by:

$$\frac{1}{||\mathbf{w}||} + \frac{1}{||\mathbf{w}||} = \frac{2}{||\mathbf{w}||} \qquad (\mathrm{B}.6)$$

[10] Recall that the distance from a point (x_0,y_0) to a line Ax+By+c=0, is equal to |Ax_0 +B y_0 + c| / sqrt (A^2+ B^2)

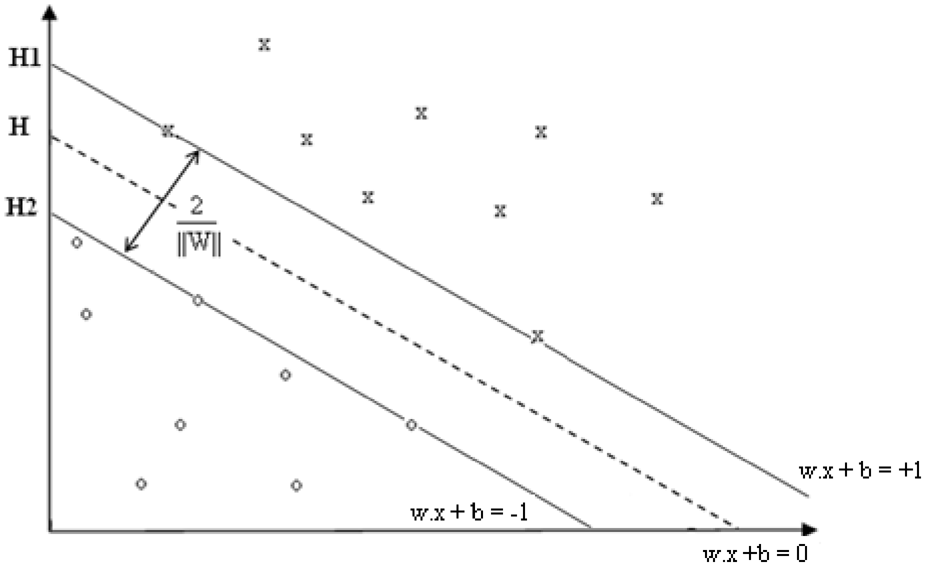


Figure B.3: Calculating the margin size

It is seen from that above that the decision boundary can be found by maximising Equation (B.6). Such a process is also equivalent to minimising the reciprocal of Equation (B.6). In other words, the margin can be maximised by minimising ${||\mathbf{w}||^2}/{2}$ while ensuring that there are no data points between H1 and H2. This is known as an optimisation problem and it is subject to the inequality constraints given in (B.4). It should also be noted that $||\mathbf{w}||^2$ instead of $||\mathbf{w}||$ to eliminate the square root function (which is an increasing function) without affecting the solution. This results in a quadratic programming problem (the objective function has quadratic terms while being constrained by linear inequalities) which is given by

$$\min_{\mathbf{w},b} \frac{||\mathbf{w}||^2}{2}$$

$$\text{subject to} \qquad \mathrm{y_i}(\langle \mathbf{x}_\mathrm{i}.\mathbf{w}\rangle + b) \geq 1 \qquad \text{for } i = 1, \ldots\ldots, \ell \qquad (\mathrm{B}.7)$$

Based on optimisation theory, the above problem which is in its primal form can be reformulated in its dual form by using Lagrangian multiplier. This is given as

$$\max_{\alpha\geq 0}\left(\min_{\mathbf{w},b} L\ (\mathbf{w},b,\alpha)\right) \tag{B.8}$$

Equation (B.4) can then be rewritten by multiplying the constraint equations using positive Lagrange multipliers and subtracting it from the objective function as follows

$$L\ (\mathbf{w},b,\alpha) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{\ell} \alpha_i[y_i(\langle \mathbf{x}_\mathrm{i}.\mathbf{w}\rangle + \mathrm{b}) - 1] \qquad , \tag{B.9}$$

where $\alpha_i, i = 1, \dots \dots \dots \dots, \ell$ are the positive Lagrange multipliers.

In order to simplify Equation (B.9), α can be assumed to be fixed while minimising with respect to **w** and b. This can be rewritten as

$$\min_{\mathbf{w},b} L\ (\mathbf{w},b,\alpha) = \begin{cases} -\infty & \text{if } \sum_{\mathrm{i}=1}^{\ell} \alpha_i\, y_i \neq 0 \\ \min_{\mathbf{w}} \frac{1}{2}\left|\left|\mathbf{w}\right|\right|^2 - \sum_{i=1}^{\ell} \alpha_i[y_i\langle \mathbf{x}_\mathrm{i}.\mathbf{w}\rangle - 1] & \text{if } \sum_{\mathrm{i}=1}^{\ell} \alpha_i\, y_i = 0 \end{cases} \tag{B.10}$$

From equation (B.10), it can be seen that when $\sum_{\mathrm{i}=1}^{\ell} \alpha_i\, y_i \neq 0$ the objective function is $-\infty$. Based on equation (B.8), it can immediately be seen that this case is not helpful if the overall aim is to maximise the function for $\alpha \geq 0$. On the other hand, when $\sum_{\mathrm{i}=1}^{\ell} \alpha_i\, y_i = 0$ it can be seen that the new objective function does not contain b. This is an interesting property which allows the said function to be minimised with respect to **w** only. To achieve this, the partial derivative of **w** is set to zero, such that

$$\frac{\partial}{\partial \mathbf{w}} L\ (\mathbf{w},b,\alpha) = 0 \tag{B.11}$$

This gives

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \qquad \text{(B.12)}$$

The above equation can now be substituted in Equation (B.9) to obtain the dual optimisation problem

$$L\ (\mathbf{w}, b, \alpha) = \frac{1}{2}\sum\nolimits_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_\mathrm{i} . \mathbf{x}_j \rangle$$
$$-\sum\nolimits_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_\mathrm{i} . \mathbf{x}_j \rangle + \sum\nolimits_{i=1}^{\ell} \alpha_i$$

$$= \sum\nolimits_{i=1}^{\ell} \alpha_i - \frac{1}{2}\sum\nolimits_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_\mathrm{i} . \mathbf{x}_j \rangle \quad , \qquad \text{(B.13)}$$

subject to $\alpha_i \geq 0$ for all i and $\sum_{i=1}^{\ell} \alpha_i y_i = 0$. $\mathbf{x}_\mathrm{i}$ for $\mathrm{i} = 1 \ldots\ldots\ldots, \ell$ and $\mathbf{x}_j$ for $\mathrm{j} = 1 \ldots\ldots\ldots, \ell$ are the training instances while $y_i, y_j, \alpha_i, \alpha_j$ are their corresponding class labels and Lagrange multipliers respectively.

There are a few important observations which can be made from Equation B.13. These can also be seen as the motivations behind the use of the Lagrangian dual for solving the original problem. First, it is seen that the constraints in Equation B.4 are replaced by constraints on the Lagrange multipliers instead. Second, it is seen that the optimisation problem in Equation B.13 is formulated only in terms of α to obtain **w**. Both these properties make the Lagrangian dual easier to handle and solve. Finally, it can also be noticed that the training points appear only in the form of inner products. This, as will be seen later, is a very important factor which allows the concept of the maximum margin linear classifier to be used in non-linear scenarios. It should also be pointed out that the value of b does not appear in the dual problem (Equation (B.13)) and must therefore be computed from the primal equation once **w** has been computed using standard techniques [146]. In the solution, all training points for which $\alpha_i > 0$ are called support vectors and lie on one of the hyperplanes H1 or H2 in Figure B.3, while all other training points have $\alpha_i = 0$. The support vectors therefore lie closest to the decision boundary and they are the critical elements of the training set. The other points have no influence on the final solution. Hence, if they were removed or moved around (but did not cross H1 or H2) and the training was repeated, the same separating hyperplane would be found.

The decision boundary $f(\mathbf{x})$ can also be reformulated by substituting (B.12) into (B.1) to give

$$f(\mathbf{x}) = \sum_{i=1}^{N_{SV}} y_i \alpha_i \langle \mathbf{x_i}.\mathbf{x} \rangle + b \qquad (\mathrm{B}.14)$$

where, $\mathbf{x}$ is the vector to classify, $\mathbf{x_i}$ are the support vectors obtained during the training stage and N_{SV} is the total number of support vectors. As before, the equation is constrained by $\sum_{i=1}^{\ell} \alpha_i y_i = 0$ and $\alpha_i > 0$. All the other symbols have the same meaning as in the above equations.

B.2 Linearly Non-separable

The linearly separable maximal margin classifier, discussed in the previous section provides the fundamental concepts of SVMs. In real-world applications, however, it is very unlikely that the data will be linearly separable in the input space [58]. Thus, in order to overcome this problem while using the same underlying concepts as before, the linear constraints in equations (B.2) and (B.3) need to be relaxed. This approach, which is shown in Figure B.4, allows more training points are allowed to lie within the margin (or even be misclassified) during the optimisation stage rather than relying only on those which lie closest to the boundary. This is commonly known as a soft margin. This is done by introducing slack variables ξ_i ,i = 1,…, ℓ in the original constraints, which then becomes

$$\langle \mathbf{w}.\mathbf{x}_{\mathrm{i}} \rangle + \mathrm{b} \geq +1 - \xi_{\mathrm{i}} \text{ when } y_i = +1 \qquad (\mathrm{B}.15)$$

$$\langle \mathbf{w}.\mathbf{x}_{\mathrm{i}} \rangle + \mathrm{b} \leq -1 + \xi_{\mathrm{i}} \text{ when } y_i = -1 \quad , \qquad (\mathrm{B}.16)$$

where $\xi_i \geq 0 \;\; \forall\, i$

As before, the above equations can be combined to give the following constraint

$$\mathrm{y_i}(\langle \mathbf{x}_\mathrm{i}.\mathbf{w}\rangle + b) \geq \ 1 - \xi_\mathrm{i} \qquad \text{(B. 17)}$$

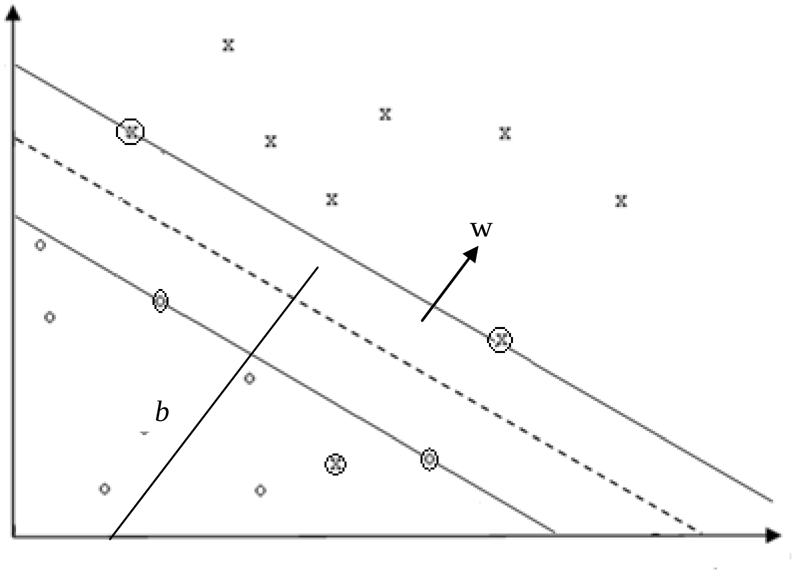


Figure B.4: Linear separating hyperplanes for the non-separable case.

It can be seen from Equation (B.17), that for an error to occur, ξ_i must be greater than 1. The resulting primal problem then becomes

$$\min_{\mathbf{w},b} \qquad \frac{||\mathbf{w}||^2}{2} + \quad \mathrm{C} \ \left(\sum_i \xi_\mathrm{i}\right)$$

$$\text{subject to} \qquad \mathrm{y_i}(\langle \mathbf{x}_\mathrm{i}.\mathbf{w}\rangle + b) \geq \ 1 \qquad \text{for } \ i = 1, \ldots\ldots, \ell \ , \qquad \text{(B. 18)}$$

where the parameter C allows the user to trade off training errors vs. model complexity.

Similarly to the linearly separable case discussed above, Equation (B.18) can be more easily solved by using the Lagrangian dual which can be simplified to give

$$L\ (\mathbf{w}, b, \alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_\mathrm{i} . \mathbf{x}_j \rangle \quad , \qquad (\mathrm{B}.19)$$

subject to

$0 \leq \alpha_i \leq \mathrm{C}$

$\sum_{i=1}^{\ell} \alpha_i y_i = 0$

It can immediately be seen that the above solution is equivalent to the optimal hyperplane for the linearly separable case in (B.13). The only exception in this case however, is that the Lagrange multipliers, α_i , are upper bounded by C. Thus, small value for *C* will increase the number of training errors, while a large *C* will lead to a similar behaviour to that of a hard-margin SVM [147]. This is because α_i is now upper bounded by C and choosing a very large of C (e.g. infinity) will lead to the original constraints of $\alpha_i \geq 0$ of the hard margin.

B.3 Non-linear SVM

The discussion in the two previous sub-sections has been restricted to finding a linear separating boundary (hard margin or soft margin) in the input space. For more complex cases where the decision boundary is not a linear function of the input data, the above methods can be generalised such that a non-linear relationship can be found using a linear machine. This can be achieved by computing a fixed non-linear mapping of the input space to obtain a higher dimensional feature space, in which a linear boundary can be used.

In order to understand this concept clearly, the example illustrated in Figure B.5 is considered. In this case, the input vectors are in a two dimensional space. This is often referred to as the input space. It can be clearly seen that a linear boundary cannot be found in this space. Thus, a non-linear transformation can be applied such that the data is mapped to a three dimensional space, known as the feature space. In other words, a vector **x** which comprises of two data points x_1 and x_2 can be mapped such that

$$\Phi(\mathbf{x}) = \begin{pmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{pmatrix} \quad , \tag{B.20}$$

where $\Phi(\cdot)$ is a mapping function such that Φ: $\mathrm{R}^d \rightarrow H$. R is the input space, d is the dimension of the input space (2 in this case) and R is the feature space.

In this scenario, once the data is mapped to a three-dimensional space, a linear boundary can be found using the same approach as in the previous subsections. However, in practice, explicit knowledge of the dimensionality of the feature space is not a necessity. This is because, as seen in equations (B.13), the objective function is formulated such that the input vectors only appear as inner products pairs. As a result, it might be possible to directly compute the inner product of the vectors in the feature space in terms of the vectors in the input space. This concept can be easily understood by considering the inner product (in the feature space) of two training vectors based on the previous example. This is given as

$$\langle \Phi(\mathbf{x}). \Phi(\mathbf{y}) \rangle = \begin{pmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{pmatrix} . \begin{pmatrix} y_1 \\ y_2 \\ y_1 y_2 \end{pmatrix} = \langle \mathbf{x}. \mathbf{y} \rangle + \; x_1 x_2 y_1 y_2 \tag{B.21}$$

Based on the above equation, it can immediately be seen that for this example, the inner product in the feature space is equivalent to sum of the inner product in the input space with another term (also defined in terms of the input vectors). This can be written in terms of a kernel function such that

$$K(\mathbf{x}, \mathbf{y}) = \; \langle \mathbf{x}. \mathbf{y} \rangle + \; x_1 x_2 y_1 y_2 \tag{B.22}$$

Equation (B.22) therefore implies that the inner products in the objective function can be replaced with a kernel function without explicit knowledge about the non-linear mapping between the input space and the feature space. Mathematically, this is given as

$$K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}). \Phi(\mathbf{y}) \rangle \tag{B.23}$$

This is usually known as the ‘kernel trick’ and provides significantly improvement in the efficiency of the optimisation algorithm. Thus, given a valid kernel function, the SVM output function (decision boundary) for non-linearly separable data is given as

$$f(\mathbf{x}) = \textstyle\sum_{i=1}^{N_{SV}} y_i\alpha_i\langle\Phi(\mathbf{x}_\mathrm{i}).\Phi(\mathbf{x})\rangle \; + \; b$$

$$= \textstyle\sum_{i=1}^{N_{SV}} y_i\alpha_i K(\mathbf{x}_\mathrm{i},\mathbf{x}) + \; b \;, \qquad \text{(B.24)}$$

where all the symbols have the same meaning as in the above equations.

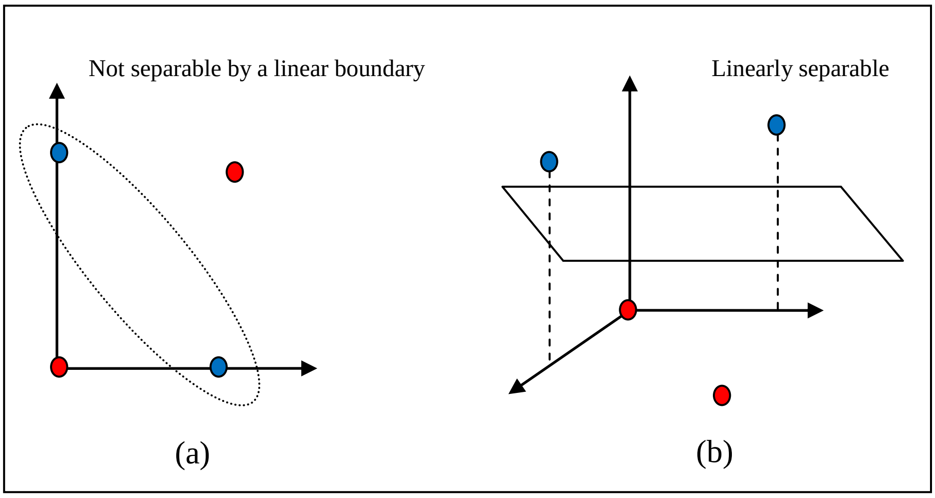


Figure B.5: Illustration of transformation to a higher dimensional space to obtain a linear boundary [147]

Examples of commonly used kernels are:

1. Linear kernel : $K(\mathbf{x},\mathbf{y}) = \langle\mathbf{x}.\mathbf{y}\rangle$
2. Quadratic kernel : $K(\mathbf{x},\mathbf{y}) = (\langle\mathbf{x}.\mathbf{y}\rangle + 1)^p$, in this case p =2 but it can also be increased to obtain a polynomial kernel.
3. Gaussian Radial Basis Function: $K(\mathbf{x},\mathbf{y}) = \exp\left(\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}\right)$ where σ is the standard deviation which defines the kernel width.

REFERENCES

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

[2] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," *IEEE Transactions on Information Forensics Security,* vol. 1, pp. 125-143, 2006.

[3] F. Alsaade, A. M. Ariyaeeinia, A. S. Malegaonkar, M. Pawlewski, and S. G. Pillay, "Enhancement of Multimodal Biometric segregation using Unconstrained Cohort Normalisation," *Pattern Recognition-Special Issue on Multimodal Biometrics,* vol. 41, pp. 814-820, 2008.

[4] A. K. Arun and A. Ross, "Multimodal Biometrics: An Overview," *Proceedings of the 12th European Signal Processing Conference (EUSIPCO)*, pp. 1221-1224, 2004.

[5] F. Bimbot, J. F. Bonastre, et. al., "A Tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing,* vol. 4, pp. 430-451, 2004.

[6] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE,* vol. 85, pp. 1437-1462, 1997.

[7] J. M. Naik, "Speaker Verification: A Tutorial," *IEEE Communications Magazine,* pp. pp 42-48, 1990.

[8] H. Gish and M. Schmidt, "Text-independent speaker identification," *Signal Processing Magazine, IEEE,* vol. 11, pp. 18-32, 1994.

[9] A. Ariyaeeinia, J. Fortuna, P. Sivakumaran, and A. Malegaonkar, "Verification effectiveness in open-set speaker identification," *IEE Vision, Image and Signal Processing,* vol. 153, pp. 618-624, October 2006.

[10] S. G. Pillay, A. Ariyaeeinia, P. Sivakumaran, and M. Pawlewski, "Open Set Speaker Identification under mismatch conditions," in *Interspeech 2009*, Brighton, United Kingdom, pp. 2347-2350, 2009.

[11] D. A. Reynolds, "A Gaussian Mixture Modelling Approach to text-independent speaker identification," PhD Thesis: Georgia Institute of Technology, 1992.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[13] 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 and Language*, vol. 20, 1-2, pp. 210-229, 2006.

[14] R. Dehak, N. Dehak, P. Kenny, and P. Dumouchel, "Linear and non linear kernel GMM Supervector machines for speaker verification," in *Proceedings of the International Conference on Spoken Language Processing, (Interspeech'07)*, 2007.

[15] 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.

[16] H. Gish, "Robust discrimination in automatic speaker identification," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'90),* vol. 1, pp. 289-292, 1990.

[17] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*: Prentice-Hall, Inc., 1993. .

[18] M. Ben, "Approaches Robustes pour la Vérification Automatique du Locuteur par Normalisation et Adaptation Hiarchique," Ph.D Thesis: University of Rennes I, 2004.

[19] B. S. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE,* vol. 64, pp. 460-475, 1976.

[20] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, pp. 254-272, 1981.

[21] F. Soong and A. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 36, pp. 871-879, 1988.

[22] W. M. Campbell, J. P. Campbell, T. P. Gleason, D. A. Reynolds, and W. Shen, "Speaker verification using support vector machines and high-level features," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 2085-2094, 2007.

[23] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "High-level speaker verification with support vector machines," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,(ICASSP'04),* vol.1, pp. 73-76, 2004.

[24] D. Klusacek, J. Navratil, D. A. Reynolds, and J. P. Campbell, "Conditional pronunciation modeling in speaker detection," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP' 03),* vol.4, pp. IV-804-7, 2003.

[25] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. A. Reynolds, and X. Bing, "Using prosodic and conversational features for high-performance speaker recognition: report from JHU WS'02," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'03)* vol.4, pp. IV-792-5, 2003.

[26] D. Reynolds, W. Andrews, J. Campbell, et. al, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'03),* vol.4, pp. IV-784-7, 2003.

[27] C. Longworth and M. J. F. Gales, "Derivative and parametric kernels for speaker verification," *Proceedings of the International Conference on Spoken Language Processing, (Interspeech'07),* pp. 310-313, 2007.

[28] 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. 308-311, 2006.

[29] M. McLaren, R. Vogt, and S. Sridharan, "*SVM Speaker Verification using session variability modelling and GMM supervectors",* ser. LNCS 4642. Springer-Verlag Berlin Heidelberg, pp.1077,1084, 2007.

[30] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine,* vol. 3, pp. 4-17, 1986.

[31] J. Fortuna, "Speaker Indexing based on voice biometrics," PhD thesis: University of Hertfordshire, 2006.

[32] J. D. Markel and A. H. Gray, *Linear prediction of speech*, Springer-Verlag New York, Inc.,1976.

[33] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of Acoustical Society of America,* vol. 87, pp. 1738-1752, 1990.

[34] P. Sivakumaran, "Robust text-dependent speaker verification," PhD thesis: University of Hertfordshire, 1998.

[35] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustic, Speech, and Signal Processing,* vol. 28, pp. 357-366, 1980.

[36] J. W. Picone, "Signal modelling techniques in speech recognition," *Proceedings of the of the IEEE,* vol. 81, pp. 1215-1247, 1993.

[37] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,(ICASSP'85),* vol. 10, pp. 387-390, 1985.

[38] A. M. Ariyaeeinia and P. Sivakumaran, "Comparison of VQ and DTW classifiers for speaker verification," *Proceedings of the European Conference on Security and Detection (ECOS'97),,* pp. 142-146, 1997.

[39] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for vector quantizer Design," *IEEE Transactions on Communications,* vol. 28, pp. 84-95, 1980.

[40] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society,* vol. 39, pp. 1-38, 1977.

[41] X. Huang, Y. Ariki, and M. Jack, *Hidden Markov Models for Speech Recognition*: Edinburgh University Press, 1990.

[42] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Procs of the IEEE,* vol. 77, pp. 257-286, 1989.

[43] L. Bahl and F. Jelinek, "Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition," *IEEE Transactions on Information Theory,* vol. 21, pp. 404-411, 1975.

[44] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the of the IEEE,* vol. 64, pp. 532-556, 1976.

[45] Y.-C. Zheng and B.-Z. Yuan, "Text-dependent speaker identification using circular hidden Markov models," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,(ICASSP'88),* vol. 1, pp. 580-582, 1988.

[46] A. E. Rosenberg, C.-H. Lee, and S. Gokcen, "Connected word talker verification using whole word hidden Markov models," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,(ICASSP'91),* vol. 1, pp. 381-384, 1991.

[47] A. E. Rosenberg, C.-H. Lee, and F. K. Soong, "Sub-word unit talker verification using hidden Markov models," *Procs ICASSP,* vol. 1, pp. 269-272, 1990.

[48] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian Mixture Models," *IEEE Transactions on Speech and Audio processing,* vol. 3, pp. 72-83, 1995.

[49] T. Matsui and S. Furui "Comparison of text-Independent speaker recognition methods using VQ-Distortion and discrete/continuous HMM's," *IEEE Transactions on Speech and Audi Processing,* vol. 2, pp. 456-459, 1994.

[50] N. Z. Tisby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Transactions on Signal Processing*, vol. 39, pp. 563-570, 1991.

[51] S. S. Haykin, *Neural networks: A comprehensive foundation*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1999.

[52] J. Hertz, A. Krogh, R. G. Palmer, and Santa Fe Institute., *Introduction to the theory of neural computation*: Addison-Wesley, 1991.

[53] C. M. Bishop, *Neural networks for pattern recognition*: Oxford University Press, 1995.

[54] H. Bourlard and C. J. Wellekens, "Links between Markov models and multilayer perceptrons," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 12, pp. 1167-1178, 1990.

[55] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'90),* vol.3, pp. 1361-1364, 1990.

[56] J. Oglesby and J. S. Mason, "Optimisation of neural models for speaker identification," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'90),* vol. 1, pp. 261-264, 1990.

[57] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 1995.

[58] N. Cristianni and J. Shawe-Taylor, *Support Vector Machines and other kernel-based learning methods*: Cambridge University Press, 2000.

[59] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, pp. 121-167, 1998.

[60] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proceedings of the IEEE International Conference*

on Acoustics, Speech and Signal Processing, (ICASSP'02), pp. 161-164, 2002

[61] V. Wan, "Speaker Verification using Support Vector Machines," PhD thesis: University of Sheffield, June 2003.

[62] J. Oglesby and J. S. Mason, "Radial basis function networks for speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'91),* vol.1, pp. 393-396, 1991.

[63] L. Minghui, X. Yanlu, Y. Zhiqiang, and D. Beiqian, "A New Hybrid GMM/SVM for Speaker Verification," in *Proceedings of the IEEE International Conference on Pattern Recognition,(ICPR'06)*, pp. 314-317, 2006.

[64] J. M. Naik and D. M. Lubensky, "A hybrid HMM-MLP speaker verification algorithm for telephone speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'94),* vol.1, pp. 153-156, 1994.

[65] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*: Kluwer Academic Publishers, 1994.

[66] J. P. Neto, C. Martins, and L. B. Almeida, "Speaker-adaptation in a hybrid HMM-MLP recognizer," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,(ICASSP'96),* vol. 6, pp. 3382-3385, 1996.

[67] S. Bengio and J. Mariethoz, "Learning the decision function for speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'01),* pp. 425-428, 2001.

[68] Campbell W.M, Reynolds D.A, and C. J.P, "Fusing discriminative and generative methods for speaker recognition: experiments on switchboard and NFI/TNO field data," in *Proceedings of the Speaker Odyssey, (Odyssey'04),* pp. 41-44, 2004.

[69] S. Fine, J. Navratil, and R. Gopinath, "Enhancing GMM scores using SVM 'hints'," in *Proceedings of the European Conference on Speech Communication and Technology'01*, pp. 1760-1767, 2001.

[70] S. Fine, J. Navratil, and R. A. Gopinath, "A hybrid GMM/SVM approach to speaker identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 417-420, 2001.

[71] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1,pp. 97-100, 2006.

[72] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, pp. 929,932, 1996.

[73] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.

[74] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," PhD thesis: Cambridge University, 1995.

[75] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust Speaker Recognition in Noisy Conditions," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 15, pp. 1711-1723, 2007.

[76] A. Drygajlo and M. El-Malikim, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 121-124, 1998.

[77] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113-120, 1979.

[78] J. S. Lim and A. V. Oppenheim, "Enhancement of bandwidth compression of noisy speech," in *Proceedings of the IEEE 67,* vol. 67, pp 1586-1604, 1979.

[79] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*. New York: Institute of Electrical and Electronics Engineers, 2000.

[80] Y. Chang Huai, K. Soo Ngee, and S. Rahardja, "Kalman filtering speech enhancement incorporating masking properties for mobile communication in a car environment," in *Proceedings of the IEEE ICME,* vol. 2, pp. 1343-1346, 2004.

[81] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proceedings of the ICASP,* pp. 177-180, 1987.

[82] H. Sorenson, *Kalman filtering: theory and application*: IEEE press New York, 1985.

[83] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication,* vol. 34, pp. 267-285, 2001.

[84] M. T. Padilla, T. F. Quatieri, and D. Reynolds, "Missing Feature theory with soft spectral subtraction for speaker verification," in *Proceedings of the International Conference on Spoken Language Processing, (Interspeech'06),* pp. 913-916, 2006.

[85] L. Besacier and J. F. Bonastre, "Frame pruning for speaker recognition," in *Proceedings of the ICASSSP,* 1998.

[86] S. G. Pillay, A. Ariyaeeinia, and M. Pawlewski, "Effectiveness of speaker-dependent feature score pruning in speaker verification," in *Proceedings of the International Symposium on Communications, Control and Signal Processing,(ISCCSP'08),* pp. 372-376, 2008.

[87] K. Yoshida, K. Takagi, and K. Ozeki, "Improved model training and automatic weight adjustment for multi-SNR multi-band speaker identification system," in *Proceedings of the International Conference on*

Spoken Language Processing, (Interspeech'04), vol. 3, pp. 1749-1752, 2004.

[88] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust Speaker Recognition in Noisy Conditions," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 1711-1723, 2007.

[89] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Transactions on Speech and Audio Processing,* vol. 2, pp. 194-205, 1994.

[90] W. Lit Ping and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proceedings of the Signal Processing and Communications Applications Conference, (ICASSP'01)* vol. 1, pp. 457-460, 2001.

[91] Z. Tufekci, "Noise robust speaker verification using parallel model combination and local features," in *Proceedings of the Signal Processing and Communications Applications Conference, (ICASSP'04)* pp. 422-425, 2004

[92] O. Bellot, D. Matrouf, T. Merlin, and J. F. Bonastre, "Additive and convolutional noises compensation in speaker recognition," in *Proceedings of the International Conference on Spoken Language Processing, (Interspeech'00),* pp. 799-802, 2000.

[93] L. Yang and W. Gong, "Multi-SNR GMMs-Based Noise-Robust Speaker Verification using 1 / f^α noises," in *Proceedings of the International Conference on Pattern Recognition, (ICPR'06),* vol. 4, pp. 241-244, 2006.

[94] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM composition in noisy environments," *Computer Speech & Language,* vol. 10, pp. 107-116, 1996.

[95] A. M. Ariyaeeinia and P. Sivakumaran, "Analysis and comparison of score normalisation methods for text-dependent speaker verification," in *Proceedings of the European Conference on Speech Communication and Technology, (Eurospeech'97),* pp. 1379-1382, 1997.

[96] D. Reynolds, "Comparison of background normalisation methods for text-independent speaker verification," in *Proceedings of the European Conference on Speech Communication and Technology,(Eurospeech'97),* pp. 963-966, 1997.

[97] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalised scores for speaker verification," in *Proceedings of the International Conference on Spoken Language Processing, (Interspeech'92),* vol. 1, pp. 599-602, 1992.

[98] K.-P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'88),* vol. 1, pp. 595-598, 1988.

[99] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalization methods in open-set speaker identification," in *Proceedings of the Speaker Odyssey*, *(Odyssey'04),* pp. 369-376, 2004.

[100] NIST, "A universal transcription format (UTF) annotation specification for evaluation of spoken language technology corpora," in *[www.nist.gov/speech/hub4_98/utf-1.0-v2.ps]*, 1998.

[101] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the European Conference on Speech Communication and Technology,(EuroSpeech'97),* pp. 1895-1898, 1997.

[102] M. Smithson, *Confidence Intervals*: Sage University Paper, 2003.

[103] M. Przybocki and A. Martin, "The NIST year 2003 speaker recognition evaluation plan,"*http://www.nist.gov/speech/tests/spk/2003/2003-spkrec-evalplan-v2.2.pdf, 2003.*

[104] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM," LDC Catalog No.: LDC93S1: Massachusetts Institute of

Technology (MIT), SRI International (SRI), Texas Instruments Inc. (TI), 1990.

[105] I. McCowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *Proceedings of Speaker Odyssey,* (*Odyssey'01),* pp. 101-106, 2001.

[106] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication,* vol. 17, pp. 91-108, 1995.

[107] J. Makhoul, "Linear Prediction: A tutorial review," *Proceedings of the IEEE,* vol. 63, pp. 561-580, 1975.

[108] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *ASSP Magazine,* vol. 3, pp. 4-16, 1986.

[109] J. Baszun, "Voice Activity Detection for speaker verification systems" in *Proceedings of the 2nd international conference on Rough sets and knowledge technology*. pp. 181-186, 2007.

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

[111] T. Quatieri, *Discrete-time speech signal processing- principles and practice*, 2002.

[112] B. S. Atal, "Automatic speaker recognition based on pitch contours," *The Journal of the Acoustical Society of America,* vol. 52, pp. 1687-1697, 1972.

[113] F. Soong and A. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 36, pp. 871-879, 1988.

[114] S. Furui, "Comparison of speaker recognition methods using static features and dynamic features," *IEEE Transactions on Acoustic, Speech, Signal Processing,* vol. 29, pp. 342-350, 1981.

[115] 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.

[116] M. Pawlewski, B. P. Milner, S. A. Hovell, D. G. Ollasson, S. P. A. Ringland, K. J. Power, S. N. Downey, and J. Bridges, "Advances in Telephony Based Speech Recognition," *BT Technology Journal,* vol. 14, pp. 127-150, 1996.

[117] D. A. Reynolds, "Automatic speaker recognition using Gaussian Mixture Models," *The Lincoln Laboratory Journal,* vol. 8, pp. 173-192, 1995.

[118] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing,* vol. 3, pp. 72-83, 1995.

[119] J. L. Gauvain and C.-H. Lee, "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.

[120] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meigner, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *EURASIP Journal on Applied Signal Processing,* pp. 430-451, 2004.

[121] R. Collobert and S. Begio, "SVMTorch: Support vector machines for large-scale regression problems," *Journal of Machine Learning Research,* vol. 1, pp. 143-160, 2001.

[122] D. A. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification using adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

[123] Z. N. Karam and W. M. Campbell, "A new kernel for SVM MLLR based speaker recognition," in *Proceedings of the International Conference on Spoken Language Processing,* **(***Interspeech'07),* 2007.

[124] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing,* vol. 4, pp. 352-359, 1996.

[125] D. A. Reynolds, "Comparison of background normalisation methods for text-independent speaker verification," *Proceedings of the European Conference on Speech Communication and Technology, (Eurospeech'97),* pp. 963-966, 1997.

[126] D. E. Sturim and D. A. Reynolds, "Speaker Adaptive Cohort Selection for TNorm in Text-Independent Speaker Verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'05),* pp. 741-744, 2005.

[127] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Proceeding of the Speaker and Language recognition workshop*, pp. 219-226, 2004.

[128] 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.

[129] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'06),* pp. 97-100, 2006.

[130] B. G. B. Fauve, D. Matrouf, N. Sheffer, J.-F. Bonastre, and J. S. D. Mason, "State-of-Art performance in text-independent speaker verification through open-source software," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 15, pp. 1960-1968, 2007.

[131] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in Channel Compensation for SVM speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'05),* pp. 629-632, 2005.

[132] A. Solomonoff, C. Quillen, and W. Campbell, "Channel compensation for SVM speaker recognition," in *Proceedings of Speaker Odyssey, (Odyssey'04)*, pp. 57-62, 2004.

[133] M. Ben and F. Bimbot, "D-MAP: A distance-normalized MAP estimation of speaker models for automatic speaker verification," in *Proceedings of the*

IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'03), pp. 69-72, 2003.

[134] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling Prosodic Features With Joint Factor Analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2095-2103, 2007.

[135] 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. 1435-1447, 2007.

[136] P. Kenny, P. Boulianne, P. Ouellet, and P. Dumouchel, "Factor Analysis simplified," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'05),* pp. 637-640, 2005

[137] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, pp. 17-38, 2008.

[138] A. Varga, H. J. M. Steeneken, M. Tornlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Speech Research Unit, Defense Research Agency, 1992.

[139] S. G. Pillay, A. Ariyaeeinia, M. Pawlewski, and P. Sivakumaran, "Speaker verification under mismatched data conditions," *IET Signal Processing,* vol. 3, pp. 236-246, 2009.

[140] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, and J. Fortuna, "On the enhancement of speaker identification accuracy using weighted bilateral scoring," in *Proceedings of International Carnahan Conference on Security Technology (ICCST 2008)*, pp. 254-258, 2008,

[141] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *Proceedings of the International Conference on Acoustics,Speech and Signal Processing, (ICASSP 2009)*, pp. 4237-4240, 2009.

[142] F. Alsaade, A. Ariyaeeinia, A. Malegaonkar, and S. Pillay, "Qualitative fusion of normalised scores in multimodal biometrics," *Pattern Recognition Letters,* pp. 564-569, 2008.

[143] N. Brummer, L. Burget, J. Cernocky, O. Glembek, and et. al, "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.

[144] J.-C. Junqua, S. Fincke, and K. Field, "The Lombard Effect: A Reflex to better communicate with others in noise," in *IEEE International Conference on Acoustics,Speech, and Signal Processing,(ICASSP'99),* 1999.

[145] M. Arcienega and A. Drygajlo, "On the number of Gaussian components in a mixture: an application to speaker verification tasks," *Proceedings of the European Conference on Speech Communication and Technology, (Eurospeech'03),* pp. 2673-2676, 2003.

[146] D. G. Luenberger, *Linear and NonLinear Programming*: Addison-Wesley Publishing Company, 1984.

[147] T. Joachims, *Learning to Classify Text using Support Vector Machines*: Kluwe Academic Publishers / Springer, 2002.

[147] A.Asano, "Support vector machine and kernel method," *Pattern information processing,* pp. 1-3, 2004.

AUTHOR'S RELATED PUBLICATIONS

JOURNAL PUBLICATIONS

1. S. G. Pillay, A. Ariyaeeinia, M. Pawlewski, and P. Sivakumaran, "Speaker verification under mismatched data conditions," *Signal Processing, IET,* vol. 3, pp. 236-246, 2009.
2. F. Alsaade, A. Ariyaeeinia, A. Malegaonkar, and S. Pillay, "Qualitative fusion of normalised scores in multimodal biometrics," *Pattern Recognition Letters,* pp. 564-569, 2008.
3. F. Alsaade, A. M. Ariyaeeinia, A. S. Malegaonkar, M. Pawlewski, and S. G. Pillay, "Enhancement of Multimodal Biometric segregation using Unconstrained Cohort Normalisation," *Pattern Recognition,* issue 41, vol. 5, pp.814-820, 2008.

CONFERENCE PUBLICATIONS

1. S. G. Pillay, A. Ariyaeeinia, P. Sivakumaran and M. Pawlewski "Open Set Speaker Identification under mismatch conditions," Procs. of the 10th Annual Conference of the International Speech Communication Association, Interspeech 2009, pp. 2347-2350.
2. S. G. Pillay, "Open-set speaker identification," *Poster presented at the Science and Technology Research Institute Showcase*, University of Hertfordshire, UK, 2008.
3. S. G. Pillay, A. Ariyaeeinia, and M. Pawlewski, "Effectiveness of speaker-dependent feature score pruning in speaker verification," in *International Symposium on Communications, Control and Signal Processing* Malta, 2008, pp. 372-376.
4. A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, S. Pillay, "On the discriminative capabilities of speech cepstral features, " *Proceedings of the COST-2101 Workshop on Biometrics and Identity Management (BIOID'08), Roskilde, Denmark, pp. 95-103, May 2008.*
5. A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, S. Pillay, "Discrimination effectiveness of speech cepstral features," *LNCS 5372 (Springer),* pp. 91-99, 2008.

PATENT

1. Ariyaeeinia A.M., Pillay S.G., Pawlewski M., "Speaker Verification," International Patent: WO/2010/049695 - PCT/GB2009/002579.