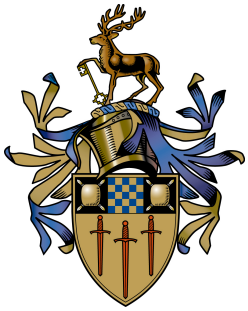
Robust Speaker Identification against Computer Aided Voice Impersonation

Zargham Haider

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey



Centre for Vision, Speech and Signal Processing, I-Lab
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

December 2011

Abstract

Speaker Identification (SID) systems offer good performance in the case of noise free speech and most of the on-going research aims at improving their reliability in noisy environments. In ideal operating conditions very low identification error rates can be achieved. The low error rates suggest that SID systems can be used in real-life applications as an extra layer of security along with existing secure layers. They can, for instance, be used alongside a Personal Identification Number (PIN) or passwords. SID systems can also be used by law enforcements agencies as a detection system to track wanted people over voice communications networks. In this thesis, the performance of the existing SID systems against impersonation attacks is analysed and strategies to counteract them are discussed. A voice impersonation system is developed using Gaussian Mixture Modelling (GMM) utilizing Line Spectral Frequencies (LSF) as the features representing the spectral parameters of the source-target pair. Voice conversion systems based on probabilistic approaches suffer from the problem of over smoothing of the converted spectrum. A hybrid scheme using Linear Multivariate Regression and GMM, together with posterior probability smoothing is proposed to reduce over smoothing and alleviate the discontinuities in the converted speech. The converted voices are used to intrude a closed-set SID system in the scenarios of identity disguise and targeted speaker impersonation. The results of the intrusion suggest that in their present form the SID systems are vulnerable to deliberate voice conversion attacks. For impostors to transform their voices, a large volume of speech data is required, which may not be easily accessible. In the context of improving the performance of SID against deliberate impersonation attacks, the use of multiple classifiers is explored. Linear Prediction (LP) residual of the speech signal is also analysed for speaker-specific excitation information. A speaker identification system based on multiple classifier system, using features to describe the vocal tract and the LP residual is targeted by the impersonation system. The identification results provide an improvement in rejecting impostor claims when presented with converted voices. It is hoped that the findings in this thesis, can lead to the development of speaker identification systems which are better equipped to deal with the problem with deliberate voice impersonation.

Key words: Speaker Identification, Voice Conversion, Gaussian Mixture Modeling, Identity Disguise, Voice Impersonation, Multiple Classifier Systems

Email: z.haider@surrey.ac.uk

WWW: http://www.ee.surrey.ac.uk/CCSR/research/ilab/

Acknowledgment

I would like to express my sincere gratitude to my supervisors Prof. Ahmet Kondoz and Dr. Stephane Villette for their support, guidance and encouragement during my PhD and without their support it would have been unbearable. I would also like to thank my colleagues and friends at I-Lab for their friendship, advise and encouragement specially Dr. DeSilva, Wasim Ahmed, Gocke Nur and Thilini Rajakaruna for their help and support during the various stages of my PhD.

I would also like to say a special thank you to my parents, my brother and specially to my sister Rabiya Abbas for their love, continuous encouragement and understanding.

I am also thankful to SUPARCO for their sustained co-operation and patience during my PhD specially to Dr. Muhammad Riaz Suddle for his invaluable guidance and support.

Abbreviations

ANN: Artificial Neural Network

BPF: Band Pass Filter

CMS: Cepstral Mean Subtraction

DCT: Discrete Cosine Transform

DET: Detection Error Trade-off

DFT: Discrete Fourier Transform

DFW: Dynamic Frequency Warping

DTW: Dynamic Time Warping

EER: Equal Error Rate

EM: Expectation Maximization

FA: False Acceptance

FAR: False Acceptance Rate

FFT: Fast Fourier Transform

FR: False Rejection

FRR: False Rejection Rate

GMM: Gaussian Mixture Model

HMM: Hidden Markov Model

HPF: High Pass Filter

LAR: Log Area Ratios

LDC: Linguistic Data Consortium

LP: Linear Prediction

LPC: Linear Prediction Coefficients

LPCC: Linear Prediction Cepstral Coefficients

LPF: Low Pass Filter

LSF: Line Spectral Frequencies

LSP: Line Spectral Pair

MFCC: Mel Frequency Cepstral Coefficients

ML: Maximum Likelihood

NIST: National Institute of Standards and Technology

NN: Neural Networks

PDSS: Power Difference of Spectra in Sub-band

PIN: Personal Identification Number

PSOLA: Pitch Synchronous Overlap Add

R-PDSS: Residual based Power Difference of Spectra in Sub-bands

RASTA: RelAtive SpecTrAl

RMS: Root Mean Squared

SD: Spectral Distortion

SID: Speaker Identification

SVM: Support Vector Machines

UBM: Universal Background Model

VAD: Voice Activity Detection

VQ: Vector Quantization

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The past few decades have seen an enormous increase in the Human-Machine interactions. From the use of a mobile phone as a personal assistant to the use of the internet as a means for information sharing, from computer chips in every other product to the security systems, mankind is reaping the benefits of this partnership. As more and more people rely on the advancements in the field of computation and digital technology, there is a greater responsibility on part of the machines for accurately identifying an individual or a group of individuals in order to grant access to certain features of a service or other benefits. Various approaches can be taken regarding the recognition task like what the entity knows, what the entity has, what the entity is or where the entity is. The traditional methods of recognition and authentication require possession of certain items like a swipe card or the knowledge of some secret information like a password or a Personal Identification Number (PIN). However, such systems are error prone in establishing a false identity once the proper inputs are presented to them regardless of who the presenter is. Biometrics is a means to prevent such identity thefts. Recognition systems based on biometrics have grown in popularity in the recent past. Each individual has certain unique physical and/or behavioural characteristics that distinguishes them from the others. Biological features such as Retina, Facial geometry, Voice, Finger prints, Hand geometry etc. are examples of such unique features. Certain biometric systems have been developed that recognize the individuals based on their physiological and/or behavioural characteristics. For these systems to perform accurately it is of utmost importance that these characteristics should be unique and permanent, easily collectable and widely acceptable while being available universally. Among all the biometrics used for recognition tasks, voice is unique, reliable and non-intrusive. Using voice as a biometric has the qualities of being user-friendly, can convey the emotions of the individual and it can be used over the existing telecommunications links for remote authentication. The ongoing research in the field of speaker identification systems is aimed at developing systems that give reliable performance under various operating conditions.

1.1 Objectives

It is generally assumed that the impostors will not make an attempt to conceal their voices from the SID systems. In order for a SID system to be trustworthy the system should not only give reliable performance in ideal conditions but it should also be resilient against deliberate impersonation attacks. The most obvious attack on a voice recognition system is voice impersonation by professional imitators. This approach however fails, as the traits of human voice cannot be easily altered by a human impersonator . In the case of computer-aided impersonation, false acceptance rate of 86% have been exhibited by the recognition system under attack in some preliminary studies found in literature. An effort will be made during this research to study how the various voice recognition techniques are affected by such deliberate impersonation attacks. The findings in this thesis can lead to the development of a speaker identification system which can have good identification performance against voices that have been deliberately altered by the use of voice conversion algorithms. To deal with the problem of analysing the performance of SID system against deliberate voice conversion attacks, it is important to identify the weaknesses and strengths of both the SID and the voice conversion systems. To this end, these exists no defined framework for SID system in literature, when dealing with the threat of computer-aided voice conversion. In this thesis the proposed objectives of research are defined as below:

- Voice conversion techniques have gained popularity with the availability of increased computational power and better statistical modelling tools. GMM based models are the obvious choice for voice conversion techniques because of their ability to model underlying phonetic classes in the speech sounds. GMM based techniques, however, are not without their disadvantages. Limited amount of training data can lead to audible artefacts in the output speech of the voice conversion systems. One objective of this thesis is to improve the quality of the output speech by reducing the native problems of the GMM based systems and the degradations resulting from the limited amount of training data available to the voice conversion system.

- A lot of research effort has been made to improve the performance of the SID systems under different operating conditions. With the emergence of easy to use voice conversion techniques, there exists no defined framework for testing the SID performance against computer altered synthetic voices. This thesis investigates the performance of the SID systems when presented with speech utterances from different impostors when they are deliberately trying to conceal their identity from the SID or targeting a speaker who is known to the SID.

- It is widely believed that the individuality of a speaker's voice is due to the differences in the shape of sound producing organs; mainly the vocal tract system. Based on this knowledge state-of-the-art in speaker identification primarily relies on the low-level characteristics by using short-time features representing the

spectral envelope of the speech spectrum. Furthermore, the preferences of the voice conversion system for certain features suggest that the use of different features would result in varying performance for the same impostor-target pair. This thesis investigates the usability of multiple classifier systems with different feature sets for the problem of speaker identification systems against intentionally modified voices.

- Apart from the use of short-time features related to the spectral envelope of the speech spectrum in SID, other levels of information can convey important information about the perceived speaker identity. The speakers are also able to identify the speaker from the linear prediction residual of speech signals. This suggests that certain speaker related information is still available in the speech signal even after the removal of the contribution by the vocal tract. One of the objectives of this thesis is to determine the presence of speaker specific information present in the speech residual. Furthermore, to eliminate the requirement of developing new techniques for SID systems, this thesis investigates the use of the residual based information with the traditional spectral envelope based features in a multiple classifier based SID system.

1.2 Original Contributions

The work reported in this thesis is carried out to meet the objectives outlined in the previous section. The major contributions in this thesis are outlined below.

- Speaker identification system has been investigated and baseline system has been implemented using GMM. The performance of the systems is tested on clean speech and is consistent with the literature.

- A voice conversion system for converting the voice of one speaker to another has been investigated. A baseline voice conversion system is implemented with the use of Line Spectral Frequencies for mapping the spectral properties of the source speaker to the target speaker using speaker specific GMM. A solution for over smoothing in GMM voice conversion systems is addressed by means of a hybrid model combining the GMM and Linear Multivariate Regression on the source model components. The voice conversion system requires huge amounts of data to find the proper correspondences between the feature vector spaces of the source and the target speakers. In practice the availability of training data at such scale is not possible. The lack of training data for the transformation function causes the output speech to be discontinuous. A posterior probability smoothing approach is presented to reduce the discontinuity between the adjacent frames of the converted speech signal. Subjective evaluations are presented favouring the modified speech.

- The performance of the speaker identification system is analysed against voice modified by the use of voice conversion techniques. The performance of the system is tested in the scenarios of identity disguise where a speaker who is enrolled in the speaker identification system has deliberately modified his/her voice to dodge the system, and in the case of targeted voice impersonation where an impostor has changed his voice characteristics, by means of a voice conversion system, to match a target speaker. Also, the performance of the system is analysed in the case of intra-gender and cross-gender voice conversions. The simulation results show that in their present form the speaker identification systems are highly vulnerable to computer-aided voice impersonation attacks.
- Previous studies have shown the improvements in the performance of multiple classifier systems in the speaker identification application. The use of multiple classifier systems have been proposed for speaker identification systems against converted synthetic voices. The use of classifiers using different feature sets, characterising different properties of the speech spectrum has been proposed for the speaker identification tasks. Specifically the use of LPCC and MFCC has been analysed in a multiple classifier system against converted synthetic voices. The performance of the system is investigated using different combination schemes. It was shown that the use of multiple classifiers can improve the identification performance of the systems. The LP-residual signal was analysed for speaker specific information and the use of Power Difference in the Spectral Sub-bands (PDSS) based features was proposed along with the use of traditional features characterising the spectral envelope in the context of multiple classifier speaker identification systems. The results showed that with the use of LP-residual based features the performance of the system improved substantially against converted synthetic voices.

1.3 Thesis Outline

The research work is mainly focused on improving the speaker identification performance against voice conversion. The thesis is organised as follows:

- In Chapter 2, anatomy of the human sound production system has been described. The contribution of different organs to speech sounds and different sounds produced by the sound production mechanism is reviewed. In order to process and extract information from the speech sounds, the speech signal undergoes various signal processing techniques. This information is generally represented in the form of features that are based on a mathematical model which try to closely approximate the human sound production mechanism. In this chapter, various features that have been used in the speech processing tasks including speech coding, speech recognition and speaker identification are described with details of their extraction from the speech waveform.

- Chapter 3 presents some of the popular speaker modelling techniques, such as GMMs, that are used for speaker recognition systems. Practical issues such as initialisation, training and testing processes are described in detail. This is followed by the introduction of the baseline speaker identification and verification systems. The experimental set up of speaker identification and verification tasks are described, explaining the details of the training and testing processed for both tasks. The identification and verification system performances are presented using clean and noisy speech samples.

- Chapter 4 describes the process of voice conversion using speaker specific GMM. Extraction of speaker specific information from the parallel speech corpus of the source and target speakers is explained in detail. The GMM based voice conversion system suffers from the phenomenon of over smoothing which is addressed in this chapter by the use of of hybrid scheme using linear multivariate regression and GMM. The smoothing of the posterior probabilities during the estimation of target speaker's characteristics is also proposed to deal with audible degradations which result from the availability of limited amount of training data for the voice conversion system. Subjective evaluations were carried out to determine the performance of the proposed technique in comparison to the traditional GMM based approaches.

- Chapter 5 details the results of intrusions into a speaker identification system using the converted synthetic voice. Two different scenarios of deliberate modifications of the speech signal are presented namely; identity disguise and targeted voice impersonation. The performance of the system is analysed in terms of the ability of the speaker identification system to identify the source and the target speakers from the converted voices. The performance of the speaker identification system is also analysed in terms of intra-gender and cross-gender voice conversions.

- Chapter 6 investigates the use of multiple classifier systems for the task of speaker identification. The concepts of contextual information extraction and complementariness are introduced. The use of GMM based classifiers using MFCC and LPCC as feature vectors is analysed in the framework of multiple classifier system against converted synthetic voices. Also the linear prediction residual of the speech signal is analysed for speaker specific information and the R-PDSS is used for the extraction of speaker specific information from the LP residual. Different combination of MFCC, LPCC and R-PDSS are analysed in improving the performance of the speaker identification system against the identity disguise and targeted voice impersonation.

- Chapter 7 provides a summary of the contributions made in this thesis towards robust speaker identification against computer aided voice impersonation and some suggestions for future work.

Chapter 2

Speech Signal Processing Techniques

2.1 Introduction

Speech is probably the most important modality in human communications. Not only does it convey information about what is being spoken but also helps to identify the speakers along with complimentary information about their physical and emotional states. In order to develop a system based on the speech signal, whether it is a speaker recognition system or a voice transformation system, it is important to understand the properties of the speech signal itself, how it can be represented and manipulated. The next section describes a review of the human speech production mechanism and how sounds are produced. A mathematical model commonly used for representing the speech production is also introduced. Later the techniques involved in the processing of the speech signal are discussed.Features describing emphasizing the properties of the speech signal related to the speaker identity are described in detail. The last section lists the distance measures used in this study.

2.2 Human Sound Production Mechanism

Human voice is unique and universally available. No two individuals sound identical due to the differences in the physical structures of their sound producing organs and the mannerisms of speaking. The physiological differences in the lengths of the vocal tract, the shape of the larynx and other parts, and the behavioural characteristics involving the use of a specific accent, intonation style, pronunciation pattern and rhythm make up a unique system that accounts for the speaker specific characteristics

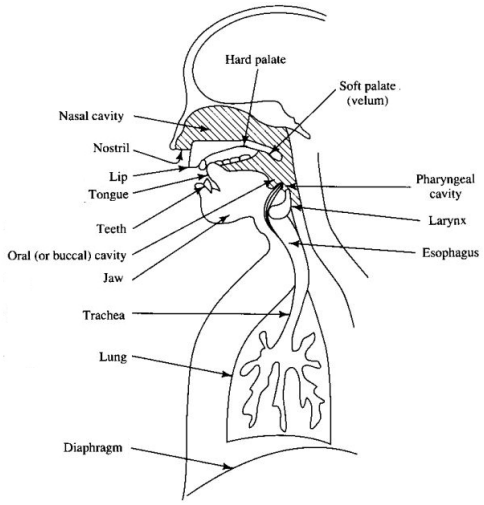


Figure 2.1: Human Sound Production Mechanism[1]

The mechanism of human speech production can be divided into three main groups namely lungs, larynx and the vocal tract [5]. The shape of the vocal tract is the most important physiological feature. The vocal tract consists of the laryngeal, oral and nasal pharynx, and the oral and nasal cavities. The vocal tract is located between the vocal cords and the lips. The cross-sectional area may vary between zero and 20 cm^2 and depends upon the process of articulation [6]. The process of articulation involves the manipulation of jaw, velum, tongue and lips. Air pressure generated from the lungs is carried over the vocal cords through the trachea. A small opening called glottis, exists between the vocal cords. Glottis normally remains open, however during the production of speech its shape is manipulated, resulting in an irregular airflow, called the glottal source or the source of the speech [7]. Speech sounds produced by the passage of the glottal source through the vocal tract and the articulators, can be broadly classified as voiced, unvoiced or mixed excitation sounds [1].

Voiced sounds are characterized by their periodicity and high energy. During the production of voiced sounds, the airflow after passing through the glottis causes the tensed vocal cords to vibrate. The period of transition from the open state of the glottis to the closing state, is termed as the fundamental period or T_0. The reciprocal of the fundamental period is the fundamental frequency given by F_0. The vibration of the vocal cords introduces quasi-periodic pulses in the airflow. The spectrum of a speech signal contains well-defined regions of emphasis or resonances and de-emphasized anti-resonances. These resonances, also known as formants or formant frequencies, are a result of various articulators forming cavities and sub-cavities in the vocal tract. The locations of these formant frequencies depend upon the shape of the vocal tract. The formants are labelled as $F_1, F_2, F_3, \ldots$ starting with the lowest frequency. Speech signal contains an infinite number of formants however, in practice only 3-5 are used in the post sampling Nyquist band [8]. Unvoiced sounds are produced if there is no vibration of the vocal cords. These sounds are produced when turbulent airflow passes a constriction in the vocal tract. Unvoiced sounds are random, white-noise like signals carrying

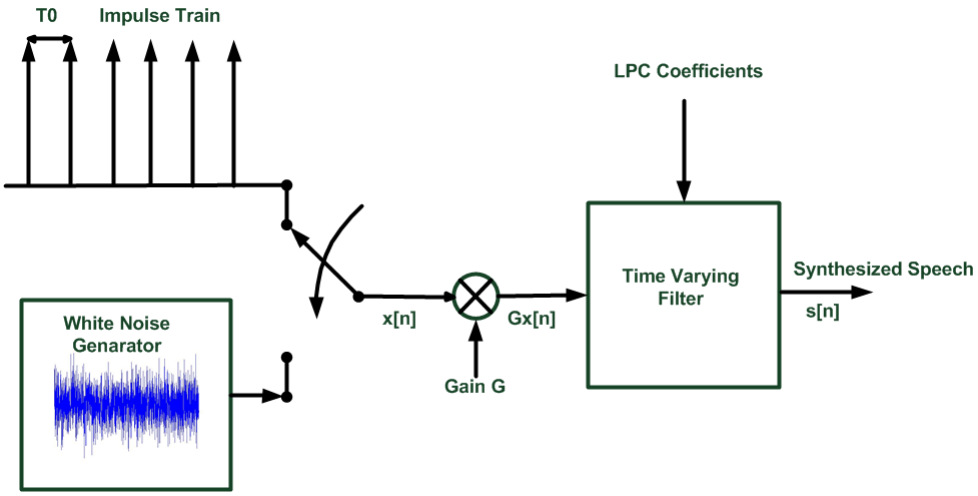


Figure 2.2: Source-Filter Model of synthetic speech production[1]

far less energy than the voiced sounds. Mixed excitation speech is produced when air from the lugs passes a constriction in vocal tract while the vocal cords are vibrating. Nasal sounds are produced when the airflow passes through the nasal cavity. Plosives are produced when the airflow is blocked and then suddenly released in the vocal tract [6, 9].

2.2.1 Synthetic Speech Production

The most widely used method of representing the human speech production mechanism is the source-filter model shown in Fig 2.2. The vocal tract is generally modelled as a time-varying all-pole filter and the glottal source is represented as a periodic impulse train for voiced segments of speech or white noise in the case of unvoiced speech [1]. The coefficients of the all-pole filter are determined by linear prediction to minimize the mean-squared error of the speech signal to be reproduced. Synthesized speech is generated by the excitation of the all-pole filter with the glottal source.

The source-filter model make certain assumption about the nature of source and excitation signals. According to the all-pole model, the excitation is considered to be independent of modulation and the all-pole filter in linear in nature. These assumption although not entirely true, serve to simplify the analysis of speech production and provide computational savings.

2.3 Speaker Characteristics

The speech signal carries different types of information. The primary information transmitted by the speech signal is the message, *what* is said, but also indicates the source of the message,*who* said it as well as the environment in which the speech signal was generated. Speaker characteristics refer to the properties of the speech

signal related to the individual and are dependent on the text of the message and the environment. The characteristics of the speech signal can be broadly classified as the following types:

Segmental Cues

Acoustic descriptions of the segmental cues consists of location and bandwidths of the formants, spectral tilt, $F0$, and energy. The segmental cues are dependent on the physiological and physical characteristics of the speech organs as well as on the emotional state of the speakers [10].

Supra-segmental Cues

The supra-segmental features describe the prosodic features associated with the mannerism of speaking such as the duration of a phoneme, intonation patterns and the amount of stress in pronunciation of an utterance or part of it. These cues are manifested as the rate of speaking, average pitch and the loudness in the speech utterances. These cues are influenced by the social and behavioural conditions [11].

Linguistic Cues

The linguistic cues are associated with the choice of particular words, accents and dialects. Such cues are very difficult to model as it would require an extensive study regarding the circumstances involved in the choice of words, variation in accents and the use of particular dialects.

The next sections discuss the preprocessing techniques for the speech signal. Later important features used in the modelling of the speech signals will be discussed.

2.4 Feature Extraction

From Sec 2.2.1, speech is a convolution of the glottal source and an all-pole filter representing the vocal tract. The process of extracting specific information from the speech waveform is called *feature extraction* or *speech parametrization*. The main objective of feature extraction is to effectively represent the speech data through a reduced data set for the modelling tasks. During feature extraction the speech signal is represented by feature vectors that can efficiently and effectively capture the characteristics of the speech signal. For the speaker modelling system to perform efficiently it is important that the extracted features produce low intra-speaker variability i.e, those arising from variation in speakers' mood, emotion, physical condition etc., and produce high inter-speaker variability i.e. effectively highlight the differences in different speakers. The

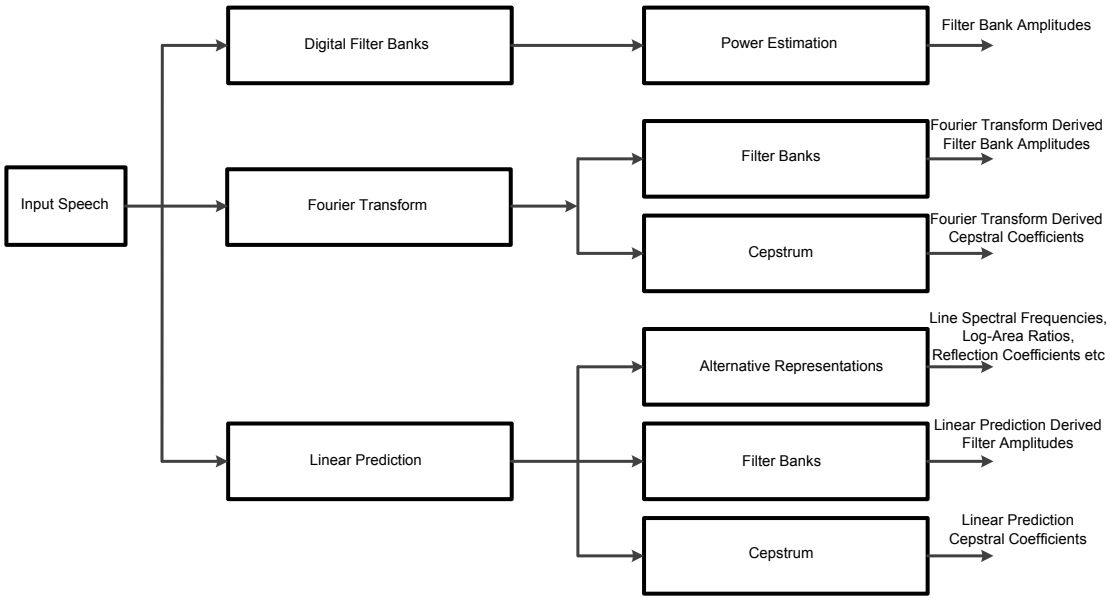


Figure 2.3: Major Spectral Analysis Techniques

short-time spectral analysis, usually carried out on a windowed segment of $20 - 30$ ms of speech, produces features that characterize the spectral information in the speech signal [8][6][1]. The major spectral analysis algorithm used in speech processing systems are depicted in Fig. 2.3 [12].

This section lists the preprocessing steps involved and a description of the various features used in this study.

2.4.1 Pre-Processing

The preprocessing steps prepare the signal for the process of feature extraction. The aim of preprocessing is to enhance the speech and improve the quality of the features that are to be extracted. In this section some of the commonly used preprocessing techniques are discussed.

2.4.1.1 Pre-emphasis

The speech spectral envelope has a high frequency roll-off due to the radiation effect of the lips [13]. This results in the high frequency components having low amplitude thus increasing the dynamic range of the speech signal. Speech analysis techniques require high computational precision to obtain the features from the high end of the spectrum. A simple solution is to process the speech with a pre-emphasis filter having a system function

$$H(z) = 1 - \alpha z^{-1} \tag{2.1}$$

which is high-pass in nature. A typical value of 0.97 is commonly used for α [1]. The use of a pre-emphasis filter reduces the numerical problems encountered in the Linear Prediction analysis which will be discussed in 2.4.3.

While synthesizing the speech, a de-emphasis filter with a frequency response opposite to Equation 2.1, and given as

$$G(z) = \frac{1}{1 - \alpha z^{-1}} \tag{2.2}$$

is used to give the synthetic speech the same spectral shape as the original speech.

2.4.1.2 Voice Activity Detection

The performance of the speech processing systems is degraded with the inclusion of silence with speech. Features extracted from the silence part of the speech model the environment rather than the speaker. Therefore, it is important to separate silence from speech, a process known as *Voice Activity Detection* (VAD). Generally, an energy based VAD is used to separate silence from speech [14, 15]. In this work, the speech databases used for recognition and impersonation tasks, described in (Sections 3.8.1.1 and 4.6.1), have been transcribed with the exact locations of speech and silence. This information was utilized in this work for the separation of silence from speech. In real life applications however, a VAD should be used to remove silence intervals. VAD is also beneficial in reducing the amount of data needed in speech processing tasks.

2.4.2 Frame Analysis and Windowing

Speech is a non-stationary signal which changes in time. However, speech is considered a quasi-stationary signal [6], therefore short-time analysis can be utilized to parametrize speech. In order to facilitate short-time analysis the speech signal is divided into smaller segments called *frames.* The use of frames for speech analysis validates the stationary assumption. These frames are often overlapped to capture the inter-frame dynamics. The short-time Fourier Transform is an important tool in the analysis of speech signals and it represents the time-varying properties of the speech signal in the frequency domain. The short-time Fourier transform can be represented as [16]

$$\mathscr{F}\left(s[n]\right) = S_k(j\omega) = \sum_{n=-\infty}^{\infty} w[k-n]s[n]e^{-j\omega n} \tag{2.3}$$

where $w[k-n]$, represents a real window function used to isolate the frame from the rest of the signal. The simplest window function is the rectangular window.

$$w[n] = \begin{cases} 1 & 0 \leq n \leq N_w - 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

where N_w is the length of the window.

The choice of a window function is important during analysis as the shape and length of the windows can affect the frequency representation of the signal. The frequency response of an ideal window should have a very narrow main lobe and no side lobes. Since, such a realization is not possible, different types of windows are used depending upon the demands of a process. Different windows have been suggested in literature as a compromise between narrow main lobe and smaller side lobes. *Hamming* and *Hanning* windows are popular choices for speech applications.

A Hamming window is defines as

$$w[n] = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{N_w - 1}\right) & 0 \leq n \leq N_w - 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

and a Hanning window is given by

$$w[n] = \begin{cases} 0.50 - 0.50 \cos\left(\frac{2\pi n}{N_w - 1}\right) & 0 \leq n \leq N_w - 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

Other popular window function with different main and side lobe characteristics are the *Bartlett, Blackman and Kaiser* window functions [1]. The length of the window should not be less than twice the smallest pitch period while it should be long enough to capture the dynamics of the speech frame appropriately. Normally a $20 - 30$ msec window is used with a frame update rate of 10 msec. The windowed speech frames are obtained by sliding the window function over the speech signal. In order to reduce the discontinuities, successive speech frames are obtained by overlapping of the windows as shown in Figure 2.4. The analysis window can be placed so as to coincide with the location of the pitch mark, a process known as *Pitch-Synchronous Analysis* as opposed to *Pitch-Asynchronous Analysis* where knowledge of the pitch marks is not required for the processing of the speech signal and the speech signal is analysed in segment having the same length. Each windowed speech frame is further processed to extract the features which are to be used in speech processing systems.

The following section introduces the Linear Prediction analysis and the extraction of *Line Spectral Frequencies* (LSF) from the *Linear Prediction Coefficients* (LPC), which are widely used in speech coding algorithms. This chapter also introduces the *Mel-Frequency Cepstrum Coefficients* (MFCC) and the *Linear Prediction cepstral Coefficients (LPCC)*. MFCC, and LPCC to some extent, are widely used in speaker and speech recognition tasks. In this work MFCC and LPCC are used for generating the

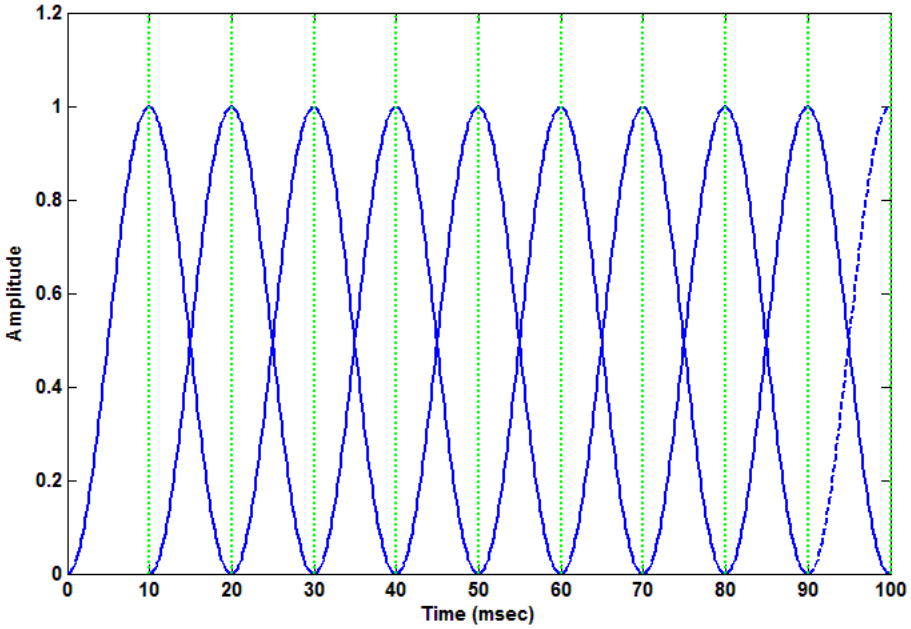


Figure 2.4: Example of Window Placements for Fixed Rate Frame Analysis

speaker models in the recognition system, while LSF are employed as the feature vectors in the voice impersonation system

2.4.3 Linear Prediction Analysis

Linear Prediction

As mentioned in Sec 2.2.1, the vocal tract is model as a linear filter. An approximation of the mathematical form of this filter is given by the system function [17]

$$H(z) = \frac{S(z)}{U(z)} = G\frac{1 - \sum_{k=1}^{q} \beta_j z^{-k}}{1 - \sum_{k=1}^{p} \alpha_j z^{-k}} \tag{2.7}$$

$H(z)$ is the pole-zero model, $S(z)$ and $U(z)$ are the z-transform of the speech and the excitation signals respectively, G represents the gain and the filter coefficients are given as α_j and β_j. Calculation of parameters based on Equation 2.7, require computation of a solution of non-linear equations [18]. Because of the numerical and mathematical difficulties introduced in this form, the all-pole model is preferred over Equation 2.7 for its computational efficiency. The system function of an all-pole model is given as

$$H(z) = \frac{G}{1 - \sum_{j=1}^{p} \alpha_j z^{-j}} \tag{2.8}$$

As mentioned in Section 2.2, there are 4-5 formants in the Nyquist band of the speech

signal and each formant is represented by a pole pair. It is common practice to use 10th order filter to effectively model the vocal tract for speech sampled at 8kHz.

The main purpose of the LP analysis is to calculate the parameters of Equation 2.8. The time-domain expression for the linear prediction estimate $\bar{s}[n]$, of the speech signal $s[n]$ has the form

$$\bar{s}[n] = Gs[n] + \sum_{j=1}^{p} \alpha_j \bar{s}[n-j] \tag{2.9}$$

The term linear prediction is assigned to this model since the current output $\bar{s}[n]$ can be 'predicted' by a weighted sum of the current input and the past p outputs $\bar{s}[n-p], \bar{s}[n-p+1], \ldots, \bar{s}[n-1]$.

Autocorrelation Method

Several techniques have been described in the literature which can be used for the computation of the LP coefficients such as *Autocorrelation, Covariance and Lattice method* [1], The autocorrelation method is most commonly used among others. LP coefficients are calculated from a windowed speech frame comprising of N samples. As mentioned previously, the signal is assumed to be stationary within the frame.

The prediction error $e[n]$, which is also known as the *residual* signal, *of* the all-pole model is given as

$$e[n] = s[n] - \bar{s}[n] = s[n] - \sum_{j=1}^{p} \alpha_j s[n-j] \tag{2.10}$$

The error signal is obtained by filtering the speech signal with the inverse of the prediction filter i.e. $\frac{1}{A(z)}$.

The optimal values of α_j can be obtained by minimizing the average squared prediction error or the energy E of the error signal. E is given as

$$E = \sum_{n=1}^{N} e^2[n] = \sum_{n=1}^{N} \left(s[n] - \sum_{j=1}^{p} \alpha_j s[n-j] \right)^2 \tag{2.11}$$

The values of α_j are computed by setting $\frac{\partial E}{\partial \alpha_j} = 0$ for $j = 1, 2, \ldots, p$. After manipulations, the p optimality equations are obtained as

$$\sum_{n=1}^{N} s[n][n-j] - \sum_{j=1}^{p} \alpha_j \sum_{n=1}^{N} s[n-i]s[n-j] = 0 \ \text{ for } i = 1, 2, \ldots, p. \tag{2.12}$$

The correlation function $R(i)$ is defined as

$$R(i) = \sum_{n=1}^{N} s[n]s[n-i] \tag{2.13}$$

From Equations 2.12 and 2.13

$$\sum_{j=1}^{p} \alpha_j R\left(|j-i|\right) = R(i) \quad \text{for } i = 1, 2, \ldots, p \tag{2.14}$$

In matrix form, Equation 2.14 can be expressed as [5]

$$\begin{bmatrix} R(0) & R(1) & \ldots & R(p-1) \\ R(1) & R(0) & \ldots & R(p-2) \\ R(2) & R(1) & \ldots & R(p-3) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \ldots & R(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix} \tag{2.15}$$

The autocorrelation matrix is Toeplitz in nature, i.e. the matrix is symmetric with identical elements along the diagonal. The solution of Equation 2.15 can be obtained by a well known method known as the *Levinson-Durbin* algorithm [16]. Levinson-Durbin is a recursive algorithm and requires no matrix inversion. There are several other methods that can be used for the optimal computation of LP filter coefficients but the Levinson-Durbin algorithm used with the autocorrelation method is the most widely used of them all. The algorithm is as follows:

$$E^{(0)} = R(0) \tag{2.16}$$

$$k_i = \frac{\left[R(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} R(i-1)\right]}{E^{(i-1)}} \quad 1 \leq i \leq p \tag{2.17}$$

$$\alpha_i^{(i)} = k_i \tag{2.18}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \quad 1 \leq j \leq i-1 \tag{2.19}$$

$$E(i) = (1 - k_i^2) E^{(i-1)} \tag{2.20}$$

The required prediction coefficients α_j are achieved after solving Equations 2.16 to 2.19 recursively. The prediction coefficients are given by

$$\alpha_j = \alpha_j^{(p)} \quad 1 \leq j \leq p \tag{2.21}$$

2.4.3.1 Alternate Representations of LPC

The LPC parameters described in Section 2.4.3, provide an accurate description of the speech spectral envelope. These parameters usually require quantization and interpolation in speech processing applications. The spectral envelope, however, is very sensitive to the variations in the LPC parameters, such as the changes introduced by the quantization process. These changes can cause instability of the LP filter, and no simple procedure exists to check for the stability of the filter based on LPC. It is common practice to use alternate forms of LPC parameters, such as LSF [19], Log Area Ratios (LAR) [20], Reflection Coefficients (RC) [21] etc., which are robust against variations introduced during the quantization process. The LSF are the most popular alternative representation of the LPC parameters. In this work LSF are used in the voice impersonation system (See Chapter 4). In the following section, the computation of LSF from LPC and some of their properties are discussed.

2.4.3.2 Line Spectral Frequencies (LSF)

Due to many desirable properties, LSF has received widespread acceptance in the speech community. In this section the origins of the LSF are explained, their conversion from LPC and their properties are presented.

The p^{th} order all-pole prediction-error filter is given as

$$H(z) = \frac{1}{A(z)} \tag{2.22}$$

where

$$A(z) = 1 - \sum_{j=1}^{p} \alpha_j z^{-1} \tag{2.23}$$

Given an even order p of the LP filter, Equation 2.23 can be written as

$$A(z) = \frac{1}{2}(P(z) + Q(z)) \tag{2.24}$$

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{2.25}$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \tag{2.26}$$

Using Equation 2.23, Equations 2.25 and 2.26 can be written as

or

$$P(z) = z^{-(p+1)} \prod_{j=0}^{p+1} (z - \gamma_j) \tag{2.27}$$

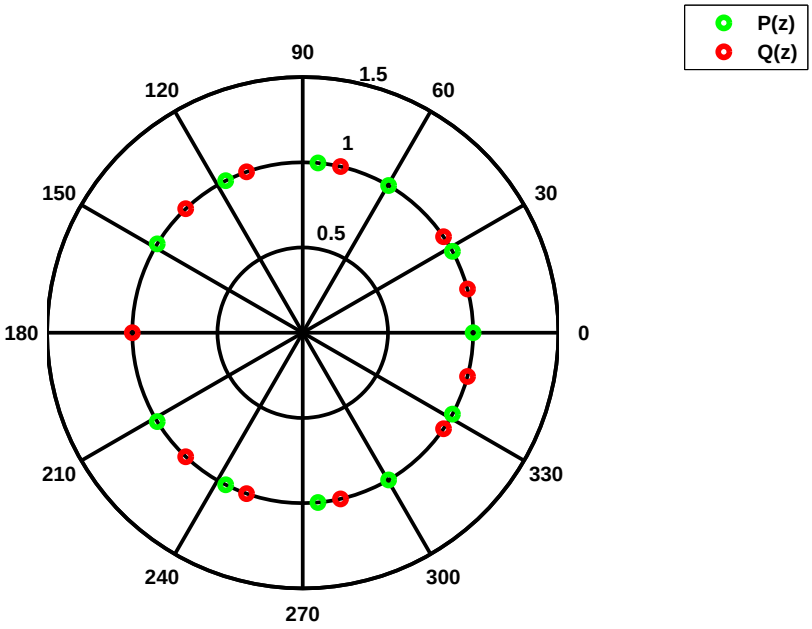


Figure 2.5: $z-$plot of $P(z)$ and $Q(z)$

Similarly

$$Q(z) = z^{-(p+1)} \prod_{j=0}^{p+1} (z - \beta_j) \tag{2.28}$$

There are p unknowns to be computed which are the roots of $P(z)$ and $Q(z)$, namely γ_j and β_j respectively. The roots of $P(z)$ and $Q(z)$ can be computed using methods such as Complex Root Method [1][13], Both γ_j and β_j occur in complex conjugate pairs and lie on the unit circle with the exception of $z^{-1} = -1$ for $P(z)$ and $z^{-1} = 1$ for $Q(z)$ as shown in Figure 2.5. The cosine arguments of these roots are known as *Line Spectral Pairs* (LSP). A unique set of p LSP parameters can describe a stable LP filter. Since the poles lie on the unit circle, the angular information is sufficient to compute the LSPs, using

$$LSP(2i) = \cos(\omega_{Q_i}) \tag{2.29}$$

and

$$LSP(2i+1) = \cos(\omega_{P_i}) \tag{2.30}$$

where $i = 0, 1, \ldots, \frac{p}{2} - 1$ and ω is the frequency associated with the LSF such that $0 \leq \omega \leq \pi$.

LSFs are computed from the LSPs using

$$LSF_i = \frac{\cos^{-1}(LSP_i)}{2\pi T} \tag{2.31}$$

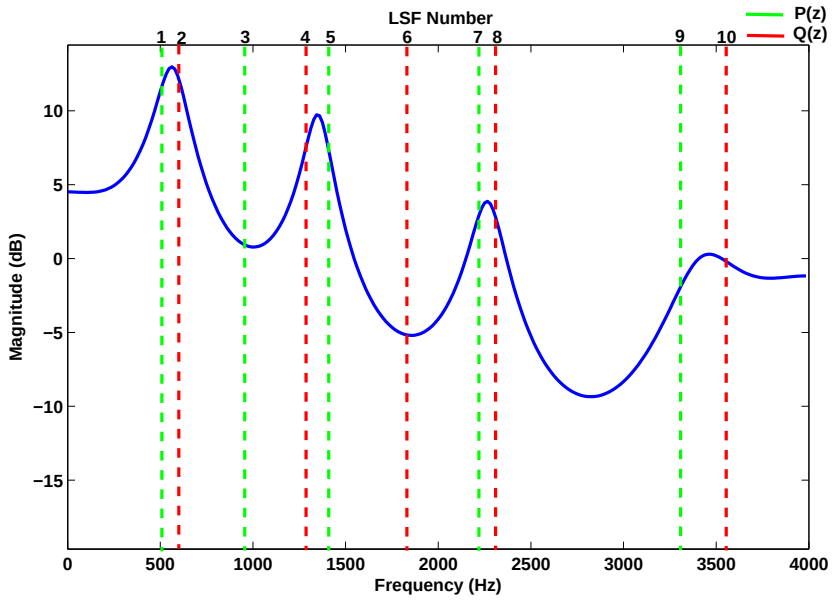


Figure 2.6: LPC spectrum plot with analysis order 10, showing the corresponding values of the LSFs

where T in the above equation is the sampling period.

Properties of LSF

- For a minimum-phase $A(z)$, all zeros of $P(z)$ and $Q(z)$ lie on the unit circle (Figure 2.5), guaranteeing the existence of LSFs for a minimum-phase $A(z)$.
- Fixed range between 0 and 4000 Hz for speech signals sampled at 8000 Hz.
- If $A(z)$ is minimum-phase, the zeros of $P(z)$ and $Q(z)$ are interlaced with each other in ascending order. For a speech signal sampled at 8 kHz, we get:

$$0 < LSF_1 < LSF_2 < LSF_3 < \ldots < LSF_p < 4000 \tag{2.32}$$

  This property can be verified easily and guarantees the stability of the corresponding LPC filter.
- Presence of a formant is indicated by the two closely grouped LSFs cf. Figure 2.6.
- The process of quantization can benefit from the inter-frame and intra-frame correlation among the LSF [22].

2.4.4 Cepstral Analysis

According to the source filter model, Section 2.2.1, speech is the convolution of the excitation signal with the impulse response of the vocal tract function. It is often desirable

to extract the excitation signal and the impulse response from the output of the linear-time varying signal so that these components can be analysed, coded, modelled or used in recognition. Since the excitation and the impulse response of the linear time-variant system are combined through convolution, the problem of separating the constituent signal is often called *Homomorphic Deconvolution* [23]. *Cepstrum Analysis* [24] is a simplified version of homomorphic deconvolution. This section gives a brief description of cepstrum analysis and a comparative analysis with Linear Prediction (Section 2.4.3). Later features based on cepstrum analysis are discussed.

Given a frame of speech data $s[n]$, generated from the convolution of the vocal tract impulse response $v[n]$ and the excitation sequence $x[n]$

$$s[n] = v[n] * x[n] \tag{2.33}$$

the cepstrum $\hat{c}[n]$ is calculated by determining the inverse Fourier transform of the logarithm of the Fourier transform of $s[n]$ [6]:

$$\hat{c}[n] = \mathscr{F}^{-1}\{\log(\mathscr{F}(s[n]))\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{S}(\omega)e^{j\omega n}d\omega \tag{2.34}$$

where

$$\hat{S}(\omega) = \log|S(\omega)| + j\arg[S(\omega)] \tag{2.35}$$

i.e. $\hat{S}(\omega)$ is the complex logarithm [6]of $S(\omega)$, the Fourier transform of $s[n]$.

If the phase angle is a continuous odd function of ω, the problem of phase uniqueness can be solved in Equation 2.34 [25]. The cepstrum from Equation 2.34 is known as the *complex cepstrum*. Although retaining the phase (or *saphe* [24]) bestows certain advantages, it is however, difficult to compute in practice and hence a *real-cepstrum* is defined as [6]:

$$c[n] = \mathscr{F}^{-1}\{\log(|\mathscr{F}(s[n])|)\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|S(\omega)|e^{j\omega n}d\omega \tag{2.36}$$

where,

$$\log|S(\omega)| = \log|V(\omega)| + \log|X(\omega)| \tag{2.37}$$

i.e. $V(\omega)$ and $X(\omega)$ are additive. Figure 2.7, shows a block level view of cepstrum analysis process.

By calculating the spectrum of the log spectrum, the vocal tract spectral envelope will appear as a low frequency component while the excitation would manifest itself as a high frequency ripple in pseudo-time, the *Quefrency.* Hence, it is possible to separate the effects of the vocal tract and excitation signals.

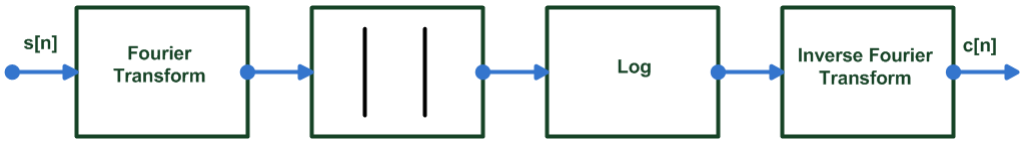


Figure 2.7: Extraction of cepstral coefficients from the speech signal $s[n]$

One the most important applications of cepstrum analysis is in the representation of the LP model. Here the signal under consideration is minimum-phase, in which case the real-cepstrum of Equation 2.36 is equal to the real part of the complex cepstrum of Equation 2.34 [8] and is therefore preferred over complex cepstrum for efficient computation.

2.4.4.1 Comparison with the LP Analysis

The spectral envelope related to the vocal tract may be obtained by multiplying $c[n]$ by a window function, also called a *lifter,* of unit height and long enough to encompass all the low frequencies pertaining to the vocal tract. The exact length of the lifter depend on the amount of detail required for the application and thus is chosen empirically.

Analysing Figure 2.8, the spectral envelopes generated from both the LP modelling (Section 2.4.3) and cepstrum analysis (Section 2.4.4) model the voiced speech spectrum well but the cepstrum generated envelopes model the spectral nulls more accurately and efficiently than the LP envelope specially in the $0-2$ kHz range. This is as expected because the cepstrum analysis does not make any assumptions about the all-pole nature of source filter and as such the cepstrum contains both poles and zeros in the analysis of a voiced frame of speech rather than just the poles as in the LP analysis.

2.4.4.2 Linear Predictive Cepstral Coefficients (LPCC)

Linear Predictive Cepstral Coefficients (LPCC) are a representation of LPC in the cepstral domain. The computation of the LPCC is a two-step process. The first steps involves determining of the LPC from speech by mathematical modelling according to the source filter theory. The process of LPC computation was given in Section 2.4.3.

Once the LPC have been computed the next step is the estimation of the cepstral coefficients. The linear prediction derived cepstral coefficients are obtained by considering the power series expansion of $\ln(H(z))$, where $H(z)$ is given in Equation 2.8. The log-transfer function in term of powers of z^{-1} is given as [26]:

$$\ln(H(z)) = C(z) = \sum_{n=1}^{\infty} c_n z^{-1} \tag{2.38}$$

where $z = \exp(j\omega t)$, $\omega =$ frequency in radians, $T =$ sampling interval and $c_n =$ amplitude of the inverse fourier transform at the n^{th}sampling instant.

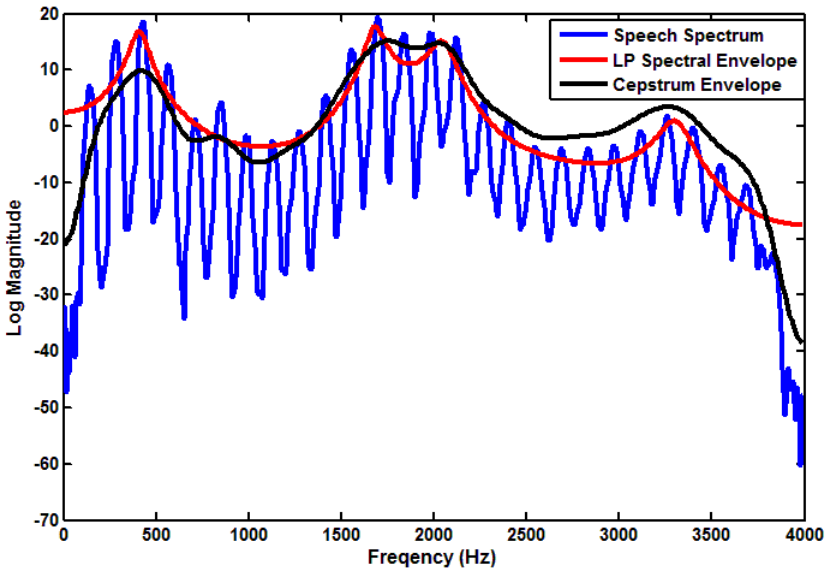


Figure 2.8: Comparative analysis of the Cepstrum and LPC spectral envelopes on a voiced segment of speech

To derive the relationship between LPC and LPCC, Equation 2.8 is substituted in Equation 2.38 and differentiated w.r.t. z^{-1}

$$\frac{d}{dz^{-1}}\left(\ln\left[\frac{1}{1-\sum_{k=1}^{p}\alpha_k z^{-k}}\right]\right) = \frac{d}{dz^{-1}}\sum_{n=1}^{\infty} c_n z^{-n} \tag{2.39}$$

simplifying

$$\frac{\sum_{k=1}^{p} k\alpha_k z^{-k+1}}{1-\sum_{k=1}^{p}\alpha_k z^{-k}} = \sum_{n=1}^{\infty} n c_n z^{-n+1} \tag{2.40}$$

rewriting

$$\sum_{k=1}^{p} k\alpha_k z^{-k+1} = \left(1-\sum_{k=1}^{p}\alpha_k z^{-k}\right)\left(\sum_{n=1}^{\infty} n c_n^{-n+1}\right) \tag{2.41}$$

equating the constants and powers of z^{-1}on both sides of Equation 2.41 gives the desired expression of the relationship between α_k's and c_n's as

$$c_n = \begin{cases} \alpha_1 & n = 1 \\ \sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\alpha_k c_{n-k} + \alpha_n & 1 < n < p \\ \sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\alpha_k c_{n-k}, & n > p \end{cases} \tag{2.42}$$

Equation 2.42 allows the computation of coefficients c_n from the p predictor coefficients. c_n can be regarded as the samples of the cepstrum function. Traditionally the cepstrum is obtained by the inverse Fourier transform of the impulse response h_n, for an all-pole

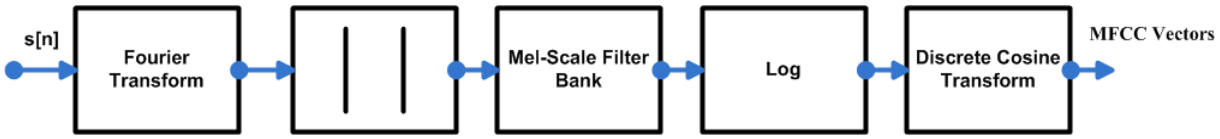


Figure 2.9: Extraction of MFCC vectors from the speech signal $s[n]$ using the mel-scale filter banks [2]

transfer function the cepstrum can be obtained from the impulse response using [26]:

$$c_n = \begin{cases} h_1 & n = 1 \\ \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) h_k c_{n-k} + h_{k,} & n > 1 \end{cases} \tag{2.43}$$

Although LPCCs benefit from the computational efficiency, they do however inherit the same assumptions about the all-pole nature of the source-filter. Similar to the LPC parameters the spectral null cannot be represented by the LPCC efficiently. Next MFCC extraction procedure is discussed which is widely used in the speech community.

2.4.4.3 Mel-Frequency Cepstrum Coefficients (MFCC)

The MFCCs are a popular choice in speaker and speech recognition applications. Mel cepstrum-filter bank is based on the perception of the human ear to the frequencies of sound, which is non-linear [27]. The filter bank is designed in a way to exploit the fact that the human ear perceives the phonetic component in the lower frequencies to be more important than in the higher frequencies [28]. The frequency resolution of the mel-scale reduces with the increase in the frequency and as such places less emphasis on the higher end of the spectrum. A block level diagram of MFCC computation is shown in Figure 2.9

The first step is the computation of the Fourier transform, $S[k]$, of the input speech sequence $s[n]$.

$$S[k] = \sum_{n=1}^{N-1} s[n] e^{-j\frac{2\pi kn}{N}} \tag{2.44}$$

where N is the number of samples of the speech frame (length of Fourier transform). The power spectrum is computed as $|S[k]|^2$ for $0 \leq k < \frac{N}{2}$, as the magnitude square of Equation 2.44, which is computed after zero padding the speech frame to twice its length to improve the frequency resolution.

The power spectrum is transformed from frequency domain to Mel-scale to emphasize the low frequency regions compared to the high frequencies. The power spectrum is multiplied by the frequency response of the Mel-scale filter. As mentioned above the filter-banks are based on the perception of sounds to the human ears. The bandwidth of these filters is also known as the *critical bands of hearing* [5]. The Mel-scale filter banks provide a mapping of linear frequencies to a representation corresponding to the

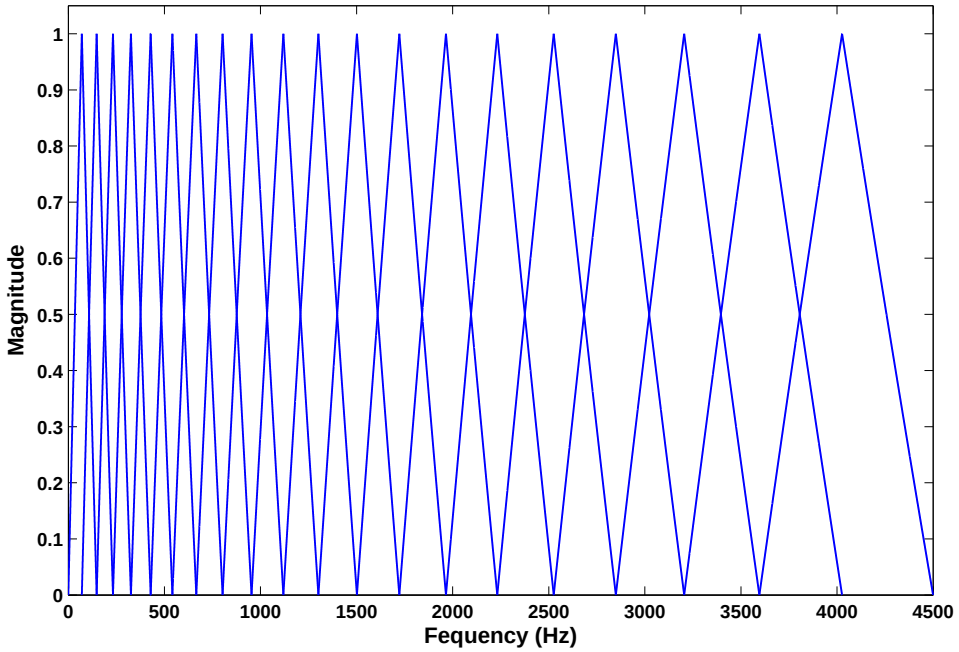


Figure 2.10: Triangular filter-banks based on the mel-scale[2]

critical bands. The filter bank consists of overlapping triangular or Hanning filters, the former being used commonly. Each filter's cut-off frequencies are determined by the centre frequencies of the adjacent filters. The frequency range of $0 - 1$ kHz is covered by 10 overlapping bands which are spaced linearly, while bands covering $1 - 4$ kHz are placed logarithmically, with logarithmically increasing bandwidths. Figure 2.10 is a graphical representation of the overlapping triangular filter bank frequency response for a 4 kHz spectrum.

In this study, different sets of filter banks defined in [29], were used. Table 2.1 lists the beginning, centre and end frequencies of the critical bands. The bandwidth of the filters is defined by the centre frequencies of the adjacent bands.

The human auditory system resolves the audio frequencies non-linearly across the spectrum. Using Mel-scale or any other filter bank with similar properties, the non-linear frequency resolution can be achieved. However, for speech and speaker recognition tasks, the design and shape of the filter banks is insignificant [6].

The energy output of each filter is calculated according to

$$E_j = \sum_{k=0}^{K-1} \phi_j(k) \left| S[k] \right|^2 \quad \text{for } 0 \leq j < J \tag{2.45}$$

where

$$\sum_{k=0}^{K-1} \phi_j(k) = \sum_{k=0}^{K-1} \left| V_j(k) \right|^2 = 1 \ , \forall j \tag{2.46}$$

Index	Lower cut-off frequency (Hz)	Centre Frequency (Hz)	Upper cut-off frequency (Hz)
1	0	100	200
2	100	200	300
3	200	300	400
4	300	400	500
5	400	500	600
6	500	600	700
7	600	700	800
8	700	800	900
9	800	900	1000
10	900	1000	1149
11	1000	1149	1320
12	1149	1320	1516
13	1320	1516	1741
14	1516	1741	2000
15	1741	2000	2297
16	2000	2297	2639
17	2297	2639	3031
18	2639	3031	3482
19	3031	3482	4000
20	3482	4000	4595
21	4000	4595	5278
22	4595	5278	6063
23	5278	6063	6964
24	6063	6964	8000

Table 2.1: Lower and upper cut-off frequencies of the mel-scale filter banks with the corresponding centre frequencies [2]

Here J represents the number of filters used and $V_j(k)$ is the frequency response of the j^{th} filter under consideration. The output of the filters is normalized to account for the differences in the bandwidths. Calculation of energy is followed by the application of natural logarithm. The energy coefficients represent the spectral envelope and also help to reduce the amount of data per frame that needs to be processed. The reduction in the amount of data, without compromising the performance of the system is known as *dimensionality reduction* [29].

The computation of log energy coefficients is preceded by the application of *Discrete Cosine Transform* (DCT) as the last step in the computation of the MFCC. The DCT produces decorrelated log-energy coefficients. These coefficients are particularly useful in speaker modelling by *Gaussian Mixture Models* (GMM) (cf. Chapter 3) since diagonal covariance matrices instead of full covariance matrices can be used [5].

The MFCCs are computed from the log-energy coefficients using:

$$c_m = \frac{1}{J}\sum_{j=1}^{J}\cos\left(m\frac{\pi}{J}(j-0.5)\right)\log(E_j) \quad , \; 0 \leq m \leq M \tag{2.47}$$

where $c_m = [c_0, c_1, \ldots, c_M]$ represent the $M+1$ MFCC coefficients, and J is the number of filters in the filter bank. In this study, a value of $M = 16$ was used in the speaker recognition tasks for speech signals sampled at 8 kHz. The coefficients c_0 is the average log-energy of the speech spectrum, corresponding to the intensity of the speech signal and the background noise and is usually not used in the set of feature vectors.

Speaker recognition system employing cepstral features can also benefit from the inclusion of the *delta* and *delta-delta* (also known as the velocity and acceleration) cepstrum coefficients. The delta and delta-delta coefficients are simply the first and second differences of the cepstral coefficients and provide the temporal information about the changing dynamics of the vocal tract. These features are concatenated to the MFCC feature vector to form a longer feature vector and this approach has been shown to improve the performance of the speech and speaker recognition systems [30].

The different parameter sets discussed so far include the LPC, LSF, LPCC and the MFCC. A large number of additional parameters can be computed from a linear transformation of any of these parameters. However the distance between two points in the multidimensional space can be made irrelevant to these linear transformations by a proper choice of distance metric [26]. Therefore, if the decision criterion is based on distance calculations between a reference and test pattern, use of linear transformation or otherwise, is immaterial. Thus, as far as the recognition performance is concerned, the feature sets which can be computed from each other by means of a linear transformation can be regarded as equivalent.

2.4.4.4 Post-Processing

An important factor that affects the performance of speech and speaker recognition systems is the presence of convolution distortion in the speech such as the distortions introduced by the microphone transfer functions and the transmission medium. When cepstral coefficients are used as the feature vectors, the linear convolution distortion becomes additive components on the cepstral vectors. The *Cepstral Mean Subtraction* (CMS) remove the stationary convolutional distortion [31, 32]. In CMS, the population mean is subtracted from each observable feature vector to remove the stationary distortions introduced by the telephone channel while *RelAtive SpecTrAl* (RASTA) processes removes the time-varying distortions from the speech signal [32]. Both CMS and RASTA are used commonly to remove the convolutional distortions introduced in the speech signal. The noise integration model [33], is another method that that generates the speech as well as noise model which are then used directly in the speaker recognition system. Score normalization techniques can also be applied at the test stage of a recognition system to minimize the affect of mismatched training and test data [34].

The next section describes some of the distance measures that are used in pattern matching application including the speaker recognition and voice impersonation tasks carried out in this study.

2.5 Distance Measures

In pattern matching applications, the differences or similarities between different sets of features can be computed by means of various distance measures. Different types of distance measures have been proposed in the literature [6, 35]. Since the output of each distance measure is different from the other, the choice of a particular distance measure depends upon the task at hand. The choice of a distance measure can be based on some minimization criterion of an error function or on the results of classification. Some of the well-known distance measure, between two arbitrary points x_k and y_k for $k = 1, 2, \ldots, K$, are listed below:

Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{K} (x_k - y_k)^2} \tag{2.48}$$

Mean Squared Error

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{P} (\mathbf{x} - \mathbf{y}) (\mathbf{x} - \mathbf{y})^T = \frac{1}{P} \sum_{k=1}^{K} (x_k - y_k)^2 \tag{2.49}$$

Manhattan Distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} |x_k - y_k| \tag{2.50}$$

Likelihood Ratio Distortion

$$d_{LR}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{y^T R_a y}}{\mathbf{x^T R_a x}} - 1 \tag{2.51}$$

In the context of speaker recognition, $\mathbf{x}$ and $\mathbf{y}$ represent the LPC parameters while $\mathbf{R_a}$ represents the toeplitz autocorrelation matrix [36].

Log-likelihood Distance

$$d_{LLR}(\mathbf{x}, \mathbf{y}) = \log\left(d_{LR}(\mathbf{x}, \mathbf{y})\right) \tag{2.52}$$

Weighted Cepstral Distance

$$d_w(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{K} \left[\omega_k (x_k - y_k)\right]^2} \tag{2.53}$$

here ω_k represents the cepstral weighting function [35].

2.6 Summary

In this chapter the human sound production mechanism along with the functions of different organs was presented. The all-pole filter method was introduced, describing a mathematical model for the production of synthetic speech. The front-end of speaker modelling systems has been explained in some detail. A typical front end includes a preprocessing stage, feature extraction which is followed by an enhancement or post-processing part. The preprocessing stage uses some signal processing techniques to prepare the speech signal for further processing. The feature extraction stage produces features which reduce the amount of speech data to be processed and provides the desired information about the speaker related characteristics. Features based on linear prediction and cepstrum analysis techniques have been presented. Extraction of LPC from a speech waveform and its commonly used alternative LSF, have been explained in detail. Cepstrum based features MFCC and LPCC have also been discussed while highlighting certain differences among these representations. In this study MFCC and LPCC are used in the speaker recognition system while LSF are employed as features

representing the speaker characteristics in the impersonation system. In order to enhance the quality of the extracted features, some post-processing techniques commonly used, were also presented for the removal of static and dynamic channel noise. In the last section some of the distance measures used in the study to measure similarities or differences among features were listed. To achieve high performance on a speaker modelling system, whether it is a speaker recognition system or a voice impersonation system, it is vital to have a front-end processing unit that produces high quality features, providing an effective and efficient representation of speaker characteristics.

Chapter 3

Speaker Modelling and Recognition

3.1 Introduction

In Chapter 2, the process of feature extraction was presented. Sequences of feature vectors were obtained from the speech signal, characterizing the properties of the speaker's voice. This chapter introduces the classification process, which utilizes the features extracted from a speaker's voice to determine the identity of the speaker. The process of classification is a two stage process, namely training and testing. During the training phase, the recognition system enrols the speakers by building a specific model for each individual speaker from the features extracted from their voice samples. During the testing stage, features extracted from a claimant's speech signal are matched against the stored models by calculating an utterance score through the use of a distance measure to determine the correspondence between the speaker models and the test utterance. The chapter is finalized by describing the simulation set up used in the text-independent speaker identification and verification baseline systems. A summary of the recognition performances for both the identification and verification systems is detailed towards the end of this chapter.

3.2 Speaker Recognition

Among all the biometric identification methods, voice as a biometric has its own unique standing among other biometrics. Voice production is a natural process and as such is non-invasive. Speech can be transmitted easily through the existing communications networks without the aid of additional transmission media, thereby allowing remote authentication. It can be acquired through simple devices such as a microphone and

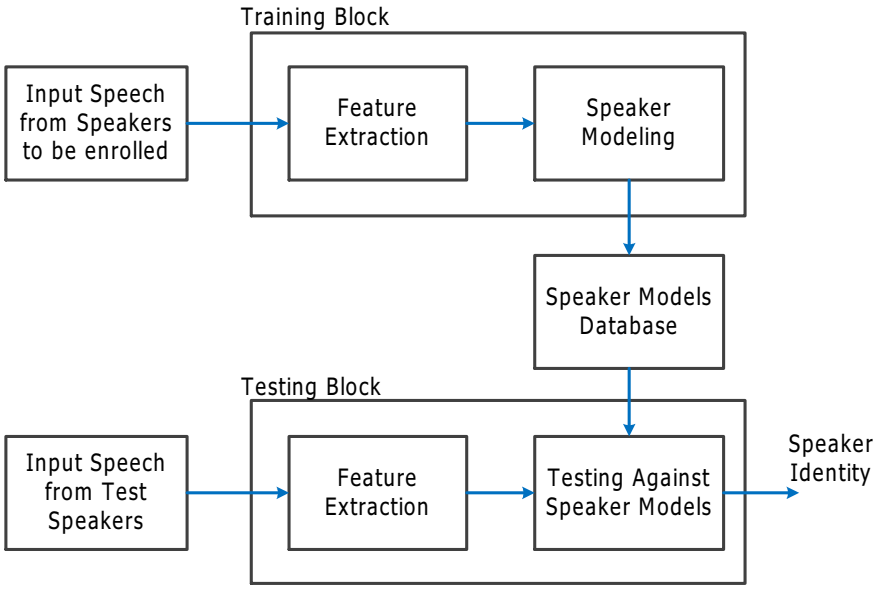


Figure 3.1: Block diagram of a speaker identification system showing the main components of the training and the testing phases.

recognition processing can be carried out by means of a computer. The speech signal not only conveys the spoken message, but also the emotions as well as the identity of the speaker. Speaker recognition is the process of identifying the originator of a speech signal or in simple words, who is speaking [31][37]. The three main processes in a typical speaker recognition system include the feature extraction, training and testing. Feature extraction is common to both training and testing. A block diagram of a typical speaker recognition system is shown in Figure 3.1.

Text-Dependent vs. Text-Independent Systems

Text-dependent speaker recognition systems use known text for both training and testing process. While text-independent recognition systems allow users to speak freely, using any text, for training and testing purposes. For text-independent speaker recognition systems the full range of speaker's vocal sounds should be used for the training process. For limited amount of training data, the text-independent systems provide better recognition performance as compared to the text-dependent recognition systems as the enrolment and testing process is not dependent on the already known text. The text-independent systems can be used in areas where limited amount of speech is available or in the law enforcement areas where the individuals are not very co-operative.

Speaker Identification Systems

Speaker identification is the process of identifying an unknown individual from a group of known speakers. The speaker identification systems can be further classified as closed-set and open-set systems.

Closed-Set Identification

In a closed-set speaker identification system, it is already decided that the unknown speaker is a part of the database and is one of the enrolled speakers. The identification has to decide which of enrolled speaker model is a best match for the feature vectors of the unknown speaker.

Open-Set Identification

In an open-set identification system, the unknown speaker is not considered to be one of the enrolled speaker a priori, which means that the speaker identification system has to decide whether the unknown speaker is a part of the known group of speaker or is an impostor. If the system fails to find a match for the unknown speaker, their claim is rejected and the unknown speaker is considered an impostor. On the other hand if a match exists then the next stage is to determine the actual identity of the speaker.

Speaker Verification Systems

Speaker verification systems, as the name suggests, verify if the unknown speaker is in fact who he/she claims to be. The decision criteria surrounding the acceptance or rejection of an identity claim must be chosen carefully so as to reduce Type-I and Type-II errors in speaker recognition systems. A type-I error, also known as False Rejection (FA), occurs when the verification system rejects speech from a speaker who is enrolled in the system. A type-II error results when the system fails to reject the speech from an impostor, in what is known as a False Acceptance (FA).

Speaker Identification and Speaker Verification systems will be discussed in more detail in the later sections of this chapter.

3.2.1 Applications

A speaker identification system aims to find the best match for the unknown speaker from a database of known speaker models. As such the application areas of speaker identification include law enforcement e.g. determining the identity of a suspect from a threat call or identifying a potential criminal from their voice.

The application areas of speaker verification systems are mainly security applications that allow access to certain facilities or services only to the authorized users. Another potential application is in monitoring the locations of prison inmates and controlling their presence in specified areas.

The use of text-dependent speaker recognition systems can further enhance the authorization process, where the unknown speaker to required to speak a particular text

that is known to the system and the speaker such as a password or a PIN. In areas where the users of the system do not have access to the text prompting devices or in situations where a suspect is unwilling to co-operate with the law-enforcement agencies a text-independent system can be implemented. A speaker recognition system can be used for national border control to monitor the movements of individuals in and out of the country, in personalizing the content of an entertainment source depending upon the preferences of the identified individual or in gaming for providing rich personalized interactive game play [38, 39].

3.2.2 Performance Evaluations

A number of factors affect the performance of a speaker recognition system including the number of speakers used for training and testing, the amount of speech material available for training and testing as well as the overall quality of the speech samples used. The performance of a speaker recognition system is evaluated in term of recognition rates, which are simply the percentage of the correctly identified speakers or error rates demonstrated as Equal Error Rate (EER) values [40] or as Detection Error Trade-off (DET) curves [41].

3.3 Speaker Modelling

The features extracted from a speaker's utterances are used to train a model during the enrolment or training phase. Once all the required speakers are enrolled, a database of the known speakers is created. An individual claiming to be a part of the group known to the database would have features extracted from their speech utterances and a similarity or difference score, depending on the type of the modelling technique used, will be calculated against all the enrolled models. The best matching model is recognized as the identity of the claimant. There are different types of methods used for modelling and testing. These methods can primarily be divided into two sub-groups: Parametric or Stochastic and Non-Parametric or Template based. Parametric methods assume a structure to characterize the parameters or, in other words, the data can be represented by a defined distribution. Non-parametric methods on the other hand make minimal assumptions about the probability density function of the parameters [7, 42].

Some of the common methods of modelling employed in the speaker recognition algorithms are discussed below.

3.4 Non-Parametric Methods

Non-parametric methods avoid making assumptions about the nature of the data and try to learn the distribution from the data itself. Without the parametric assumptions, these methods require considerably more data to approximate the optimal distribution as compared to the parametric method which fit data to a restricted parametric model [43]. Some of the more common methods used in speaker recognition systems are discussed below.

3.4.1 Support Vector Machines (SVM)

Support Vector Machines (SVM) are binary classifier systems that use a hypothesis space of linear functions in high-dimensional feature space [44]. The data under inspection is mapped onto a higher-dimensional feature space by means of non-linear mapping. Classification is performed by constructing hyperplanes to separate the data belonging to different classes [44, 3]. The SVM classifier is obtained by a sum of kernel function $K(.,.)$ given as

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(\mathbf{x}, \mathbf{x_i}) + b \tag{3.1}$$

where $\mathbf{x_i}$ represent the support vectors, N is the number of support vectors, α_i and b are the solutions of the quadratic programming problem, t_i is a target value for each support vector and can take values of -1 and $+1$ depending upon the class that the support vectors belongs to, $\alpha_i \geq 0 \,\text{for}\, i = 1 \,\text{and}\, \sum_{i=1}^{N} \alpha_i t_i = 0$. The classification decision is made by comparing the value of $f(x)$ to a threshold. Even, though SVMs are linear classifiers, they can be used for non-linear data separation by the help of kernel function. In order to achieve better separation among data with non-linear boundaries the input space is mapped onto a higher-dimensional space, called the *feature space.* The choice of a kernel function for a particular application is a decision that requires utmost attention when designing classifiers based on SVMs. Some of the simpler kernel functions used in the literature are the dot product, polynomial kernels, and the Radial Basis Function (RBF) [45, 3].

For a real-valued function $K(x_1, x_2)$ to fulfil the *Mercer's Condition*, the following must be satisfied

$$\int K(x_1, x_2)\, g(x_1)\, g(x_2)\, dx_1 dx_2 \geq 0$$

for a square integrable function $g(x)$ i.e. $\int g(x)^2\, dx$ is finite.

The two-class data to be classified is assumed to be separable and there is a linear class boundary. The SVM algorithm classifies the data by locating the maximal margin hyperplane to classify the data belonging to the two classes [3]. Figure 3.2 shows an

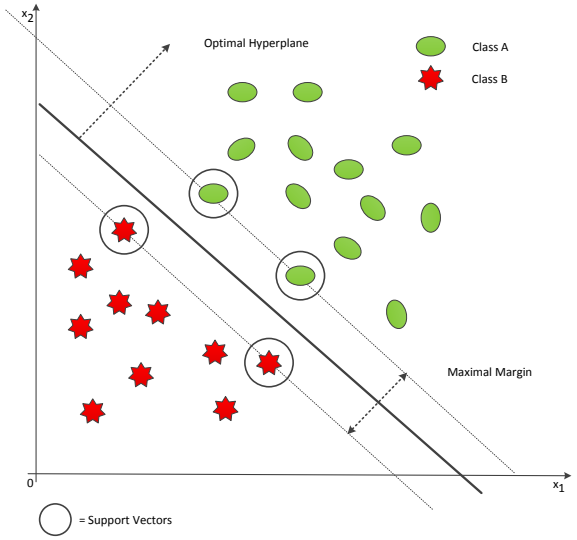


Figure 3.2: Optimal separating hyperplane in two-dimensional space demonstrating the classification criteria for SVM [3]

optimal hyperplane separating a two-dimensional space with maximal margin. Maximal Margin is the hyperplane that can segregate two data clusters and lies in the middle of the two clusters.

Application areas employing SVM include speaker recognition [44], face recognition, handwritten digit recognition and language recognition to name a few [45][46]. In [47][48] SVM have been used directly for speaker recognition. GMM-SVM hybrid classifiers have also been shown to have promising recognition performance [49][50][51]. In these systems, SVMs are used to separate and classify the likelihood values of the client and impostor speaker.

More information on SVMs can be obtained from [44][3].

3.4.2 Neural Networks (NN)

Neural Networks (NN), also known as Artificial Neural Networks (ANN), are systems modelled on the basis of the neural architecture associated with the human brain structure [52][53]. A typical NN can contain any number of layers of units called *neurons.* The neurons in each layer are connected via weights. The value of the weights is adjusted during the training phase.

Neurons are tasked with the following operations:

- Receive inputs from input sources.
- Calculating the weighted sum of the inputs by combining them.
- Perform a non-linear operation on the previous result.

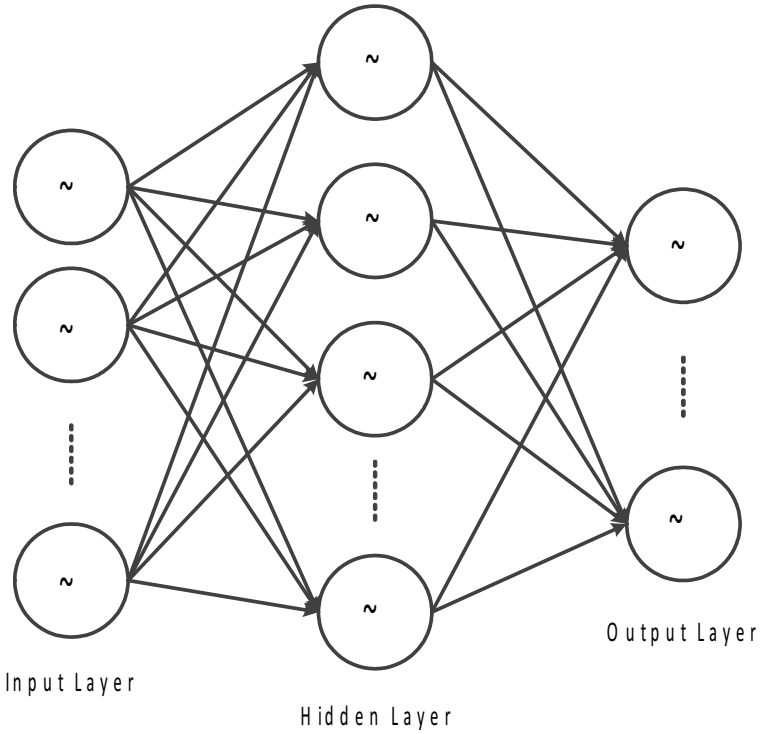


Figure 3.3: A Typical Neural Network Architecture

- Based on the calculated weights, produce outputs.

Non-linearity in the data can be modelled by NNs for a better representation of the data [6]. A typical NN consists of an input layer of neurons for accepting the inputs, one or more hidden layers for combining the inputs and calculating the weights, and an output layer for generating the output as a weighted combination of the outputs from the hidden layer(s). Figure 3.3 shows a typical NN architecture. NNs can also be applied in areas of data clustering and classification and have been applied successfully used in speech and speaker recognition systems.

In a speaker recognition system, each speaker might be represented by a separate NN. Training of the NN is performed by an adjustment of network weights so that each output value of the NN is 1 for the speaker that it is modelling [54] and an output 0 for any other speaker. Similar to any other speaker modelling technique, the identity of the claimant is decided by the NN that produces the highest score for the input speech from the unknown speaker. During speaker verification stage the output of the selected NN as a result of the identification stage is compared against a predetermined trained threshold and a decision of acceptance or rejection is made. A large NN has also been used for all the enrolled speaker in a speaker recognition system in [54]. The training and testing phase of such a NN is the same as the system employing one NN per speaker.

NNs offer certain advantages in pattern matching and classification systems such as the ability to model non-linearity in the data and the capacity of adaptive learning by virtue of a flexible structure [52][6]. During the training stage the weights are constantly

updated and with each update the outputs are recalculated. This process is repeated a number of times until the desired output is achieved. NNs are computationally expensive. The long training times are a major disadvantage of NNs. Factors such as the number of hidden layers and the number of neurons in each layer are some of the factors play their part in the increased computational costs and training times associated with NN. It has been reported in [55] that the NNs are limited in their performance compared to other parametric approaches. [52][39][53] are some further sources of information about NN.

3.4.3 Vector Quantization (VQ)

VQ is a lossy data compression method that aims to divide the given data into non-overlapping clusters. The centre of each cluster is called the centroid, which is the mean value of all the data vectors belonging to that cluster. Each data cluster is represented by its centroid vector in subsequent processing.

The process of quantization is a method of limiting the infinite range of the sampled data vectors to a finite set. This finite set consists of the centroids of the clusters. The VQ can be used in speaker recognition systems with the aid of VQ codebooks which cluster the data vectors from a speaker by a finite set of representative feature vectors. For each speaker to be enrolled in the speaker recognition system, the feature data vectors are represented by a speaker-specific VQ codebook, which divides the feature vector space of the speaker into non-overlapping clusters. Each feature vector is represented by the centroid of its associated cluster, reducing the actual number of feature vectors to be processed thereby reducing the complexity of the system. Different clustering algorithm are used in literature, with *k-means* [56] and Linde, Buzo and Gray (*LBG*) [57] among the most popular. The LGB algorithm minimizes the weighted mean square error during clustering analysis while performing quantization over the feature vectors involved in training. The codebook used in quantization process using the LBG algorithm is described as follows [57]:

1. Initialization: The codebook design procedure is initialized by calculating the mean value of the N vectors seen in training. The mean value represents the code-vector or the centroid of the training vectors. $C_1(0)$, which is the code-vector of the first codebook is given as

$$C_1(0) = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{3.2}$$

   where x_n is the n^{th} vector in the training. This is the design stage $M = 1$.

2. Splitting: Each codevector in the codebook is split into two. A small perturbation value of ϵ is used to rearrange the codevectors so that the new codebook C_{M+1}

is given by

$$C_{M+1} = (1+\epsilon)\, C_M\,(k) \tag{3.3}$$

$$C_{M+1}\left(2^{M-1}+k\right) = (1-\epsilon)\, C_M\,(k) \tag{3.4}$$

where $k = 1, 2, \ldots, 2^{M+1}$ and $\epsilon < 1$. The value of the M is incremented by 1.

3. Optimization: The splitting stage is followed by optimization which is a two step process:

   - Partitioning: Each codevector is assigned to a codevector $C_M\,(k)$ from the codebook, which minimizes the distortion $||x_n - C_M\,(k)\,||$, where $||.||$ is the norm.
   - Updating: The codebook entries are updated by calculating the mean of the training vectors belonging to a cluster, reducing the quantization error in each of the clusters.

   The optimization process is repeated many times until the average distortion within the cluster is below a predefined threshold.

4. Steps 2 and 3 are repeated until the codebook is populated with the desired number of codevectors.

During the testing phase, the features extracted from an unknown speaker are matched against the codebook entries of all the enrolled speakers. The identity of the speaker is taken as the codebook that generates the minimum accumulated distortion. Because of the non-overlapping nature of the codebooks, each input feature vector is assigned to only one class. This restriction can result in performance degradation in speaker recognition systems based on codebooks. The codebook method can be used for both text independent/dependent speaker identification and recognition.

3.5 Parametric Methods

Parametric methods consist of models which assume a structure characterized by parameters. Parametric methods assume that the given data can fit a statistical distribution. The parameters of the distribution can in turn be adjusted to fit the data. These assumption result in faster computation times as compared to non-parametric methods mentioned in Section 3.4. This section discusses some the commonly used parametric methods in speaker modelling with more emphasis on Gaussian Mixture Modelling as it is the main modelling technique used in this study.

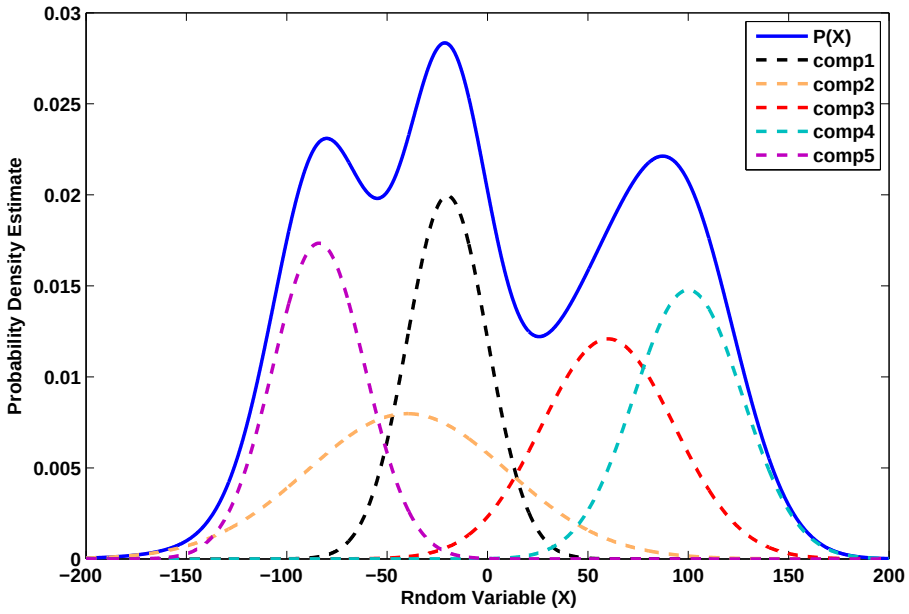


Figure 3.4: Example of Density Modelling by a 5 component 1-D GMM

3.5.1 Gaussian Mixture Models (GMM)

As we speak, various factors including the vocal tract shape, glottal flow, fluid and anatomical dynamical variations influence the manner in which we produce speech [5]. All these factors affect the speech production and cause the speech to be non-deterministic in nature, which can be modelled by the GMM. The probability density functions of the multidimensional Gaussian distribution can be used to represent the speaker-specific spectral characteristics [58][14][55]. GMM can model any distribution, meaning GMM do not impose any restriction on the type of distribution it can model. Figure 3.4 illustrates the process of density modelling with GMM. While used in speech or speaker processing system each component of the GMM models some broad sound class and contains information about speaker-specific vocal tract anatomy [55]. A GMM can contain any number of components to model the data. A probabilistic model is generated by the GMM for the set of sounds a speaker can produce. The remainder of this section details the GMM model description, training of the model parameters and their use in a speaker recognition system.

3.5.1.1 Model Description

In models based on the GMM, each speaker is represented by a separate model λ_s. Each λ_s includes the probability density parameters namely μ_m^s represent the mean vector of the component m, Σ_m^s is the covariance matrix and p_m are the component weights:

$$\lambda_s = \{p_m^s, \mu_m^s, \Sigma_m^s\}, \quad m = 1, 2, \ldots, M \tag{3.5}$$

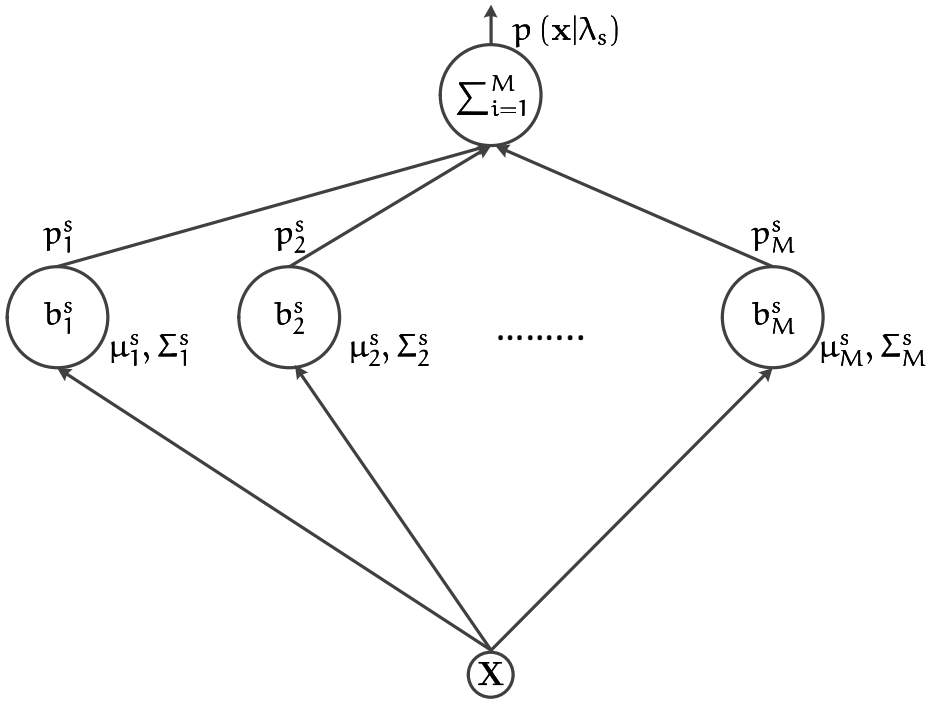


Figure 3.5: An M Component Gaussian Mixture Density

where M is the number of components in the mixture and s represents a speaker from the S enrolled speakers. In a speaker recognition system employing GMMs, the feature vectors extracted from the input speech of the enrolled speakers are modelled by the Gaussian mixture densities while each mixture model represents the speaker. The GMMs are computed as a weighted sum of mixture component densities i.e:

$$p\left(\mathbf{x}|\lambda_s\right) = \sum_{m=1}^{M} p_m^s b_m^s\left(\mathbf{x}\right) \tag{3.6}$$

where $\mathbf{x}$ is a multidimensional feature vector, b_m^s represent the component densities, p_m are the mixture weights, M is the number of components in the mixture where $m = 1, 2, \ldots, M$ and s represents one of the enrolled S speakers. The process is depicted in Figure 3.5

The component Gaussian densities b_m^s are given as:

$$b_m^s\left(\mathbf{x}\right) = \frac{1}{\left(2\pi\right)^{D/2}\left|\Sigma_m^s\right|^{1/2}} \exp\left\{-\frac{\left(\mathbf{x}-\mu_m^s\right)^T\left(\Sigma_m^s\right)^{-1}\left(\mathbf{x}-\mu_m^s\right)}{2}\right\} \tag{3.7}$$

where D is the dimension of the vector $\mathbf{x}$, μ_m^s,Σ_m^s and p_m are the mean vector, covariance matrix and weight vector of the m^{th} component density of the speaker s, respectively. The component weights are bounded by the property $\sum_m^s p_m = 1$. $\left(\Sigma_m^s\right)^{-1}$ represents the inverse matrix operation performed on the covariance matrix of the m^{th} component of the s^{th} speaker while $|\Sigma_m^s|$ is the determinant of the covariance matrix [14].

The covariance matrices can be chosen as either a diagonal matrices or full covariance matrices. Selection of either the full or diagonal covariance matrix depends on the type of application and the required accuracy. Full covariance matrices represent the densities more accurately but are subject to computational overhead raised by the matrix inversion operation in Equation 3.7 whereas diagonal covariance matrices are easily invertible but are inferior in density representation as compared to full covariance matrices. During GMM modelling, the covariance matrix can be one of the following types [14]:

- Global Covariance: All speaker models have a single covariance matrix.
- Grand Covariance: Each speaker model has its own covariance model.
- Nodal Covariance: Every Gaussian component of each speaker model has its own covariance model.

In speaker recognition systems, diagonal covariance matrices are sufficient to model the probability densities of the feature vectors representing the speaker characteristics [58][55]. A speaker model with an M^{th} order full covariance matrix can be represented by an equivalent model consisting of higher order diagonal covariance matrices [34]. The diagonal covariance matrices are computationally less extensive as compared to full covariance since they do not require full matrix inversion. In [58][55] the diagonal covariances have been shown to sufficiently represent models based on full covariance matrices. In this work, the speaker recognition system employs diagonal covariance matrices.

Training a GMM requires the computation of the parameters $p_{m,}\mu_m^s$ and Σ_m^s from the given feature vectors belonging to a speaker. These values are calculated by an iterative algorithm known as the *Expectation-Maximization* (EM) [59], which is discussed below.

3.5.1.2 Expectation Maximization(EM)

Maximum Likelihood (ML) is the most widely used technique for the estimation of GMM parameters. ML aims to maximize the conditional probability or the likelihood $p(\mathbf{x}|\lambda_s)$ of the GMM from the given set of feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$. The EM algorithm uses these ML estimates to iteratively update the GMM parameters from the provided feature vectors until the model likelihood value converges. The algorithm is employed to determine the correct parameters that will monotonically increase the likelihood values of the GMM. In other words $p(\mathbf{X}|\lambda_s^{i+1}) \geq p(\mathbf{X}|\lambda_s^i)$, where i is the iteration number. For each iteration the GMM parameters are updated as follows [42][55]:

For the m^{th} component of every GMM of a speaker s, where $m = 1, 2, \ldots, M$:

- Mixture weights:

$$\bar{p}_m^s = \frac{1}{T}\sum_{t=1}^{T} p\left(m|\mathbf{x}_t, \lambda_s\right) \tag{3.8}$$

- Mean vector:

$$\bar{\boldsymbol{\mu}}_m^s = \frac{\sum_{t=1}^{T} p\left(m|\mathbf{x}_t, \lambda_s\right)\mathbf{x}_t}{\sum_{t=1}^{T} p\left(m|\mathbf{x}_t, \lambda_s\right)} \tag{3.9}$$

- Variances:

$$\bar{\boldsymbol{\sigma}}_m^s = \frac{\sum_{t=1}^{T} p\left(m|\mathbf{x}_t, \lambda_s\right)x_t^2}{\sum_{t=1}^{T} p\left(m|\mathbf{x}_t, \lambda_s\right)} - (\bar{\boldsymbol{\mu}}_m^s)^2 \tag{3.10}$$

- A posteriori probability

$$p\left(m|\mathbf{x}_t, \lambda_s\right) = \frac{p_m^s b_m^s(\mathbf{x}_t)}{\sum_{k=1}^{M} p_k^s b_k^s\left(\mathbf{x}_t\right)} \tag{3.11}$$

The iterative algorithm is terminated if $p(X|\lambda_s^{i+1}) - p(X|\lambda_s^i)$ is equal to a pre set threshold or if the user defined maximum number of iterations is reached. With the convergence of the likelihood values the EM algorithm stops and the updated parameters represent the speaker's GMM model. 5-10 EM iterations are generally adequate for parameter convergence.

The EM algorithm can be initialized by clustering the given feature vectors through an unsupervised clustering method such as the k-means [56]. In this work the k-means algorithm was initialized through random selection of candidate cluster mean vector from the given feature vectors. The Gaussian mixture components were initialized to be equally likely by setting each weight to be $\frac{1}{M}$, obtaining equally probable weights while the covariance matrix initialization was performed by using an identity matrix i.e. setting each diagonal element of the matrix to be 1 and each off-diagonal element to 0. It has been shown that such an initialization scheme can provide similar recognition performance compared to more elaborate phonetic segmentation methods based on HMMs [55].

3.5.1.3 Variance Limiting

During the GMM training, small variance values can affect the likelihood values of cause performance degradation. The small variance values can arise either because of noisy or insufficient data. It is therefore, necessary to apply some form of variance limiting during the training of the GMM models. Variance limiting can be applied as follows:

$$\bar{\sigma}_m^{s2} = \begin{cases} \sigma_m^{s2} & \text{if } \sigma_m^{s2} > \sigma_{min}^2 \\ \sigma_{min}^2 & \text{if } \sigma_m^{s2} \leq \sigma_{min}^{s2} \end{cases} \tag{3.12}$$

where σ_m^2 represents the m^{th} element of the variance vector, and σ_{min}^2 is the variance limiting value. The σ_{min}^2 is determined empirically and typically selected to be in the range of 0.01 to 0.1 [55]. The value of σ_{min}^2 should be chosen carefully as too high a value can cause masking of the actual variance values, hence degrading the model and the recognition performance. On the other hand a small value of σ_{min}^2 may not be sufficient to achieve required variance limiting and such may be ineffective. The values of the variances must be checked for every update obtained from the EM iteration.

3.5.1.4 Model Order

Selection of the number of Gaussian mixture components needed for appropriate modelling of the speaker characteristics is an important factor in the design of GMM based speaker recognition systems. A small number of components may not be able to adequately represent the speaker characteristics while too many components with limited training data can wrongly model the data resulting in poor modelling [55].

The training process is followed by the testing process, which involves matching the unknown test utterances from a claimant speaker to the stored models. Speaker identification and verification processes used in this study are based on GMM and are explained in the following section. The speaker recognition system based on GMM used in this study will be discussed in more detail in Section 3.2.

3.5.2 Hidden Markov Model

Hidden Markov Models are statistical models capable of representing the stationary as well as temporal characteristics. The assumption in HMM modelling is that the speech signal can be characterized as a parametric random process and the parameters can be estimated accurately [6]. HMMs can model both the speech sounds and their sequencing in the temporal domain.

HMMs model the speech feature vectors as a group of processes. A HMM models two stochastic processes: a hidden Markov chain, a process which is not directly observable and an observable process. The probability of following a particular transition depends only on the present state and not on the past states or transitions as defined by the Markov property. In a model based on HMMs, the temporal variations are dealt with by a hidden Markov chain while an observable process deals with the spectral variations in the feature vectors. A HMM contains a number of interconnected states with transitions among each state [35]. Changes in the signal are represented by a set of states with observation probabilities B_i and A_{ij} are the sequence of transition probabilities of the Markov chain [6, 60]. The probability density function (pdf) of each state statistically represents the feature vectors. The most commonly used pdf function for the HMM states is a multidimensional Gaussian pdf which was explained in Section 3.5.1.

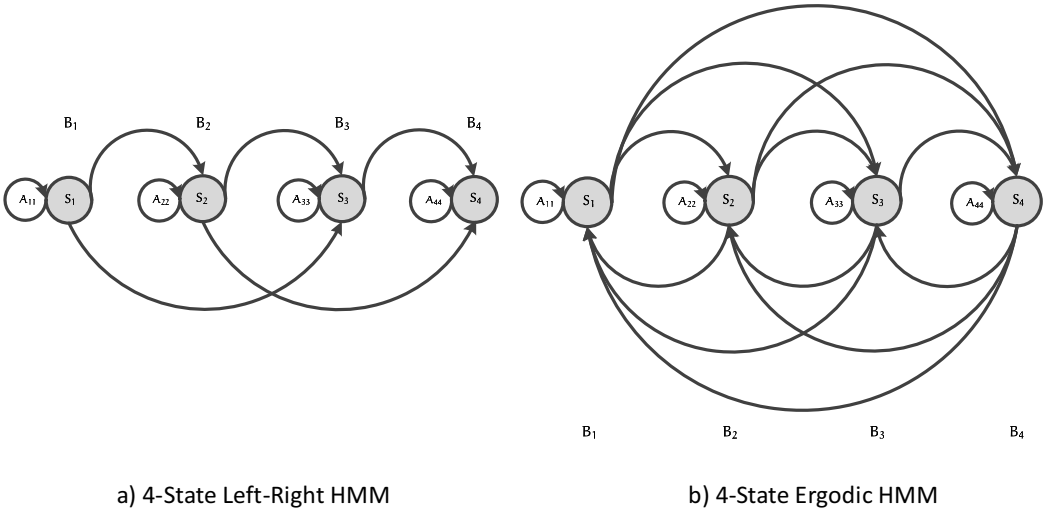


Figure 3.6: HMM Topologies

Depending upon the permitted transition among states, HMMs can be classified as either ergodic or left-to-right HMMs [61, 6]. In an ergodic HMM, all the states are interconnected and it is possible to make a transition from one state to another in a single step i.e. the state transition probabilities for an ergodic HMM are non-zero. For a left-to-right HMM the states move only from the left to the right with increase in time. Left-to-right HMM are used for signals that exhibit slowly varying properties such as a speech signal. Figure 3.6 shows a 4-state left-right and a 4-state ergodic HMM.

HMMs can be used for text-dependent and text-independent speaker identification and verification processes. A text-dependent process is modelled by a left-to-right HMM. To incorporate the flexibility of random text i.e. text-independent system, an ergodic or circular HMM is employed. Figure 3.6 provides a depiction of left-to-right and an ergodic HMM.

3.6 Speaker Identification

A speaker identification system determines the identity of the speaker from a group of known speakers. Feature vectors obtained from the utterances of the unknown speakers are matched against the GMM parameters of the enrolled speaker models. The model that gives the highest likelihood value is taken as the identity of the unknown speaker. A level diagram of the speaker identification system is depicted in Figure 3.7.

The likelihood of each known speaker is calculated by the *Maximum A posteriori Probability* (MAP) classification method. Given the feature vectors of the unknown speaker, the likelihood of each enrolled speaker model is given by the Bayes' rule as:

$$\hat{S} = \arg\ \max_{1 \leq k \leq S}\ Pr\left(\lambda_k | \mathbf{X}\right) = \arg\ \max_{1 \leq k \leq S}\ \frac{p\left(\mathbf{X} | \lambda_k\right)}{p\left(\mathbf{X}\right)} pr\left(\lambda_k\right) \tag{3.13}$$

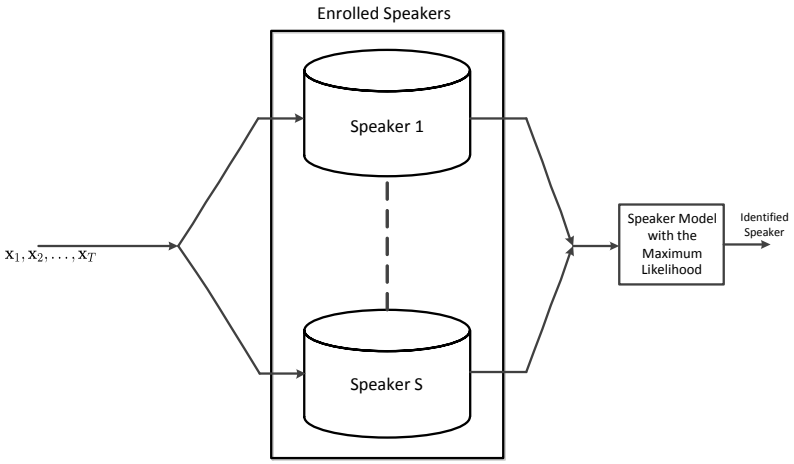


Figure 3.7: Speaker Identification System

where $\hat{S}$ is the identified speaker, $\boldsymbol{X}$ is the set of feature vectors $\mathbf{X} = \{\mathbf{x}_1,\mathbf{x}_{2,\ldots,}\mathbf{x}_T\}$, $pr\,(\lambda_k)$ is the prior probability of the speaker model λ_k and $p\,(\mathbf{X})$ is the prior probability of the training vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$.

In the context of the work carried out in this thesis, all speaker have been assumed to have equal a priori probabilities i.e. $pr\,(\lambda_k) = \frac{1}{S}$, where S is the number of enrolled speakers, also the training data $\mathbf{X}$ from all the S speaker models is also assumed to be equally probable with $p\,(\mathbf{X}) = \frac{1}{S}$, these assumptions lead to

$$\hat{S} = \arg\, \max_{1\leq k\leq S}\, p\,(\mathbf{X}|\lambda_k) \tag{3.14}$$

Each frame of speech data is considered to be independent from the others. The value of $p\,(\mathbf{X}|\lambda_k)$, i.e. the likelihood of the unknown speaker, can be calculated as a product of the likelihood values of each frame

$$p\,(\mathbf{X}|\lambda_k) = p\,(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}\,|\lambda_k) = \prod_{t=1}^{T} p\,(\mathbf{x}_t|\lambda_k) \tag{3.15}$$

or with the use of logarithm we have

$$\hat{S} = \arg\ \ \max_{1\leq k\leq S}\ \sum_{t=1}^{T} \log p\,(\mathbf{x}_t|\lambda_k) \tag{3.16}$$

which gives the identity of the claimant.

The recognition performance of the identification system is measure by means of the identification error rates as follows

$$\%\,\text{Error} = \frac{N_E}{N} \times 100 \tag{3.17}$$

where N_E represents the number of misclassified tests while N is the total number of

tests carried out by the identification system.

3.7 Speaker Verification

Speaker verification determine the actual identity of the claimant or the hypothesized speaker. The speaker verification system aims to determine whether the feature vectors from the unknown speaker, or the claimant in this case, match with the model selected through the speaker identification stage. As such, the speaker verification forms a binary decision with acceptance or rejection as the possible outcomes.

The verification process defines two hypothesis. Considering a set of feature vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, belonging to the unknown speaker. The first hypothesis H_0 states:

- H_0 : $\mathbf{X}$ belongs to the claimed speaker

and the second hypothesis H_1 is defined as:

- H_1 : $\mathbf{X}$ is not from the claimed speaker.

The decision is based on the result of the following likelihood test

$$\text{Likelihood Ratio} = \frac{p\left(\boldsymbol{X}|H_0\right)}{p\left(\boldsymbol{X}|H_1\right)} \begin{cases} \geq \theta \text{ accept} & H_0 \\ < \theta \text{ reject} & H_1 \end{cases} \tag{3.18}$$

where $p\left(\boldsymbol{X}|H_i\right)\ i = 0, 1$ represent the probability density function for the hypothesis H_i evaluated for the measurement $\mathbf{X}$, also referred to as the likelihood of the hypothesis H_i given the measurement. nd θ is the decision threshold for accepting or rejecting H_0.

The likelihood ratio test of Equation 3.18 can be re-written as

$$\text{Likelihood Ratio} = \frac{p\left(\boldsymbol{X}|\lambda_c\right)}{p\left(\boldsymbol{X}|\lambda_{\bar{c}}\right)} \tag{3.19}$$

where $\mathbf{X}$ is the feature vectors from the unknown speaker, $p\left(\mathbf{X}|\lambda_c\right)$ is the likelihood of the features vectors given that it belongs to the claimed speaker, and $p\left(\mathbf{X}|\lambda_{\bar{c}}\right)$ is the likelihood that the feature vectors $\mathbf{X}$ do not belong to the claimed speaker. The log-likelihood ratio can be written as

$$L\left(\boldsymbol{X}\right) = \log p\left(\boldsymbol{X}|\lambda_c\right) - \log p\left(\boldsymbol{X}|\lambda_{\bar{c}}\right) \tag{3.20}$$

where

$$\log p\left(\boldsymbol{X}|\lambda_c\right) = \frac{1}{T}\sum_{t=1}^{T} \log p\left(\boldsymbol{x}_t|\lambda_c\right) \tag{3.21}$$

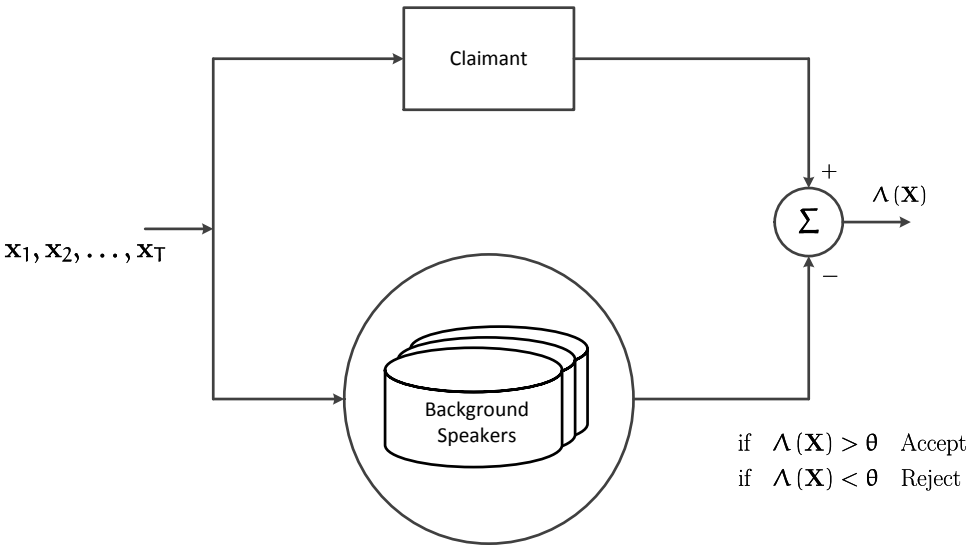


Figure 3.8: Speaker verification process using the universal background model

and

$$\log p\left(\boldsymbol{X}|\lambda_{\bar{c}}\right) = \log\left(\frac{1}{m}\sum_{k=1}^{m} p\left(\boldsymbol{X}|\lambda_k\right)\right) \tag{3.22}$$

m represents the total number of background speakers. The speaker verification process is depicted in Figure 3.8.

3.7.1 Background Speaker Selection

The process of speaker verification requires models for the alternative speaker and the hypothesized speakers. The process of background speaker selection must be carried out carefully to properly represent the alternative speakers. There are two known methods for the creation of alternative hypothesized speaker modelling in the process of speaker verification. The first method utilizes a set of known speaker specific models to determine the alternative speaker. A particular set of background speaker models are used for each speaker in the database. For a large database, requirement of increased storage space and high computational costs pose problems in the application of this method for the purpose of speaker verification. For this method different approaches have been presented in [55, 62, 63]. In [55] the alternative speaker models are created by using a combination of speakers who have a similar or dissimilar voice properties to the hypothesized speaker. The selection process for the background speaker in [55] is presented below:

- Compute GMMs for all the speakers in the database

- Pair-wise distance between each GMM, the pair-wise distance $d(\lambda_i, \lambda_j)$ is computed as:
$$d(\lambda_i, \lambda_j) = \log \frac{p(\boldsymbol{X}_i|\lambda_i)}{p(\boldsymbol{X}_i|\lambda_j)} + \log \frac{p(\boldsymbol{X}_j|\lambda j_j)}{p(\boldsymbol{X}_j|\lambda_i)} \tag{3.23}$$
- Calculate the n farthest and n closest speakers from the hypothesized speakers
- Select $\frac{m}{2}$ farthest and $\frac{m}{2}$ closest speakers that are maximally spread from each other ($m < n$)

The two stages of background speaker selected from the above formulation are known as *Maximally Spread Close (MSC)* and *Maximally Spread Far (MSF)* set respectively. The number of speakers m must be selected carefully, which reduces the computational requirements and leads to an effective representation of the possible impostor group. The result of speaker verification is computed as a likelihood ratio test, Equation 3.18, resulting in either acceptance or rejection of the claimed speaker.

3.7.1.1 Universal Background Model (UBM)

The second method employs a generalized alternative model for all the hypothesized speakers. This method is known as the *Universal Background Model (UBM)* [64]. The speaker-independent model is formed by using a number of different speakers to represent the alternative hypothesized speakers.

When one large alternative speaker model is used for representing the background speaker model, the speech used for training must be chosen so that it represents the existing speaker features. Multiple background speaker models can be used depending upon the requirements of the application. During the generation of a UBM the training and the testing data must be chosen carefully. In case of gender-dependent experiments, two single-sex UBMs are required, one based on male speech and the other on female speech only. For the gender-independent case one UBM is used consisting of both the male and female speech. However, UBMs can be tailored to better represent the characteristics of the enrolled speakers in the database. This reduces the mismatch between the training and the testing data as well as allows for better speaker modelling. A model order in the range of 512-2048 mixtures can generally represent the underlying desired speech characteristics of the database. There exists no general method of generating the UBMs. UBMs are created by pooling the speech from different sets of speakers which represent the general characteristics of the speech features. UBMs must be generated in a manner to ensure that they do not favour a sub population over rest of the population i.e. in case of a gender-independent experiment, equal number of speech features should be used from male and the female speech, in order to avoid any bias towards a particular gender [65].

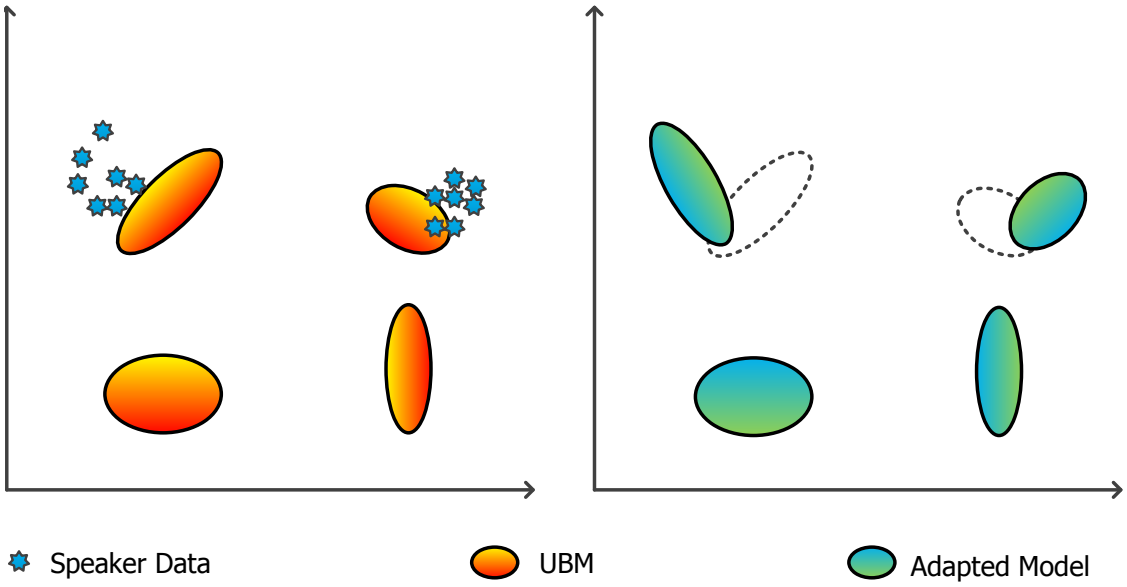


Figure 3.9: Adaptation of a speaker's models using the universal background model (UBM)

3.7.1.2 Adaptation of Speaker Model

A good representation of the speech features can be obtained by a large and a well-trained UBM. The UBM can be changed for the representation of the hypothesized speakers. MAP estimation and the training speech of the input speaker can be used to adapt the parameters of the UBM to model the hypothesized speaker [66]. The adaptation of the UBM parameters for modelling the hypothesized speaker provides a strong link between the two models. This coupling provides higher recognition performance and simplifies the speaker scoring time as described below.

The hypothesized speaker model can be obtained from the UBM through the following steps:

1. Calculate the estimates of the count, the first and the second moment of the hypothesized speaker's training data for each UBM mixture.
2. Adapt the model using the combination of of the newly estimated statistics from the first step with the statistics of the UBM.

The first step probabilistically maps a speaker's training data onto the UBM mixtures. The next step calculates the adapted model parameters by the use of UBM mixture parameters and the training data statistics.

The process of speaker adaptation is described below after [65]:

The count, first and the second moments of the hypothesized speaker, with feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ and a UBM are computed as follows:

$$\eta_i = \sum_{t=1}^{T} Pr\left(i|\mathbf{x}_t\right) \tag{3.24}$$

$$M_i(\mathbf{x}) = \frac{1}{\eta_i} \sum_{t=1}^{T} Pr(i|\mathbf{x}_t)\, \mathbf{x}_t \tag{3.25}$$

$$M_i(\mathbf{x}^2) = \frac{1}{\eta_i} \sum_{t=1}^{T} Pr(i|\mathbf{x}_t)\, \mathbf{x}_t^2 \tag{3.26}$$

where η_i is the count, $M_i(\mathbf{x})$ and $M_i(\mathbf{x}^2)$ are the first and the second moments respectively, and $pr(i|\mathbf{x}_t)$ is the probability of the i^{th} component of the UBM mixture given the vector $\mathbf{x}_t$ can be derived as

$$Pr(i|\mathbf{x}_t) = \frac{p_i^s b_i^s(\mathbf{x}_t)}{\sum_{j=1}^{K} p_j^s b_j^s(\mathbf{x}_t)} \tag{3.27}$$

Now the adapted weights, means and covariance vectors can be formulated as

$$\hat{p_i} = \left[\frac{\alpha_i \eta_i}{T} + (1-\alpha_i)\, p_i\right] \gamma \tag{3.28}$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i M_i(\mathbf{x}) + (1-\alpha_i)\, \boldsymbol{\mu}_i \tag{3.29}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \alpha_i M_i(\mathbf{x}^2) + (1-\alpha_i)\left(\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2\right) - \hat{\boldsymbol{\mu}}_i^2 \tag{3.30}$$

Here α_i is the coefficient of adaptation and γ is the scaling factor. The value of α_i is calculated as

$$\alpha_i = \frac{\eta_i}{\eta_i + r} \tag{3.31}$$

Here r is the fixed relevance factor whose is value is determined empirically and is fixed between 8 and 20 [65]. γ is the normalization factor ensuring that the values of the adapted weights sum to unity. The value of α_i is dependent on the data and controls the balance between the old and the new estimates. The adaptation process depends upon the speaker data. Only the components of the UBM mixture that have sufficient correspondence with the speaker data, are adapted. A UBM represents the wide range speaker-independent speech sounds and the adaptation process modifies the UBM to represent the speech classes derived from the speaker-dependent training speech.

The log-likelihood ratio, Equation 3.20, of the hypothesized speaker is calculated from the hypothesized speaker model and the UBM. Since the hypothesized speaker model is an adapted version of the UBM, a method called fast-scoring can be utilized. When a new test-set is presented to the system only a small number of UBM components will be close enough to affect the final likelihood values, since the speaker's adapted GMMs are obtained from the components of the UBM, the same components would represent the speaker in the large mixture model of the UBM. The likelihood values can be estimated using the best scoring top C components. The likelihood values can now be found as:

- Estimate the likelihood values from the UBM components,

- Using only the best scoring C components, calculate the likelihood values
- Calculate the adapted speaker model likelihood result using only the C components
- Calculate the speaker's likelihood values.

A typical value of $C = 5$ has been suggested in [65]. This speeds up the computation, as it requires only $C + M$ calculation instead of the $2M$ computations required in the case of an M^{th} order UBM.

3.7.1.3 Error Measures and Decision Criteria

A recognition decision is made after the computation of the likelihood values. Here, decision making is a binary process with the two possible outcomes being acceptance or rejection. The speaker verification system compares the values of the speaker's likelihood values with the threshold θ. If the value of the likelihood is greater than θ, the claim is accepted and is rejected when the likelihood value falls below θ. Similar to any decision making system, a speaker verification system can produce type-I and type-II errors. A type-I error occurs when the verification system rejects speech from a speaker who is enrolled in the system, this type of error, in the context of speaker recognition, is known as *False Rejection (FR).* A type-II error results when the system fails to reject the speech from an impostor, resulting in what is known as a *False Acceptance (FA).* In simple terms FR is the case when an enrolled speaker is considered an impostor by the system and is rejected while FA is the scenario where an impostor is accepted as a true speaker by the system. The *False Acceptance Rate (FAR)* and *False Rejection Rate (FRR)* associated with the speaker recognition systems are defined as

$$FAR = \frac{I_A}{I_T} \tag{3.32}$$

$$FRR = \frac{C_F}{C_T} \tag{3.33}$$

where I_A is the false acceptances, I_T is the total number of impostor verification attempts, C_F is the number of false rejections and C_T is the total number of claimant verification attempts. The value of the threshold θ should be chosen so as to minimize the overall error score of the system. Selection of a suitable value for θ depends upon the application e.g. for increased security the value of θ can be chosen so as to reduce the FA occurrences.

Equal Error Rate (ERR) is one way of reporting the verification score of the system [40]. The value of θ is selected to obtain a value of EER such that the rate of false acceptances is equal to the rate of the false rejections. The most common and widely

used method of representing the middle ground between the FAR and FRR values is the *Detection Error Trade-off (DET)* curve [41]. The DET curve is obtained from the results of the speaker verification experiments which are presented below along with the simulation set up used in this study for the task of speaker recognition.

Having described the fundamental concepts in speaker identification and speaker verification, we now present the implementation of the base line speaker identification and speaker verification system in the following sections. The performance of this baseline speaker identification system is also analysed with the TIMIT speech corpus. Later in the thesis, the performance of the system will also be observed against synthetic converted voices that have been obtained by the use of voice conversion techniques.

3.8 Speaker Identification Implementation

This section describes the implementation of the speaker identification system that will provide high recognition performance and forms the basis of the work carried out in the following chapters of this thesis. The speaker identification system utilizes GMM (Section 3.5.1) for speaker modelling and evaluations. The following sections describe the process of speaker identification along with the description of the speech corpus, the process of feature extraction and the method of performance evaluation.

3.8.1 Speech Corpus

The development and evaluation of a speaker recognition system requires the availability of a speech corpus. Some of the speech corpora which have been widely used in literature for the task of speaker recognition include TIMIT [67], NTIMIT [68], YOHO [69, 70], Switchboard [71] and KING [72]. The National Institute of Standards and Technology (NIST) [73] has been carrying out speaker recognition evaluations since 1996. It provide recommendations for fair measurement grounds to evaluate the performance of a speaker recognition system under criteria defined by NIST [74], detailing the guidelines for determining the best speaker recognition methods and put forth the direction for the ongoing research. NIST has been providing yearly updated speech databases to its participants. The speech corpus used in this thesis for evaluating the speaker recognition system is the standard American English Database TIMIT (Texas Instruments / Massachusetts Institute of Technology) and is provided by the Linguistic Data Consortium (LDC) [75].

3.8.1.1 TIMIT Corpus

The TIMIT speech corpus [76] was designed to provide a large speaker database with diverse range of population, containing rich phonetic content. The database consists

Sentence Type	Unique Sentences	Total	Sentences/Speaker
Dialect (*sa*)	2	1260	2
Compact (*sx*)	450	3150	5
Diverse (*si*)	1890	1890	3
Total	2342	6300	10

Table 3.1: TIMIT Corpus Sentence Assignments

of 630 speakers from 8 different dialect regions of the United States. The database consists of speech from 438 male and 192 female speakers. Each speaker is designated 10 speech files with an average duration of 3 seconds per file. Depending upon the phonetic content all the speech files are divided into three different groups and labelled accordingly. The *sa* files represent the dialect sentences, the phonetically-compact sentences are labelled as *sx*, while the phonetically-rich sentences are designated as *si* sentences. There are two *sa* sentences which are common to all the speakers of the database. Each speaker is assigned 5 *sx* sentences, and the *sx* sentences are shared by 7 speakers making a total of 450 phonetically-compact *sx* sentences. The *si* sentences are unique to all the speakers with no overlap. Each speaker utters 3 *si* sentences. A breakdown of the different speech files in the corpus is shown in Table 3.1

The speech files were recorded with high quality microphones in a quiet environment. All the speech files were recorded in one session to avoid inter session variations in the speech of the same speaker. All the speech material has been recorded with a sampling frequency of 16 kHz.

3.8.2 Preparing the Speech Material

In the experimentation carried out in this thesis, two versions of the TIMIT corpus were used for a closed-set speaker identification system. The TIMIT-16, consisting of speech files sampled at $16kHz$ and the TIMIT-8, where all the speech material is sampled at $8kHz$. Normally 20 to 30% of the speech material in the corpus is used for the testing purposes while the remaining $70\,to\,80\%$ is used for training purposes. All the 630 speakers of the corpus were enrolled in the system and the sentences of each speaker were segregated as training and testing material which will be discussed later in the section.

The feature extraction process begins with the removal of silence regions from the speech. As was mentioned in Section 2.4.1.2, separation of silence from speech is essential otherwise the extracted features will model the environment rather than the characteristics of the speaker. The TIMIT corpus provide complete transcripts of the speech files. These transcripts include details of the speech and the silence intervals. It is, therefore, easier to separate speech from the silence and is the method of voice activity detection employed in this thesis. After the removal of the silence regions from the speech samples, the files are prepared for the training and testing of the speaker identification system as below:

- The *sa* and *si* sentences were concatenated together, providing approximately 15 seconds of speech for each speaker. The combination of the *sa* and *si* files was used as training material for the speaker recognition system.
- The 5 *sx* files per speaker were concatenated to give on average 15 seconds of speech that was used in the testing phase of the recognition system.

The concatenated speech files are analysed using 20 msec Hamming window, as was mentioned in Section 2.4.2, corresponding to 320 samples of speech sampled at 16 kHz or 160 samples for 8 kHz sampled speech. A frame update rate of 10 msec, corresponding to 160 samples for 16 kHz and 80 samples for speech sampled at 8 kHz, was used. Each analysis segment is multiplied by a Hamming window to reduce the discontinuities at the boundaries. Each windowed segment of the speech signal undergoes the process of extracting MFCC, which was described in Section 2.4.4.3, and is briefly revisited below.

The length of the windowed segment of speech is increased from N samples to $2N$ samples by means of zero-padding, to improve the frequency resolution of the signal. After the computation of the DFT, Equation 2.44, the energy coefficients are computed as an inner product of the mel-scale filter banks, Figure 2.10, and the magnitude of the Fourier transform of the windowed speech segment. Logarithm is applied to the energy coefficients and finally the MFCC are obtained by evaluating the DCT, Equation 2.47, on the log spectral energy values.

For speech sampled at 16 kHz, 24-dimensional MFCC vectors were extracted from each segment of windowed speech, covering a frequency range of $0 - 8000$ Hz. For TIMIT-8 experiments 16-dimensional MFCC were used for each windowed segment of speech, encompassing the frequency range of $0 - 4000$ Hz. As was mentioned in Section 2.4.4.3 the zero order MFCC represents the average energy of the speech frame as is not included in the set of the feature vectors.

3.8.3 Speaker Modelling

For each speaker that has to be enrolled in the speaker identification system, a corresponding model was built to provide a characteristic representation of the speaker-specific properties. Each speaker model was constructed using 32-component GMM, as defined in Section 3.5.1, the mixture components were initially set to $\frac{1}{M}$ where $M = 32$ the number of mixture components. Diagonal-nodal covariance matrices were used with the matrix values initialized by an identity matrix with a variance limiting value of 10^{-2}. The component means were initialized by randomly selecting 32 MFCC vectors as component means and then using a single pass of the k-means algorithm. The model parameters were estimated by iteratively running the EM parameter estimation algorithm as described in Section 3.5.1.2. The number of iterations was limited to a maximum of 10 , which is sufficient for the convergence of the likelihood values [55].

Database	Identification Performance (%)
TIMIT-16	99.7
TIMIT-8	98.4

Table 3.2: Identification Performance of the Speaker Identification System with TIMIT-16 and TIMIT-8

3.8.4 Performance Evaluation

The trained speaker models were stored as a computer file representing a database of the enrolled speakers. After completion of the training process and the generation of speaker specific model, the system performance was evaluated in the testing phase. The process of testing for a speaker identification begins by extracting the feature vectors from the speech of the unknown speaker, claiming to be one of the enrolled speakers. The feature extraction is performed as was described in Section 2.4.4.3. The feature vectors of the unknown speaker are compared with the stored models of the enrolled speakers and log-likelihood values are generated according to Equation 3.16. The speaker model which gives the highest log-likelihood value for the test vectors of the unknown speaker is selected as the best matching model and the identity of the unknown speaker is taken as the identity of this best matching model. The identification performance for the TIMIT-16 and TIMIT-8 databases is given below:

Table 3.2 shows that the TIMIT-16 has a high identification performance of 99.7 (%). This is expected as the TIMIT corpus is a clean, almost-ideal and phonetically rich files which does not have any inter session variations. The TIMIT-8 achieves an identification performance of 98.4 (%). The small drop in performance is due to the loss of the high frequency components and the lesser number of mel-scale filter banks as was mentioned in Section 3.8.2.

3.9 Speaker Verification implementation

This section describes the implementation of the speaker verification system and the evaluation measures. The verification system experiments are performed on the TIMIT-16 and TIMIT-8 speech corpus. The simulations detail the performance of the speaker verification system on clean speech. The preparation of the speech material in these experiments is described below:

- Approximately 24 *sec* of training speech was accumulated for each speaker by concatenating eight speech files including the two *sa* files, three *si* and five *sx* files from each speaker.
- The test speech for each speaker in the test set contains two *sx* sentences averaging up to 3 *sec* of speech per speaker.

The test set of the TIMIT speech corpus contain 56 female and 112 male speakers for the speaker verification simulations with a minimum of two male and one female speaker from each of the eight dialect regions of the US and 2 sentences per speaker. This provides a total of 336 male and female speaker tests. The impostor attacks were carried out using the two speech files for every speaker, giving 56×55 female and 112×111 male, or a total of 31024 attack sets.

3.9.1 Background Speaker Modelling

The speaker verification experiments were performed with two background models which were designed to be gender-dependent. The test set of TIMIT corpus does not contain an equal number of male and female speakers, therefore a different test set was designed for the evaluation of the speaker verification system separately for the male and female speakers. Usage of two gender-dependent background models ensure that the final models will not be biased towards a particular gender. It has been suggested in [65] that one hour of speech is sufficient for modeling the background speakers and the same amount of speech material has been used in these experiments for modelling the background speakers. The GMM representing the UBM consists of 1024 components. 1024 components can adequately model the alternative speakers and can provide high recognition performance [77]. The mixture weights were initialized to $\frac{1}{M}$ where $M = 1024$, the number of mixture components. An identity matrix is used to initialize the nodal-covariance matrices and variance limiting was set to 0.01. The components means were initialized by a single pass of the k-means algorithm where the initialization seeds for k-means were 1024 randomly selected MFCC. The parameters of the models were estimated by the EM algorithm. Since the number of components is substantially large, the maximum number of EM iterations has been limited to 20 instead of 10 as was the case in modelling the speakers in Section 3.8. This is to allow the likelihood values of the UBM to converge [77].

3.9.2 Performance Evaluation

The creation of the UBM is followed by adaptation of every speaker model from the GMM-UBM as was discussed in Section 3.7.1.1. The adapted speaker models are stored to create a speaker database. During the testing phase, feature vectors extracted from the unknown speaker are used to compute the likelihood values from the top 5 best scoring components using the fast scoring technique described in Section 3.7.1.1. The results of the verification experiment are reported as EER values for both test sets of male and female speakers.

Database	Male	Female
TIMIT-8	1.34	1.79

Table 3.3: Equal Error Rate (EER) for TIMIT-8 Male and Female Speech

3.10 Conclusion

This chapter presented the most common techniques employed in the task of speaker recognition. The GMM are well known for their high recognition performance as has been demonstrated in [34, 40, 55, 65] as well in this chapter. Since the research work in this thesis deals with the performance of speaker recognition systems against impersonation attacks, GMM have been used for modeling the speaker in a speaker recognition system as have also been used in speaker modelling and transformation in the voice impersonation system which will be discussed in detail in Chapter 4. Speaker recognition system consists of speaker identification and speaker verification system which have been described in this chapter. The speaker identification and speaker verification systems use GMM for modelling the speakers. The verification system using adapted GMM and the criteria for speaker selection in the UBM were also described. The decision making process as well as the error measures used have also been mentioned in the later parts of the chapter.

The performance of the baseline identification and verification systems have been described. The chapter also details the structure, content and experimental set up of the TIMIT corpus which has been used to evaluate the performance of the recognition system. The performance of the speaker recognition system has been described in terms of ERR. Chapter 4 details the process of changing the voice of an individual in order to impersonate a speaker that is already enrolled in the speaker recognition system.

Chapter 4

Computer Aided Voice Impersonation

4.1 Introduction

The speech signal carries a wide range of information: linguistics, segmental, supra-segmental, paralinguistic etc. The speech signal not only conveys the message of the speaker but it also carries with it the identity of the speaker. Voice is a unique and non-intrusive attribute. Voice individuality is not only imperative because it helps to identify the person but it also enriches our daily lives [11]. Voice impersonation is an act of disguising ones voice and to try to mimic speech produced by another speaker. Voice conversion is a technique to change the speaker's individuality, i.e. to reshape speaker's voice characteristics in order to change the perceived identity of the speaker, so that an utterance appears to have been spoken by a different speaker. The voice conversion technology finds numerous applications in speech synthesis such as in text-to-speech conversion for creating new computer voices without the need of recording additional human voices. It also allows for customized voice conversions in the entertainment industry thus eliminating the need for skilled mimickers. In the area of speech recognition it is desirable to get rid of any speaker specific information in the speech signal before the recognition process, and therefore some form of speaker normalization will greatly aid the speech recognition performance. Voice conversion, because of its close relationship to speaker adaptation techniques, can be employed in these cases to convert all the input speakers to a single generic speaker [78]. Voice conversion techniques can also be used for the aid of the people suffering with some form of hearing and speech impairments [79, 80]. Different approaches have been presented in literature for voice conversion consisting of techniques dealing with the mapping of spectral characteristics of one speaker onto the spectral properties of another [81, 82].

This chapter describes the process of voice conversion, starting with the factors which

contribute to individuality in a speaker's voice. Different methods have been proposed in literature for the determination of an optimum conversion function between the spectral properties of different speakers, some of the well-known techniques are briefly revisited in this chapter. In the later half of the chapter, a voice conversion system based on the GMM modelling is described, with a detailed description of the processes and procedures involved for determining the relevant phonetic correspondences between the speech samples of two different speakers. Two techniques have been proposed in this chapter dealing with the shortcomings in the performance of the voice conversion systems. The first approach deals with the problem of over smoothing in GMM based voice conversion systems: the second addresses the discontinuities arising from the training of the conversion function with limited amounts of training data. The chapter concludes by presenting the results of a subjective experiment conducted on the outputs of a conventional GMM based voice conversion system and the system with the proposed changes. The results indicate a preference for the output of the modified system over the traditional GMM based voice conversion systems.

4.2 Factors Affecting Voice Individuality

The perceived speaker identity is a consequence of combining several factors. It has been reported in literature that the supra-segmental features such as the speaking rate, the duration of pauses during conversations and the evolution of pitch contour contribute greatly to the perceived speaker identity [83][84][85]. Also the voice individuality is dependent on the linguistic style of the speech, such as the choice of particular words, the use of a certain dialect and the selection of a particular accent. It is however, difficult for a machine to model these features as high-level considerations are involved. Also the meaning of the spoken text and the intention of the speaker strongly affect the prosodic features, which causes hindrance in the automatic processing of these features, specially in cases where the text of the utterance in not known beforehand. The average value of these features, however, is strongly linked with the speaker specific information [86][83][84]. Also using the spectral envelopes of the corresponding segmental level features can lead to effective speaker discrimination [83][87]. In the view of these findings, most of the commonly used speaker recognition techniques employ classification of the statistical distribution of the spectral envelopes [43][88]. Generally, the overall shape of the spectral envelope and location and bandwidth of the formants are considered to be the most speaker defining features.

According to literature, some of the factors which contribute to the voice individuality and the perceived identity of a speaker are listed below [89][90][11]:

- Spectral envelope shape and spectral tilt
- Absolute values of formant frequencies

- Average speech spectrum
- Formant trajectories
- Formant bandwidth
- Pitch frequency
- Pitch contour
- The glottal wave shape

The voice individuality of a speaker is not entirely dependent on any one of these factors but on a combination of these, where the importance of each factor varies from one speaker to the other [11, 82].

4.3 Voice Conversion

Voice conversion techniques transform the speech signal generated by a speaker in a way to alter the characteristics of his/her voice. In terms of psychoacoustics, the correlation between the spoken text and the perceived speaker identity is largely unknown. It is, however, easier to modify the speech signal uttered by an individual, if the desired modifications are carried out with reference to another speaker. Voice conversion refers to techniques that attempt to modify the characteristics of a speech signal uttered by a speaker, so that it appears to have been spoken by another speaker [91].

4.3.1 Applications

There are a number of applications for voice conversion mentioned in literature. Some of the more popular ones are listed below:

- The most popular application of voice conversion is in text-to-speech conversion [92]. Voice conversion can be used to alter the characteristics of the standard speaker to adapt or personalize synthesized voices in corporate dialogue systems [93].
- Voice conversion techniques can be used to build a concatenation speech synthesis system by normalizing the high quality speech databases to increase the available speech data [94].
- Cross-language voice conversion can be used in entertainment industry for dubbing tasks in films and music [95].
- Speech from people suffering from dysarthiya can be modified by voice conversion techniques to enhance the intelligibility and naturalness of the otherwise impaired-speech [96].

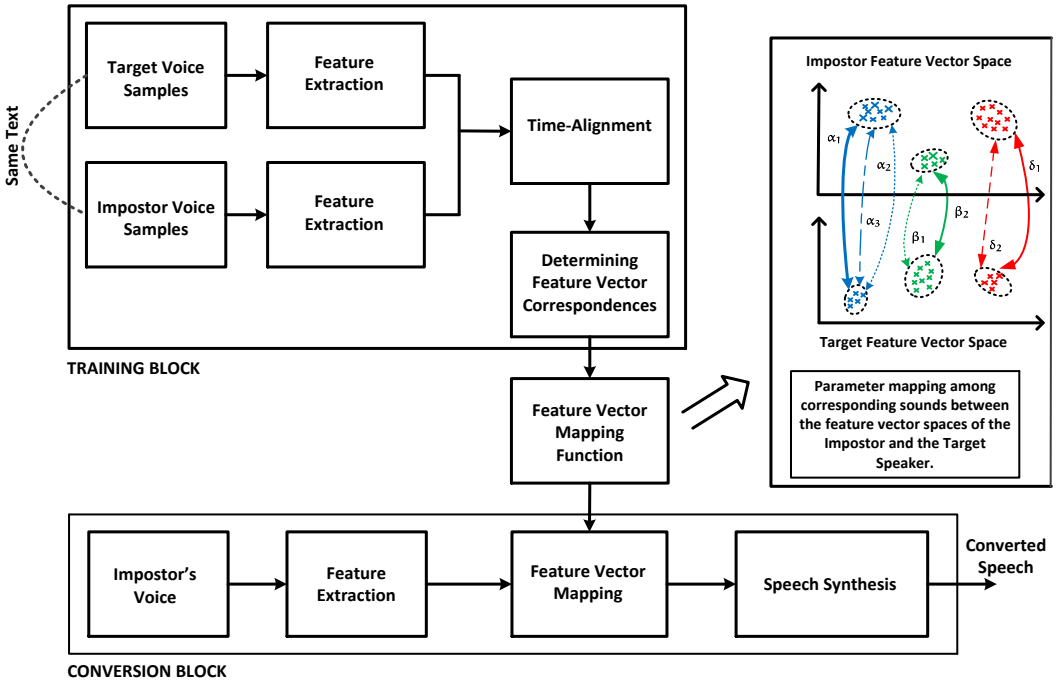


Figure 4.1: Block Level Diagram of a Voice Conversion System

- Voice conversion techniques can be used in speech recognition systems to normalize the voices of the incoming speakers to a standard speaker. The inclusion of the voice conversion systems in speech recognition has been shown to improve the recognition performance [97].

4.4 Components of A Voice Conversion System

A typical voice conversion system, shown in Figure 4.1, has two main parts: Training and Conversion. This section briefly describes the purpose and procedures carried out by the parts of a typical voice conversion system.

4.4.1 Training

In the training mode, the voice conversion system analyses the speech samples taken from the source (impostor) and the target speaker. The analysis is carried out with reference to a particular speech model. Commonly used speech models are based on linear prediction, Section 2.4.3, and therefore result in parameters that characterize the spectral envelope [98, 4]. Systems that attempt to go beyond the spectral transformation have also been proposed in [99, 100]. A training stage in a typical voice conversion system is shown in Figure 4.2.

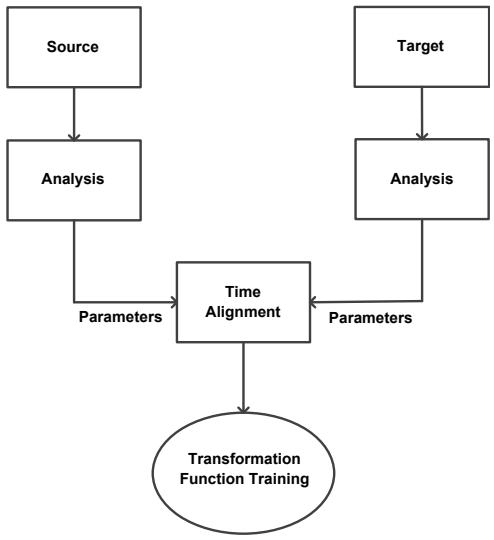


Figure 4.2: Training Stage of the Voice Conversion System

The first stage in the training process is the speech analysis. During the analysis stage, parameters representing the source and target speech are extracted. The analysis stage is followed by the application of techniques which try to determine the correspondences between the speech sounds of the source-target pair, leading to the generation of the training data. These correspondences are obtained by grouping the source and target features together which represent the same sound or phonetic class. Such a grouping can be achieved by time-alignment or classification using techniques such as Dynamic-Time-Warping (DTW) [6], unsupervised Hidden Markov Models (HMM) or forced alignment [100]. The training data obtained from the alignment procedure is used to estimate a transformation or conversion function. The aim of a conversion function is to find statistical relationships between the features representing the source and target speech sounds. Different implementations of the conversion function have been proposed in the literature e.g. using a mapping codebook [4], neural networks [101] and Gaussian Mixture Models [98, 82].

4.4.1.1 Speech Corpus

A speech corpus provides the speech data required for the training of the conversion function and for evaluating the performance of the voice conversion system by objective and/or subjective experiments. The optimum size and content of the speech data depends on the requirement of a particular application of the voice conversion system. It can contain just the vowels [102], words [103, 4], short read sentences [104, 92] or hours of read speech [100].

The number of speakers along with the amount of speech data is an important aspect for the design of any speech corpus. A larger speaker population is advantageous for the design and evaluation of a voice conversion system as it aids a better representation of the general population as well as providing a sufficiently rich set of prosodic choices

for the context.

4.4.1.2 Modelling and Feature Extraction

Any speech processing system requires some meaningful representation of the speech signal. Selecting a particular representation or model depends on the requirements of the application. In the context of voice conversion systems, an optimal model should be able to generate a variety of speech that is intelligible, accurate and sounds natural with respect to the speaker individuality. These criteria demand a speech model that should have high degrees of freedom, however, the transformation function is normally trained on a low-dimensional parameter set obtained from limited amount of training data. These conflicting requirements demand a compromise between the transformation function and speech model.

It was mentioned in Section 4.2 that the voice individuality is represented by all the acoustic cues. It was also mentioned that the segmental features and the average value of the supra-segmental features in particular are sufficient to obtain a high degree of speaker recognition by humans. Also, in Section 3.6, it was shown that the parameters representing the spectral envelope alone contain enough information for effective speaker discrimination by automated speaker identification systems. Based on these findings, the voice conversion system almost always focuses on the conversion of spectral envelope parameters, to alter the characteristics of the source speaker to match the properties of the target speaker's parameters. Besides the transformation of the spectral envelope parameters the average value of the source speaker's $F0$, energy, and the speaking rate are adjusted to match those of the target. Similar to most of the other speech processing systems, the speech signal in a voice conversion system can be processed in short segments called frames (Section 2.4.2) or conversion of entire phonetic units [81].

The Source-Filter model, Section 2.2.1, provides a successful representation of the speech signal for voice conversion systems. Speech, according to this model, is produced by fitting a spectral envelope over the magnitude spectrum of the excitation signal generated by the lungs. The vocal tract is estimated as a slowly varying spectral envelope and often the parameters of the source-filter model are computed by means of linear prediction. LPC, introduced in Section 2.4.3, represents the coefficients of the time-varying filter and are seldom used in their original form as they are very sensitive to even the smallest of variations in their values. Several alternative representation of the LPC, some of which were mentioned in Section 2.4.3.1, are used in speech processing systems. The alternate representations have properties which are more desirable e.g. interpolation and the capacity to localize errors in their values.

The LPC residual signal is obtained by inverse filtering the speech segment with its corresponding LPC filters. Since the LPC coefficients represent the vocal tract, inverse filtering the speech signal removes the contribution of the vocal tract. The output

of the inverse filtering operation is the glottal excitation waveform (see Section 2.2). The source excitation signal can be used without any modifications in the synthesis of transformed speech [103, 98]. This approach results in a more natural sounding speech. It has been shown in literature that the excitation waveform contains speaker specific information [92, 105]. Several approaches have been proposed in literature to modify the source speaker's residual in addition to the transformation of the spectral envelope. *Dynamic Frequency Warping* (DFW) is a technique that works directly on the magnitude spectrum [103]. DFW attempts to find a non-linear mapping of the frequency axis in an effort to find the changes in the speaker characteristics. However, this technique was found to be inferior to the traditional spectral envelope mapping algorithms [103]. A codebook based transformation of the source LPC residuals have also been suggested in [104, 100, 106], by using a weighted combination of excitation filters for each class of spectral envelope transformation. This approach is a two-stage spectral conversion as both the spectral envelope represented by the LPC and the LPC residual are transformed using the same single classification. In [99], a neural network has been used as a transformation function. During the conversion stage the weights of the neural network are transformed along with the parameters representing the spectral envelope.

4.4.2 Conversion

During the conversion stage, the transformation function estimated during the training stage is used to transform the source features to target features. The predicted features are then used for generating the final transformed speech signal at the speech synthesis stage. The conversion stage of the voice conversion system is shown in Figure 4.3.

The prosodic features such as $F0$ contours, speaking rate etc. can be adjusted to match the average prosody values of the intended target speaker. As mentioned in Section 4.2, it is difficult to model the supra-segmental cues such as the intonation patterns since it involves the extraction and manipulation of high level information. Although some progress has been made in developing the intonation models [107, 108], these models however, require significant manual effort, are controversial, difficult and inaccurate [109], These factors make the transformation of prosodic features in voice conversion systems unsuitable for obtaining satisfactory results. In this thesis, the focus is only on the transformation of the segmental features.

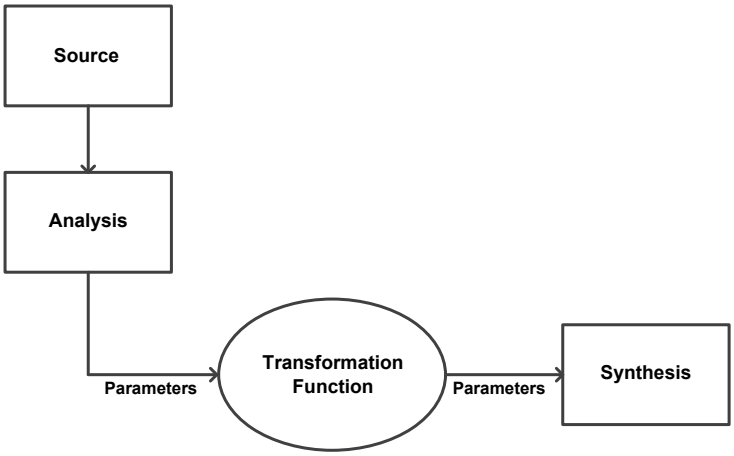


Figure 4.3: Voice Transformation Stage of a Voice Conversion System

4.5 Conversion Function Training

The role of a conversion function is to find correspondences between the feature vector spaces of the source and the target speakers. The differences between the feature vectors of the source and target speakers arise due to the differences in the physical characteristics of the sound producing organs as well as the variations in the linguistic units even when producing the same utterances. Before training of conversion function, it is important to group the feature vectors linguistically or to time align the feature streams. Such grouping of the feature vectors provide the necessary associations between the source and the target features which are required for the training of the transformation function. These associations have been determined by means of DTW [16] as in [110, 81, 103, 111], unsupervised HMM [100, 104], forced-aligned speech recognition [112] or the use of a phonetic classifier [100, 106]. Some of the commonly used methods used in literature for the training of the conversion function are described below.

4.5.1 Mapping Codebooks

One of the earliest approaches adopted for the voice conversion systems is a technique known as Mapping Codebook [81, 4]. The codebook entries, or codevectors, of the source codebook have a one-to-one correspondence with the entries in the target codebook. The speaker specific codebooks are generated by the use of a VQ algorithm such as k-means [56] or LBG [57]. The VQ algorithm partition the feature space into non-overlapping regions and all the feature vectors which fall into these regions are represented by the centroid of the region. A histogram is generated by measuring the one-to-one correspondence of the source and target codevectors by using the DTW algorithm. The histogram is then used as a weighting function to produce converted source vectors by a weighted linear combination of the target codevectors. Figure 4.4

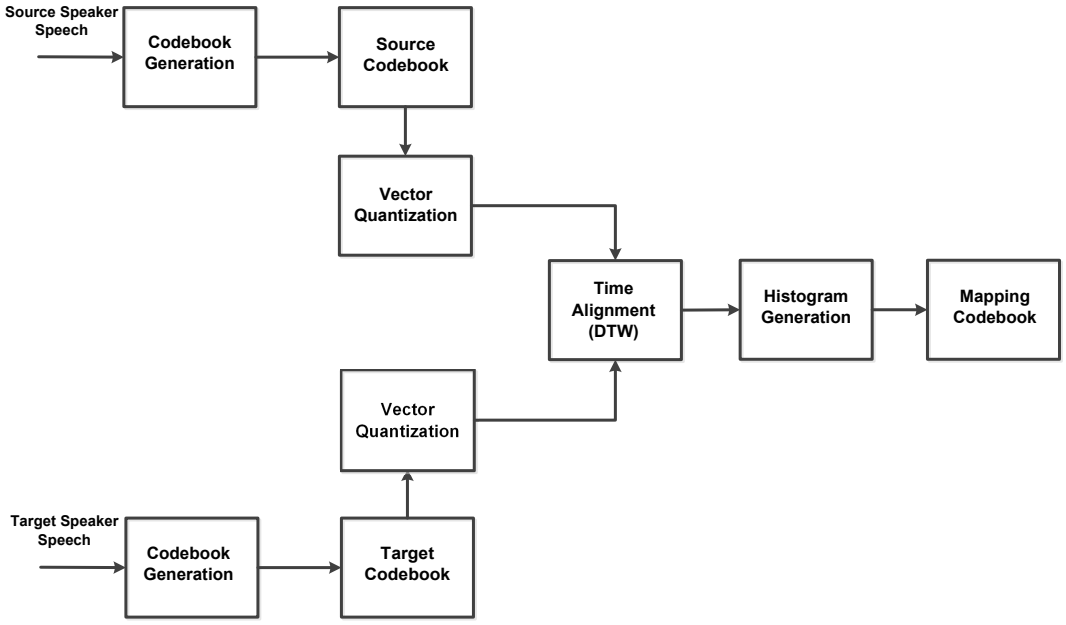


Figure 4.4: Vector Quantization based Voice Conversion [4]

shows a block level diagram of a mapping codebook based voice conversion system. This technique, however, has a fundamental problem that the entire feature space is represented by a discrete set of codevectors, resulting in discontinuities in the converted speech signal. Researchers have proposed several methods to reduce the discontinuities. One such method is the technique called weighted or fuzzy-VQ [11]. According to this method the feature vectors are represented by a combination of neighbouring codevectors instead of just a single codevector. This leads to an improved quality of the converted speech as the discontinuities in the feature vector stream are significantly reduced.

4.5.2 Discrete Conversion Function

Several researchers have proposed local functions for representing the relationship between the source and target feature vectors. These functions are considered local as they represent the relationship between the source and target feature space of one class of speech sound. An example of a discrete conversion function is DFW [103]. The proposed algorithm consists of two conversion approaches: linear regression and DFW. The optimal values of both the linear regression and DFW are calculated for each class. This method, however, fails to remove all the speaker specific characteristics for speaker independent vowel normalization [103, 113]. *Pitch Synchronous Overlap Add* (PSOLA) is a method that allows modifications of $F0$ values along with the conversion of spectral envelopes. PSOLA involves extracting and converting the parameters representing the spectral envelope of the source speech signal at Glottal Closure Instant (GCI), also known as the pitch marks. The discrete conversion functions can produce an infinite number of target feature vectors. However, the performance is degraded due to the

discontinuities in the output speech which occur as a result of the discrete nature of the conversion.

4.5.3 Continuous Conversion Function

In order to deal with the discontinuities arising in the discrete conversion function, researchers have proposed various *continuous* conversion functions. An Artificial Neural Network (ANN) is an example of a continuous conversion function. ANN with a non-linear hidden layer(s) have the ability to model any arbitrary mapping [53, 52]. ANN with back propagation have been used to transform the formant frequencies and have been shown to generalize the unseen data properly.

GMM have been used by several researchers as a probabilistic approach to feature mapping. One of the best known technique employing GMM for voice transformation was presented by [82]. In this approach the parameters of a mixture of locally linear conversion functions are estimated through the solution of normal equation for a least squares problem representing the correspondences between the source and target speakers feature vectors. It has been shown that the GMM is as good as or better than the other voice conversion techniques e.g. ANN, VQ, fuzzy-VQ and linear regression [114]. GMM have also been computed from the joint density estimates of the source and target feature vectors [98, 112]. Estimation of the GMM parameters from the joint density allows for a more judicious allocation of mixture components and have been shown to reduce the problems in numerical computations during the inversion of large and ill-conditioned matrices.

The following section details the process of voice conversion and the transformation of the spectral envelope starting with a description of the speech database used in this thesis for voice conversion.

4.6 Spectral Envelope Conversion

A typical voice conversion system and its main components were presented in the previous section. The section briefly described the different methods which are presented in the literature for the training of the conversion function. This section describes the implementation of a voice conversion system based on the transformation of the spectral envelope parameters. The spectral envelope conversion is performed on the parameters representing an all-pole model, using a conversion function based on a Gaussian mixture regression model. The speech database, extraction of features and the training of the transformation function are explained in this section, beginning with a description of the speech database used.

4.6.1 VOICES Speech Corpus

The VOICES speech corpus was designed by Kain *et al.* [92]. The corpus consists of 12 speakers, each reading 50 phonetically rich sentences. The sentences have been taken from the TIMIT [76] (Section 3.8.1.1) and the Harvard Psychoacoustics Sentences [115]. The recording of the sentences was carried out in three stages. In the first stage the speakers were asked to read the prompted sentences naturally resulting in sentences that were not constrained in timing or intonations. In the second stage the speakers were told to listen to the utterance spoken by a template speaker and then to mimic the sentence on their own. Stage 3, the speakers were asked to listen and speak along with the template speaker's speech and then a recording was made of the same sentence immediately afterwards. Recording of two mimic sentences provide an opportunity to estimate the intra-speaker variability. The speech waveform and the corresponding laryngograph signal were recorded simultaneously, at 22 kHz with 16-bit encoding, for free and mimicked versions of each sentence. Pitch marks, calculated from the laryngograph signal, and time marks, the output of a forced-alignment algorithm, are packaged with the corresponding waveforms. The provisions of time marks assist in finding the proper phonetic correspondences between speech produced by different speakers.

For the training of the voice conversion system, out of the 50 sentences per speaker, 40 sentences are used for the training and 10 sentences are used for the testing of the system. The 50 sentences amount to 5 minutes of speech data per speaker, resulting in approximately 15,000 features. Each speaker is used as a source and target twice. Out of a possible 90 speaker 5 combinations each for male-male, male-female, female-male and female-female speakers are used as source-target pairs.

The selection of speakers is followed by the analysis of the speech waveform to extract the features representing the speech spectral envelope. Feature extraction aims to reduce the amount of speech data needed for processing while providing an efficient and effective representation of the properties of the speech signal. The following section describes the pitch-synchronous analysis of the speech waveforms for the extraction of parameters representing the spectral envelope.

4.6.2 Analysis

This section details how the speech parameters representing the spectral envelope signal are extracted from the speech signal. The analysis of the speech waveforms begin with the removal of silence from the beginning and the end of the speech signal. The silence regions are removed using the *sox* utility, which is a freely available open source program. The speech waveforms are sampled at 22 kHz with a 16-bit encoding as mentioned in the previous section. The database used in these simulations also provides pitch marks that are computed from the corresponding laryngograph signal.

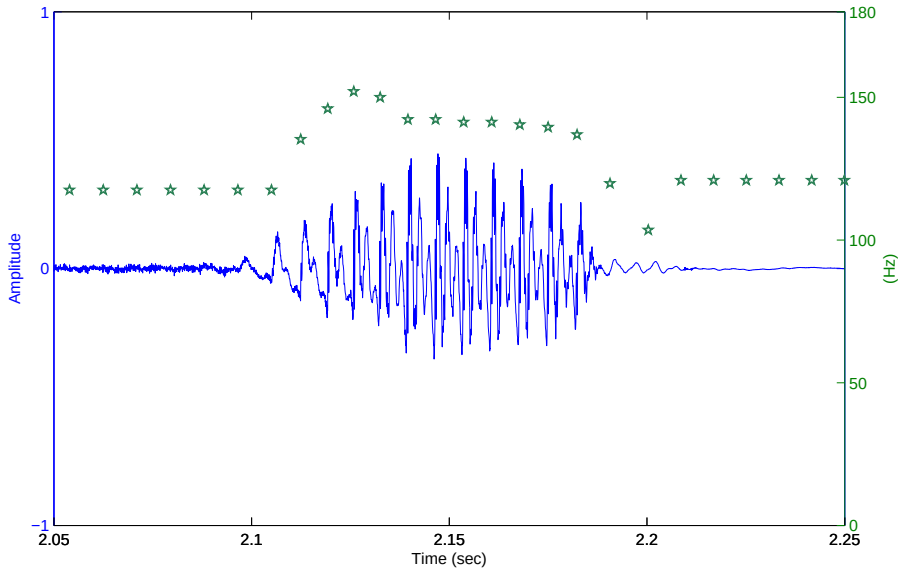


Figure 4.5: A segment of speech signal with the corresponding pitch marks in the voiced and unvoiced regions

The speech waveforms are analysed, processed and synthesized considering small segments of the speech waveform at a given time. This results in the speech signal being partitioned into small overlapping frames, s_n. The speech frames are computed synchronously with $F0$, a process known as the pitch-synchronous analysis. Each frame is two pitch cycles long, centred on the current pitch mark.The database contains the pitch marks for the voiced segments of the speech signal. During the simulations, the provided pitch marks are extended to the unvoiced regions with a constant frame update rate of 125 Hz.. Figure 4.5 gives an example of the speech signal and the corresponding pitch marks. Any errors at the beginning and at the end of the frames are not significant at either the analysis or synthesis stage since the successive frames overlap with each other.

The perceptual quality of the speech analysis/synthesis systems can be improved by considering the non-linear frequency resolution of the human ear to soundwhich is greater for lower frequencies than for the higher end of the spectrum [116]. A scale that represents this property is the Bark scale. The relationship between the BARK scale frequency f' (Bark) and the linear frequency f (Hz) is given as [92]:

$$f' = 6 \log \left(\frac{f}{1200} + \sqrt{\left(\frac{f}{1200} \right)^2 + 1} \right) \tag{4.1}$$

and the inverse relationship is given by

$$f = 600 \left(e^{\frac{f'}{6}} - \frac{1}{e^{\frac{f'}{6}}} \right) \tag{4.2}$$

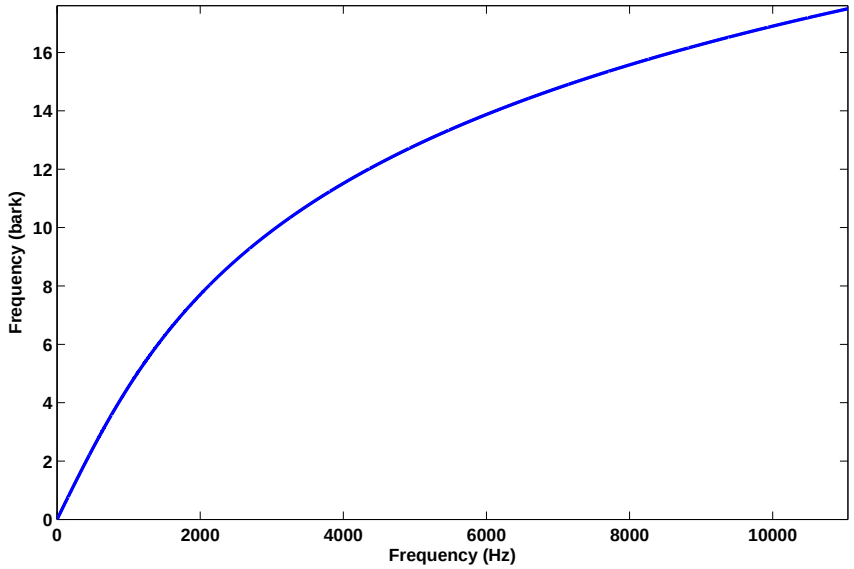


Figure 4.6: Frequency Conversion Between Bark and Linear Scale

Figure 4.6 shows the frequency conversion between the linear and the Bark scale.

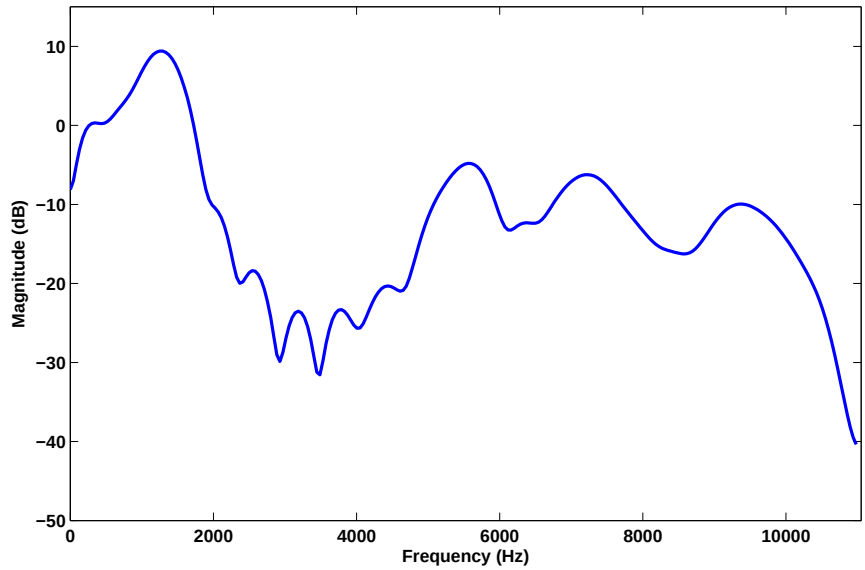
An all-pole model fitted with the BARK scale representation of the spectrum has higher resolution at the lower end of the spectrum, with a loss of detail at the higher frequencies. The non-linear spectral warping of the speech spectrum has been used successfully in the literature for speech coding and spectral modification tasks [5, 117]. The warping of the spectrum is carried out by re-sampling the magnitude spectrum according to the BARK scale warping of the linear frequencies using cubic spline interpolation [118]. The non-linear warping of the spectrum, according to the BARK scale, is shown in Figure 4.7.

The power spectral density $S_x(\omega)$ and the auto-correlation function $R_x(\tau)$ of a real and stationary signal $x(t)$, form a Fourier transform pair.

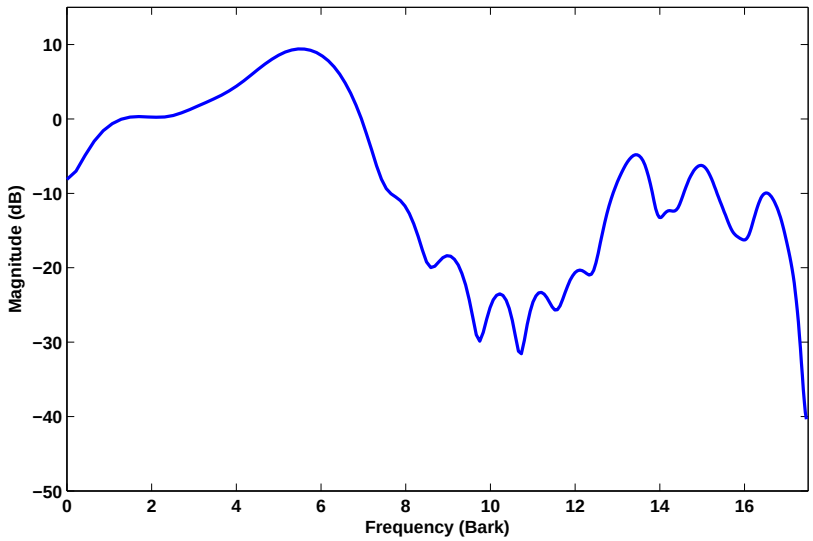
$$S_x(\omega) = \int_{-\infty}^{\infty} R_x(\tau)\, e^{-j2\pi\omega\tau} d\tau \tag{4.3}$$

In accordance with Equation 4.3, the auto-correlation sequence r_n, for the frame s_n is computed from the corresponding warped power spectrum S_n. The LPC filter coefficients α_k are computed by applying the Levinson-Durbin algorithm to the auto-correlation sequence r_n. The linear prediction analysis of the speech signal for the extraction of the LPC filter coefficients was described in Section 2.4.3. The all-pole model fit is displayed in Figure 4.8 for the warped and unwarped spectra.

The computed filter coefficients α_k of the all-pole filter $A(z) = 1 + \sum_{k=1}^{p} \alpha_k z^{-k}$ are converted to LSFs, as was described in Section 2.4.3.2. The LSFs are used extensively in speech coding [119, 120] and speech compression systems [121]. Good interpolation properties of the spectral features are crucial for the voice conversion system, as the

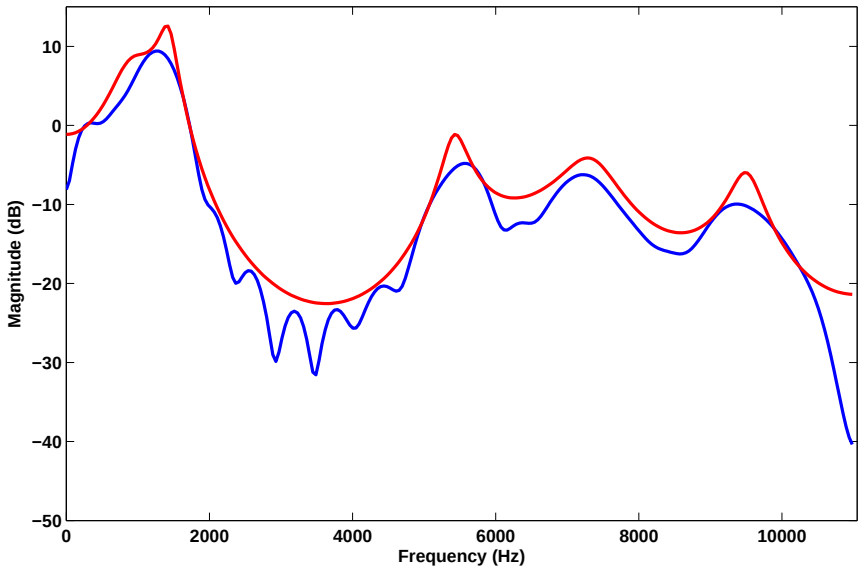


(a) Unwarped Speech Spectrum

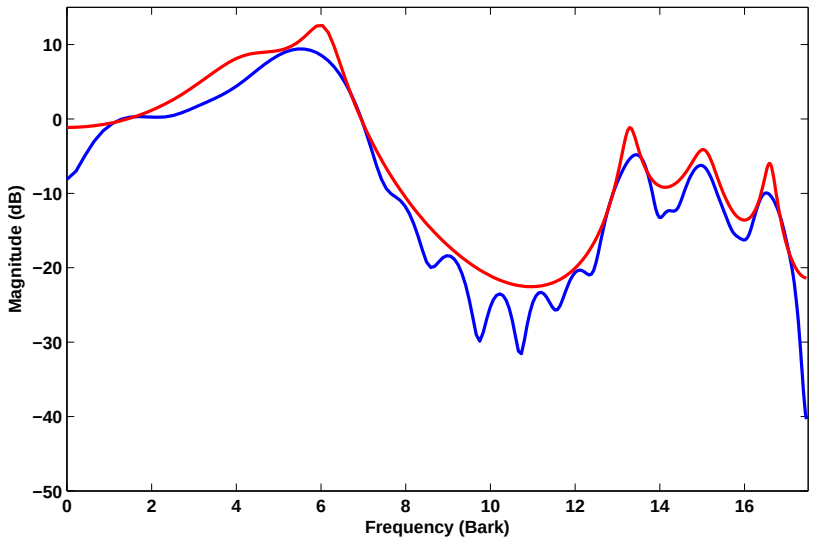


(b) Bark-Warped Spectrum

Figure 4.7: Bark-warped and Unwrapped Speech Spectrum



(a) All-pole Model Fit (red) to the Linear Magnitude Spectrum (blue)



(b) All-pole Model Fit (red) to the Warped Magnitude Spectrum (blue)

Figure 4.8: All-pole Model Fits to the Linear and Wraped Magnitude Spectra

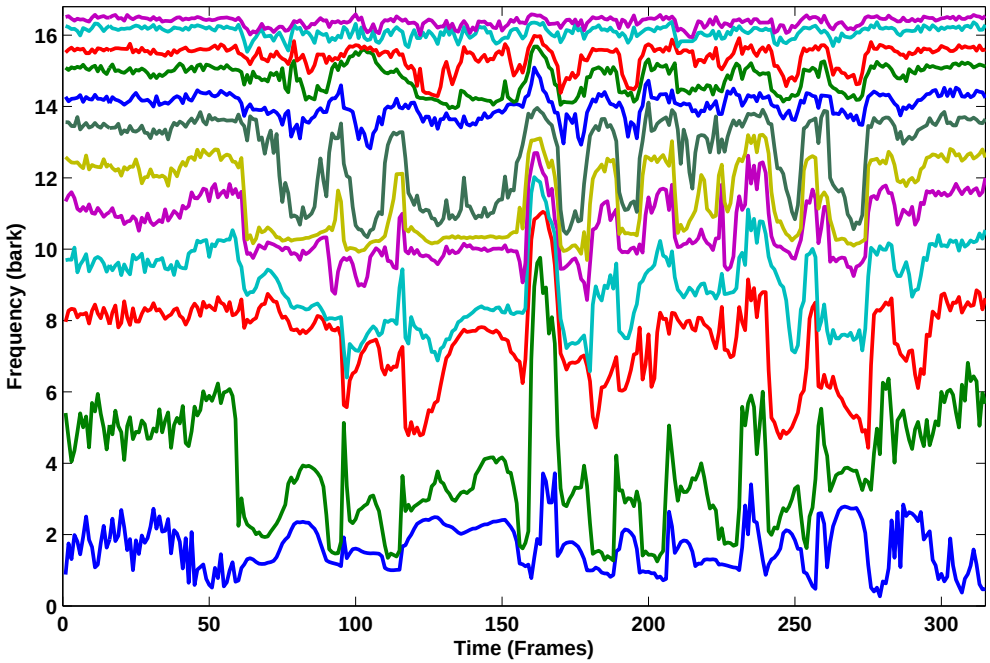


Figure 4.9: Bark-warped LSF trajectories of an example sentence '*smash light bulbs and their cash value will diminish to nothing* '

conversion function is approximated using a weighted sum of linear transformations on these features to estimate the target features.

Figure 4.9 shows the evolution of the LSF trajectories over an example sentence. LSF trajectories are closely related to the movement of the LPC poles, such as the presence of two closely spaced LSF corresponding to the presence of a spectral peak with a narrow bandwidth which indicates the presence of a formant.

For a frame-based system, features extracted from one frame represent a small portion of the speech signal. A sequence of such frames can describe a whole sentence or utterance. Due to the variations in the durations of the linguistic units uttered by different speakers, the stream of the feature vectors from the source and the target speakers must be aligned. This allows the conversion function to learn the correspondence between the source and target features representing the same phonetic content.

4.6.3 Time Alignment

Time-alignment procedures are performed on each source/target speaker pair. The purpose of time alignment is to modify the source and/or target feature vector stream in such a way that the re-arranged stream appears to be representing the same linguistic units on a frame-by-frame basis. Time-alignment was carried out by deleting or repeating the target feature vectors to match the source feature vector stream within the same phonetic content. As an alternative approach, the feature vectors from the

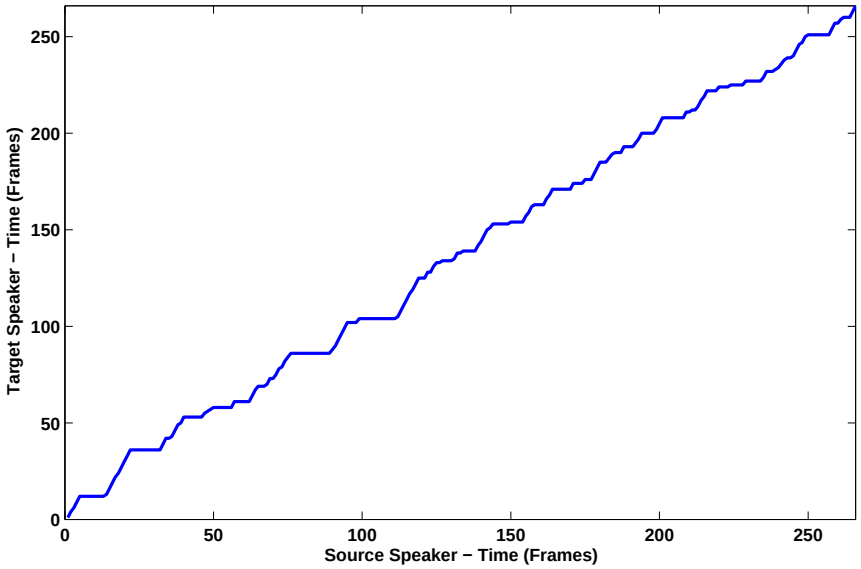


Figure 4.10: Time-alignment on an Example Utterance

shorter region of one speaker can be stretched to the length of the longer region in the other speaker's utterance. The choice of the alignment procedure does not play a significant role in finding the correspondences between the source and target feature vectors. Spectral Distortion (SD) is used to compute the differences between the source and target features vector streams. The spectral distortion measure is defined as

$$SD\left(A,B\right)=\frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(20.\log\left|S_A\left(\omega\right)\right|-20.\log\left|S_B\left(\omega\right)\right|\right)^2} \tag{4.4}$$

where A and B are the two feature stream and $S_A(\omega)$ and $S_B(\omega)$ represent the N-point spectrum of A and B.

Figure 4.10 shows the time-alignment between the source and target feature vector streams of an example sentence.

The aligned, p dimensional, N source and target vectors, x and y, are collected as:

$$X_{pN}=\left[x_s^1,\,x_s^2,\,x_s^3,\ldots,\,x_s^N\right] \tag{4.5}$$

and

$$Y_{pN}=\left[y_t^1,\,y_t^2,\,y_t^3,\ldots,\,y_t^N\right] \tag{4.6}$$

It was mentioned previously that the silences are not included in the modelling of the speech waveforms, since features extracted from the silence regions of the speech signal will model the environment rather than the speakers themselves. The silence regions are excluded from the time alignment procedure in this work. The number of vectors

accumulated in both X and Y depends upon the size of the training data available and N is larger than 15,000. It is known that the spectral envelope parameters from the unvoiced frames carry little to no speaker specific information. In these experiments the source-target aligned feature pair consisting of only the voiced frames from both the source and the target speaker are selected for training the GMM.

4.6.4 Training the Conversion Function

The conversion function aims to map the source speaker's feature vectors X to an approximation of the corresponding target feature vectors Y. In these experiments the conversion function is trained using the GMM based training kernel suggested by [82] and modified by [98]. GMM allows the implementation of a locally linear and probabilistic conversion function with the benefits of fast and accurate estimate of the fewer model parameters than conversion functions based on techniques such as the principal component analysis and neural networks [122]. GMM is suitable for the task of voice conversion [114] and has been used successfully in speaker recognition system (Chapter 3).

The parameters of the GMM are computed using the EM algorithm which was described in Section 3.5.1.2. For numerical stability, during each EM iteration a small value $\epsilon = 0.001$ is added to the diagonal elements of the covariance matrix. This technique allows regularization of the matrix density and sets a lower bound on the covariance values.

GMM can define the underlying class within each component. The correspondence between the source vectors x_t and the target vector y_t can be defined by means of the conditional probabilities. The conversion function F is estimated by computing the parameters of a GMM by modelling the joint probability density estimates of the source and target vectors x_t and y_t as $p(Z) = P(X, Y)$, where

$$Z_{2pxN} = \begin{bmatrix} X_{pxN} \\ Y_{pxN} \end{bmatrix} \tag{4.7}$$

X and Y are the aligned stream of LSF computed as the output of the time-alignment procedure. The joint density estimate takes into consideration, the observations containing both the source and target feature vector. This leads to a more judicious choice of mixture allocation [92] as opposed to the density estimation considering only the source feature vectors [98]. The linear regression used as the conversion function is given by [98]:

$$y_t^{'} = F(\mathbf{x}) = \sum_{m=1}^{M} p(\lambda_m|\mathbf{x}) \left[\mu_m^y + \Sigma_m^{YX} \left(\Sigma_m^{XX}\right)^{-1} (\mathbf{x} - \mu_m^x)\right] \tag{4.8}$$

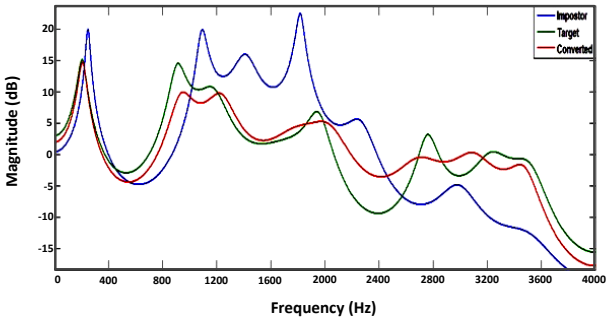


Figure 4.11: Source (impostor), Target and Converted Spectral Envelopes

where

$$\Sigma_m = \begin{bmatrix} \Sigma_m^{XX} & \Sigma_m^{XY} \\ \Sigma_m^{YX} & \Sigma_m^{YY} \end{bmatrix} \tag{4.9}$$

where Σ_m^{XX} and Σ_m^{YY} are the auto-covariance of source vectors X and target vectors Y respectively and Σ_m^{XY} and Σ_m^{YX} are the cross-covariance of X on Y and Y on X respectively for the m^{th} mixture component.

and

$$\mu_m = \begin{bmatrix} \mu_m^X \\ \mu_m^Y \end{bmatrix} \tag{4.10}$$

The conditional probability $p(\lambda_k|\mathbf{x})$ in this case is given by

$$p(\lambda_k|\mathbf{x}) = \frac{p_k N\left(\mathbf{x}; \mu_k, \Sigma_k^{XX}\right)}{\sum_{m=1}^{M} p_m N\left(\mathbf{x}; \mu_m, \Sigma_m^{XX}\right)} \tag{4.11}$$

4.6.5 Conversion

During the conversion stage, the source feature vectors X are converted to Y', which is the approximation of the target speaker's feature vectors Y. Figure 4.11 gives an example of the spectral envelope conversion.

For each frame, the extracted Bark scale LSF are converted using Equation 4.8. Only the voiced segments were used for the conversion process. The converted LSF vectors are then used to determine the LPC parameters for each frame. The LPC parameters at this stage represent the BARK warped spectral envelope. The Bark warped spectrum of

the converted speech frame is computed by multiplication of the spectral envelope with the spectrum of the source LPC residual estimated during the analysis stage. Inverse Bark warping is applied to the converted speech spectrum according to Equation 4.2 to estimate the final converted spectrum of each frame. The energy of the speech frame is normalized and made equal to the energy of the corresponding source speech frame. The synthesis of the speech waveform from the converted speech frames is described in the next section.

4.6.6 Synthesis

In order to synthesize a complete speech waveform the individual converted frames need to be grouped together. The parameters describing the speech frame are considered to be constant within each frame and in order to avoid discontinuities the frame are added by means of Overlap-Add (OLA). The OLA also allows for simple $F0$ and time modifications [123]. The computation of the k converted speech frames is followed by their weighting, overlapping and addition as follows to provide the spectral envelope modified speech waveform $\hat{s}[n]$:

$$\hat{s}[n] = w_s^{k-1}[n]\hat{s}^{k-1}[n] + w_s^k[n - T_0^k]\hat{s}_s^k[n - T_0^k] \tag{4.12}$$

where T_0^k is the fundamental pitch period for the k^{th} frame and $w_s^k[n]$ is the synthesis window function following the property

$$w_s^{k-1}[n] + w_s^k[n - T_0^k] = 1 \tag{4.13}$$

Figure 4.12 shows an asymmetric trapezoidal window that satisfies the property of Equation 4.13 and is used for the process of OLA as the complimentary synthesis window function.

The conversion of the pitch contour is carried out as:

$$f_0^{t'}(t) = \mu_t + \frac{\sigma_t}{\sigma_s}(f_0^s - \mu_s) \tag{4.14}$$

where $f_0^{t'}(t)$ is the converted source F0 values. σ_s and μ_s are the standard deviation and mean of the source instantaneous F0 values, $f_0^s(t)$, while σ_t and μ_t represent the standard deviation and the mean of the target F0 values $f_0^t(t)$.

4.6.7 Conversion Performance

The performance of the spectral envelope conversion system is dependent on two values: Number of mixture components M and the analysis order p. The effectiveness of the

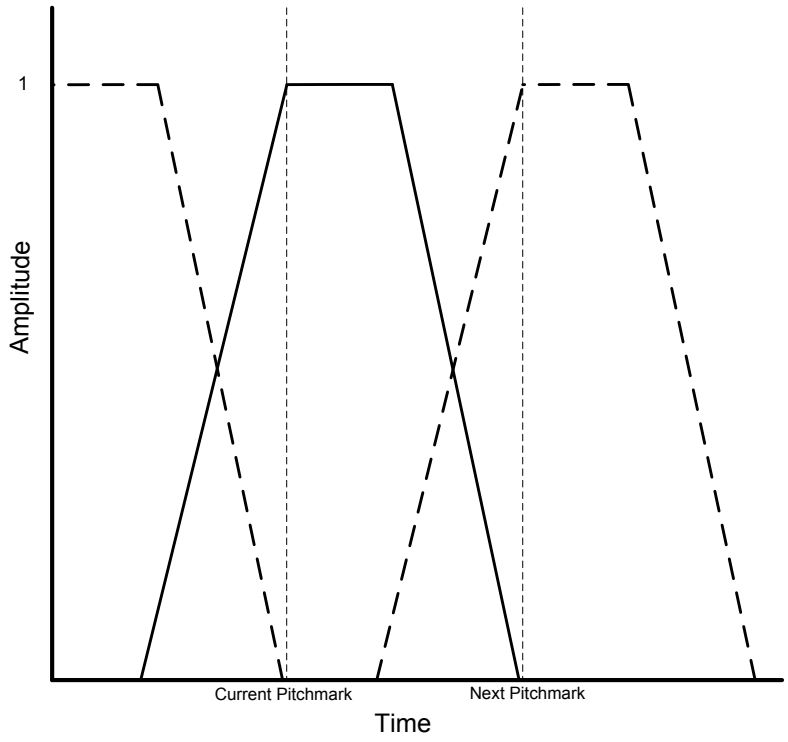


Figure 4.12: The Trapezoidal Window Used at the Synthesis Stage

conversion system was tested by selecting different combinations of both p and M. The values of M are varied between $1, 2, 4, 8, 16, 32, 64$ while the values of p are varied between $8, 12, 16, 20, 24$. In order to measure the performance of a voice conversion system, different objective and subjective measures have been proposed by researchers over the years. One of the most widely used objective measure is the Spectral Distortion (Equation 4.4). The SD represents the average spectral difference between two frames belonging to the feature vector streams of two different speakers. Figure 4.13 and Figure 4.14 depict the spectral distortion measure of Equation 4.4 between male→male, female→male, male→female and female→female source-target speaker pairs. It can be noted that the conversion error SD(Trg,Conv) decreases with an increase in the number of mixture components M, for any particular value of p.This is to be expected as an increase in the number of mixture components will result in a more accurate modeling of the underlying data. Theoretically, using each feature vector as a mixture component will result in a degenerate look-up table, with the assumption of a one-to-one function.

In a voice conversion system, the performance of the conversion is measured between the source, target and converted utterances. The performance index used for the evaluation of the voice conversion system is the one proposed in [92]:

$$P_{SD} = 1 - \frac{\text{SD(Trg,Conv)}}{\text{SD(Trg,Src)}} \tag{4.15}$$

Here SD(Trg,Conv) and SD(Trg,Src) are the spectral distortion measures between the converted-target speaker pair or the conversion distortion and the source-target speaker

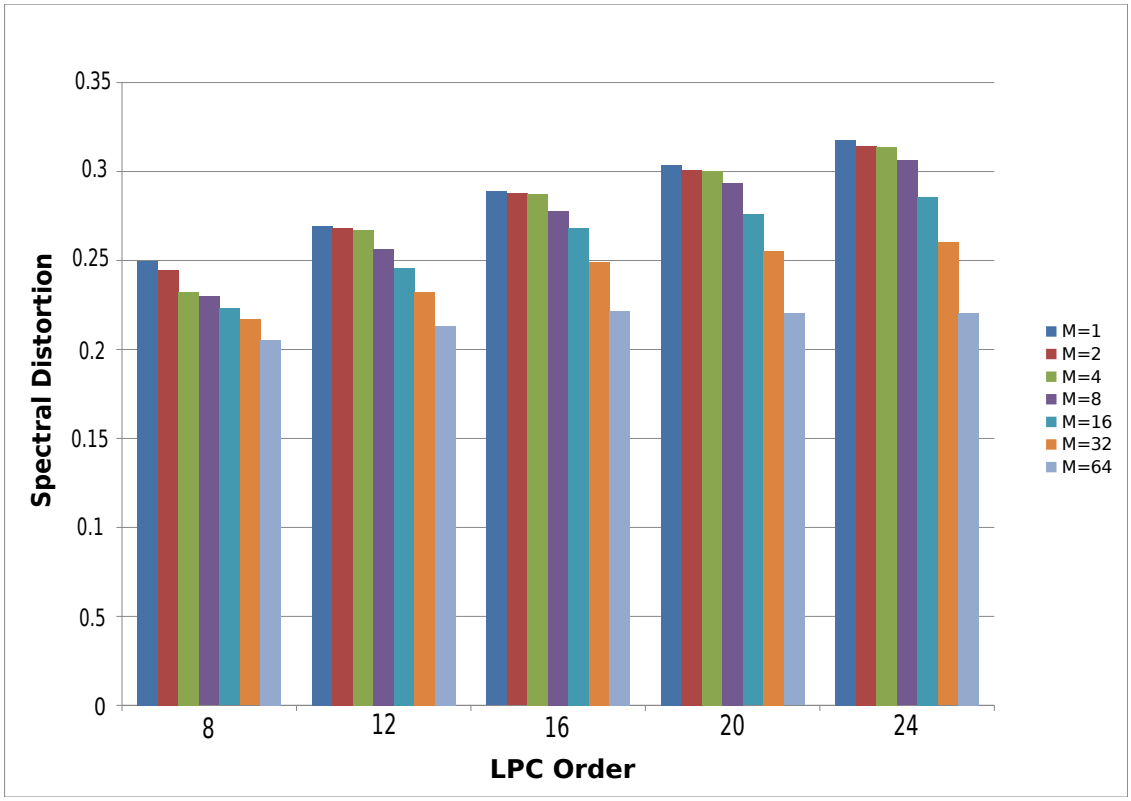
pair or the inter-speaker distortion, respectively. P_{SD} will be equal to zero if the conversion error equals the inter-speaker error suggesting poor conversion performance. Conversely P_{SD} will approach 1 when the conversion distortion approaches zero, in practice the conversion distortion cannot be equal to zero, since there are many ways in which an utterance can be spoken. Figure 4.15 and Figure 4.16 represent the conversion performance in terms of the P_{SD}. Similar to the performance evaluations obtained using the SD measure, for a particular value of p, the improvement in the performance of the system is marginal for values of $M = 1, 2$ and 4, with the highest value of P_{SD} obtained with $M = 64$ in all the cases. However, it has been demonstrated in [92] that the choice of a particular value of M is dependent on the analysis order p, and generally for a GMM using full covariance matrices values of M greater than 64 should not be used to avoid potential over-fitting problems.

This section described the voice conversion system used in this thesis. It was shown that a voice conversion system utilizing a GMM based conversion function can adequately map the source speaker's parameters to a target speaker's feature vectors. The next section describes the problem of over smoothing that arises because of the use of weighted combinations of target feature vectors to obtain the converted speech.

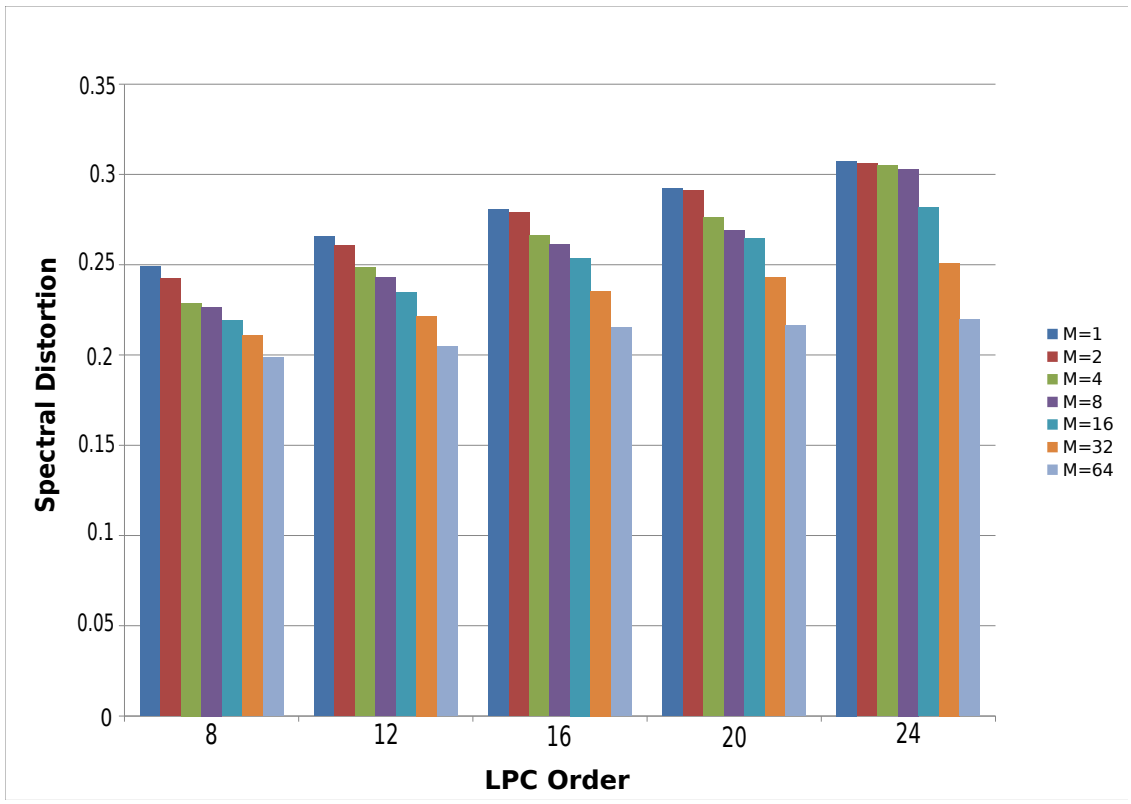
4.7 Over Smoothing in GMM based Voice Conversion

For voice conversion systems, finding a proper balance between simple and complex models for representing the source and target speech parameters presents a major challenge, especially when the amount of training data is limited. Model fitting tasks and regression commonly suffer from the bias-variance dilemma [124]. Simple models for voice conversion tasks may not be adequate to model the underlying correspondences between the source and the target feature vectors and result in the phenomenon of statistical smoothing. On the other hand, over-fitting may occur as a consequence of using complex models. Using a complex model to determine the relationships between the source-target feature vector pairs, with too many degrees of freedom, could emphasize the minor variations in the training data, resulting in poor prediction on new data while providing satisfactory results on the training set.

GMM based voice conversion systems utilize a locally linear probabilistic model of Equation 4.8 to estimate the feature vectors containing the properties of the target speaker. The converted feature vectors are obtained by a linear weighted combination of target feature vectors obtained from the target speech during the training process.

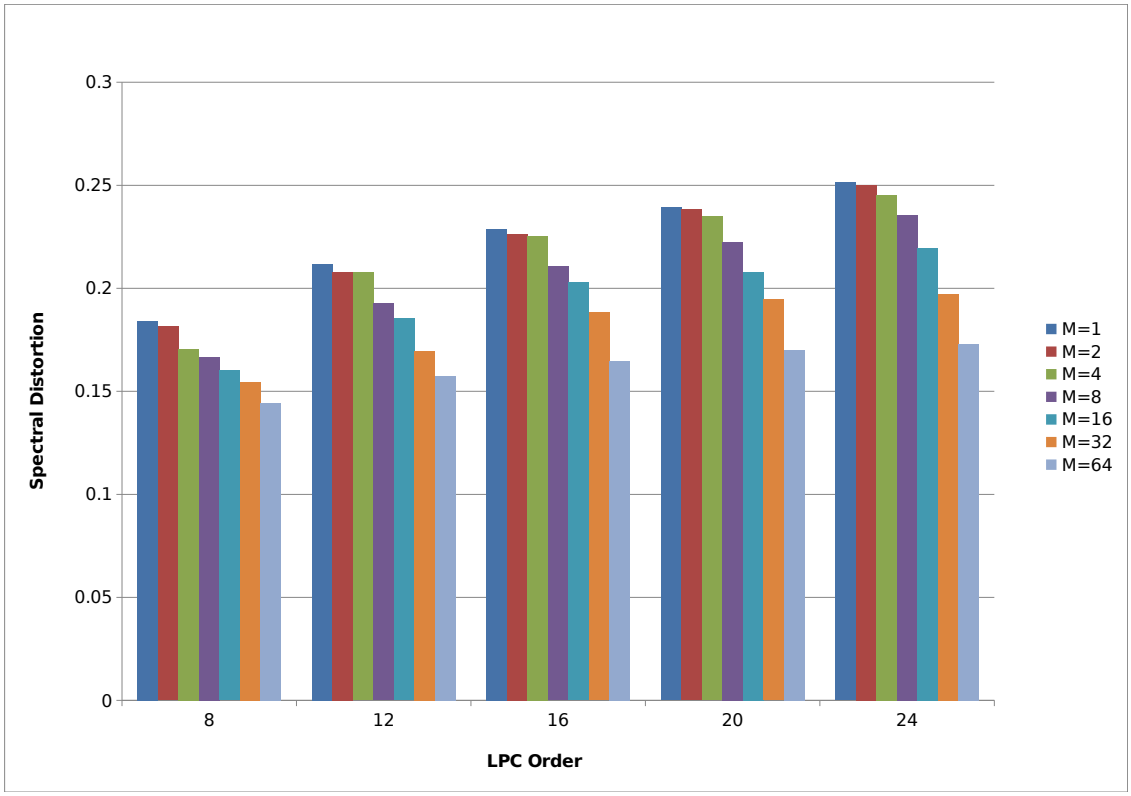


(a) SD for Male-Male Source-Target Speaker Pair

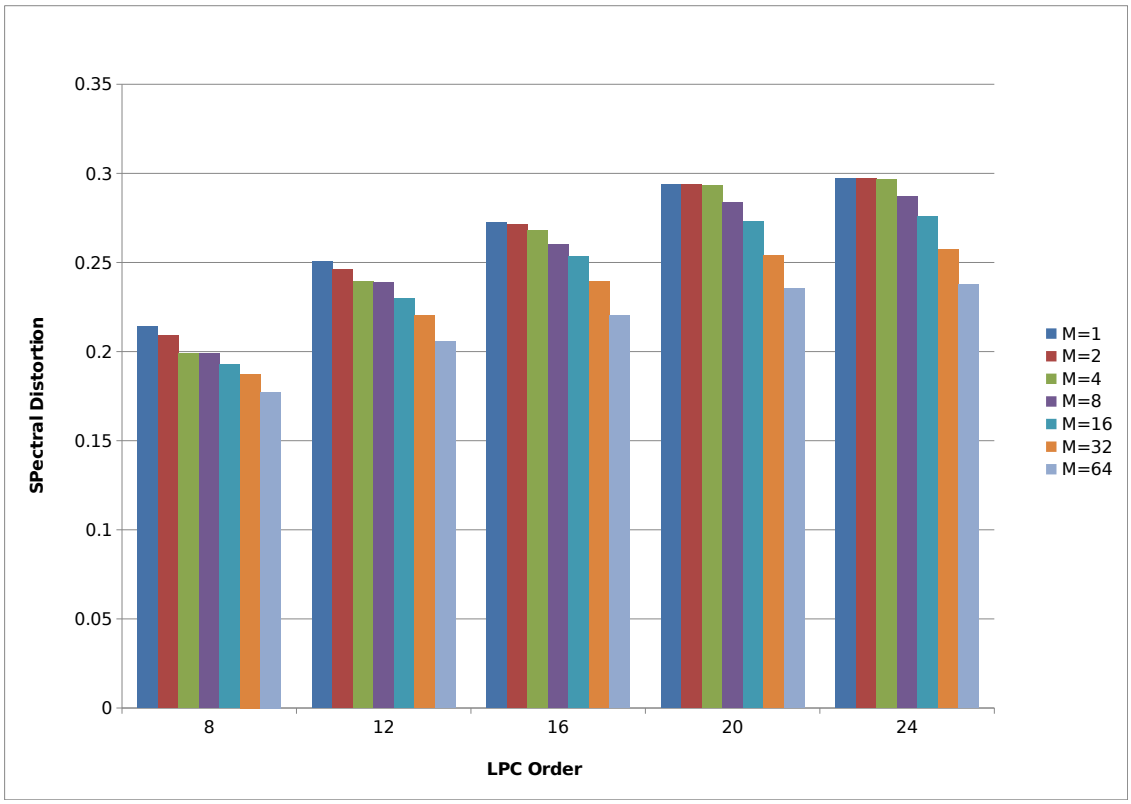


(b) SD for Female-Male Source-Target Speaker Pair

Figure 4.13: Spectral Distortion Measure for Male-Male and Female-Male Source-Target Speaker Pairs

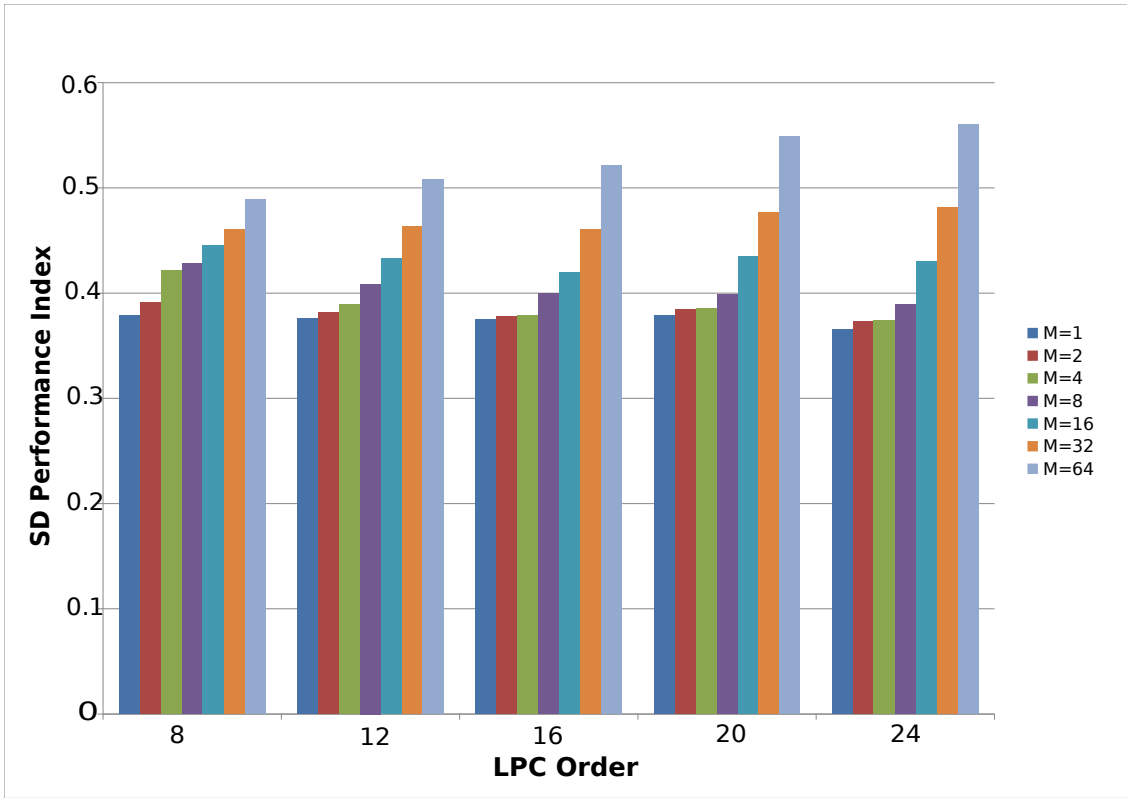


(a) SD for Male-Female Source-Target Speaker Pair

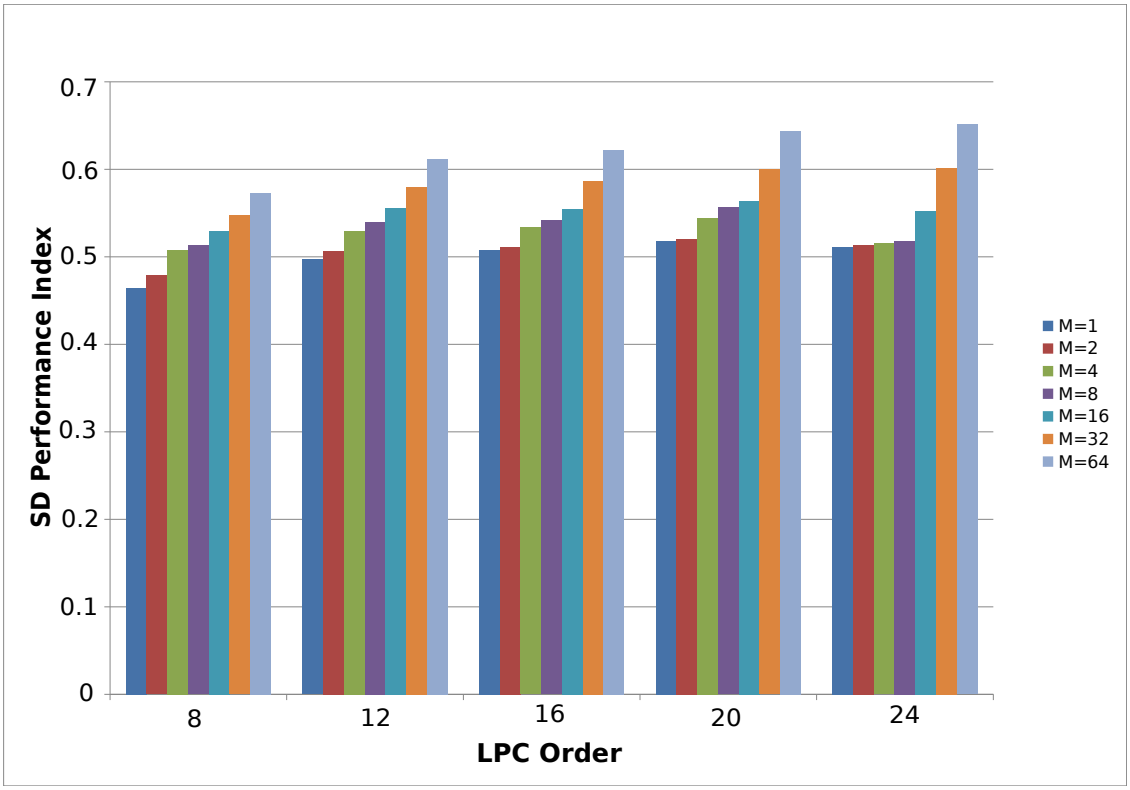


(b) SD for Female-Female Source-Target Speaker Pair

Figure 4.14: Spectral Distortion Measure for Male-Female and Female-Female Source-Target Speaker Pairs

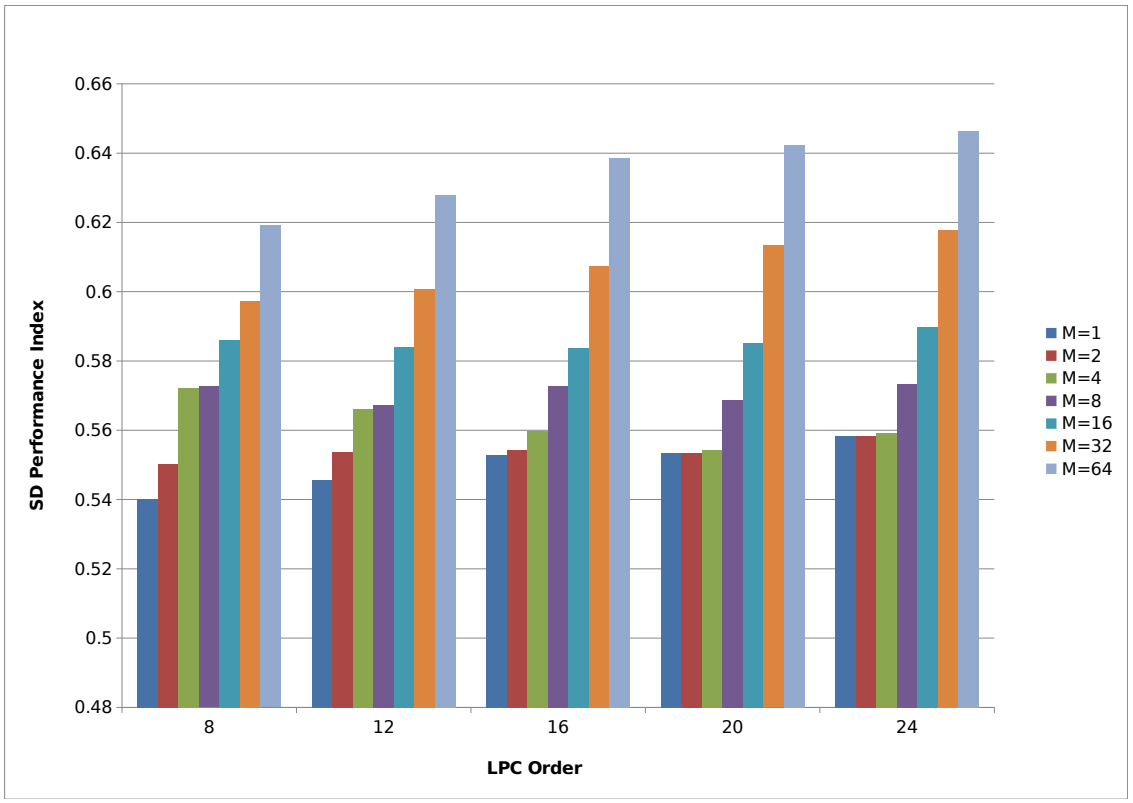


(a) P_{SD} for Male-Male Source-Target Speaker Pair

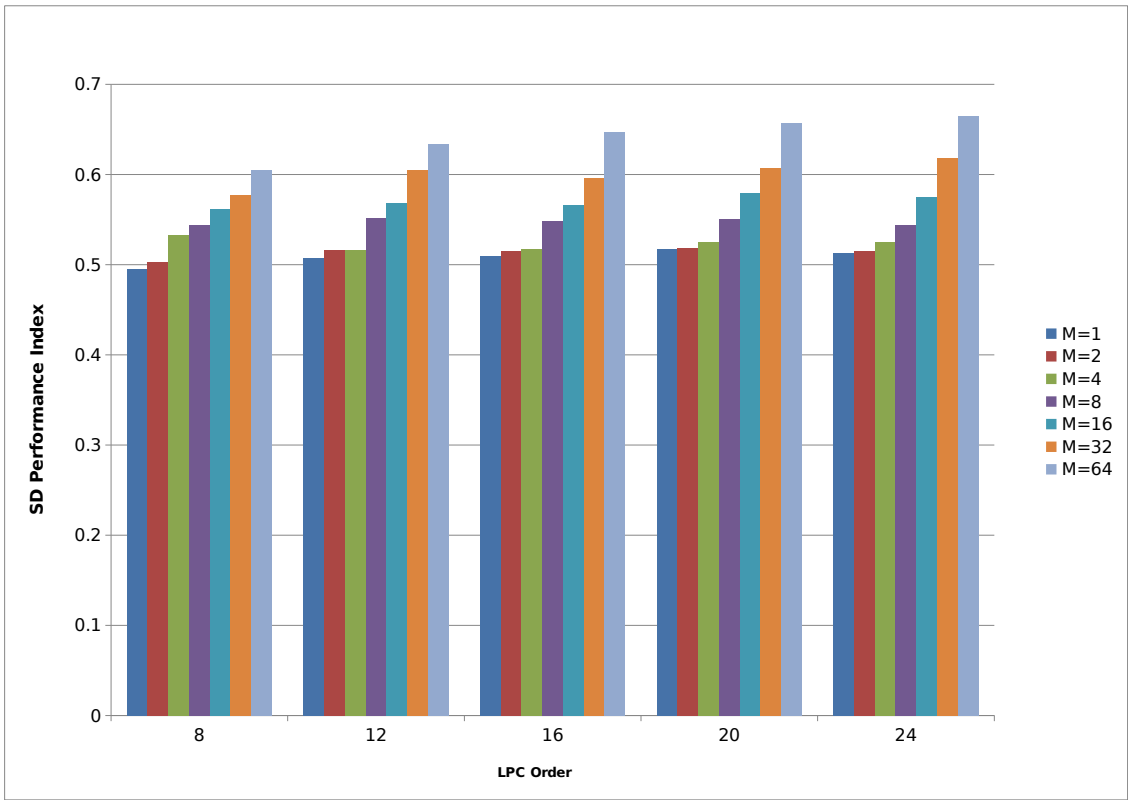


(b) P_{SD} for Female-Male Source-Target Speaker Pair

Figure 4.15: P_{SD} for Male-Male and Female-Male Source-Target Speaker Pairs



(a) P_{SD} for Male-Female Source-Target Speaker Pair



(b) P_{SD} for Female-Female Source-Target Speaker Pair

Figure 4.16: P_{SD} for Female-Female and Female-Male Source-Target Speaker Pairs

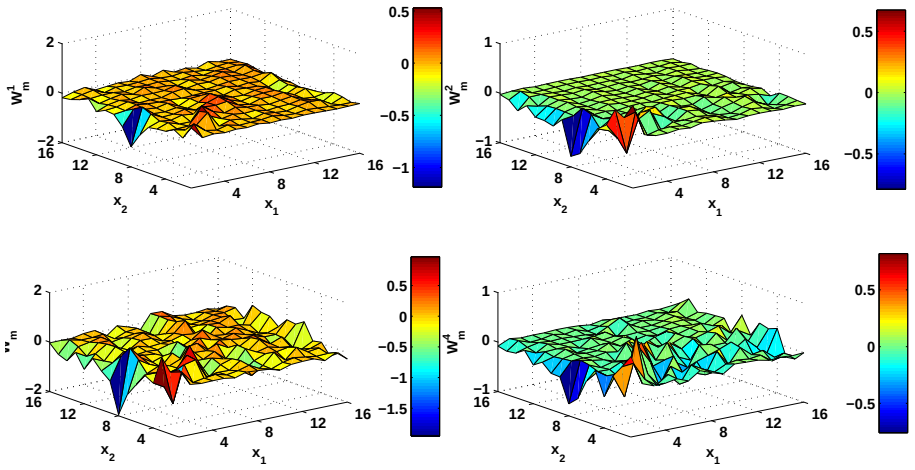


Figure 4.17: W_m for an Example Mixture Component

The conversion function of Equation 4.8 can be re-written as [122]:

$$F(x) = \sum_{m=1}^{M} p(\lambda_m|x)\left[W_m x + b_m\right] \tag{4.16}$$

$$W_m = \Sigma_m^{yx}\left(\Sigma_m^{xx}\right)^{-1} \tag{4.17}$$

$$b_m = \mu_m^y - \Sigma_m^{yx}\left(\Sigma_m^{xx}\right)^{-1}\mu_m^x \tag{4.18}$$

$$b_m = \mu_m^y - W_m\mu_m^x \tag{4.19}$$

The joint density represents the maximum likelihood estimate of the target feature vectors given the source feature vectors. The value of W_m, which is the product of the matrix representing the cross-covariance between the target and the source feature vectors and the inverse of the covariance matrix of the source feature vectors, can become extremely small for the mixture components. Figure 4.17 shows the values of W_m for a GMM with $M = 4$. A small value of W_m represents low correlation among the feature vectors of a particular source-target speaker pair. Furthermore, if the feature vectors within the same component density are linearly dependent, the inverse of the covariance matrix does not exist and hence the conversion function of Equation 4.8 cannot be used. The use of diagonal covariance matrices, instead of full covariance matrices, present a simplified alternative but the converted speech is limited in quality as it is obtained by transforming each source vector entry independently of the others.

The ill-conditioning of the covariance matrices is generally avoided by the use of variance-limiting or by the addition of a small offset vale ϵ during each EM iteration of the GMM training process. If the size of the training set is large enough i.e. $\geqslant 35,000$ vectors, it has been reported that the source and target source vectors exhibits the same covariance [125], in which case $W_m \approx 1$ and the conversion function of Equation 4.16 can be re-written as

$$F(x) = \sum_{m=1}^{M} p(\lambda_m|x)\left(x - \mu_x + \mu_y\right) \tag{4.20}$$

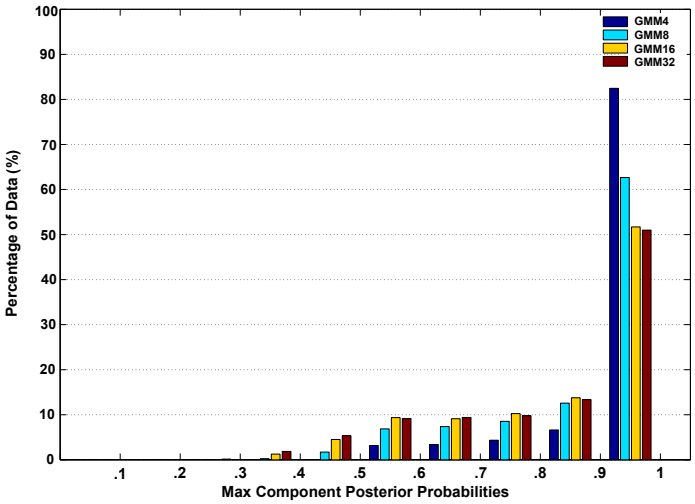


Figure 4.18: Frame-Wise Posterior Probability Ranges as Percentage of the Data for Analysis Order 24

Equation 4.20 represents a codebook type conversion. The converted vectors are represented as the weighted combination of the source vectors which are offset by the difference of the mean vectors of the source and the target component densities. The posterior probabilities are used as the weights during the combination. In practice, however, the availability of a training set with sufficient vectors to guarantee $W_m \approx 1$ presents a practical constraint for most of the target speaker pairs as it would require many hours long parallel speech material from both the target and the source speaker. Also the usage of variance limiting or variance boosting techniques, although it can ensure high enough values in the covariance matrix, can falsely increase the correlation between the source and target feature vectors leading to inaccurate regression. Therefore with limited amount of training data, it would not be possible to produce high quality converted speech using Equation 4.20.

The values of the component posterior probabilities present another challenge in obtaining high quality converted speech. A single component of the GMM is usually dominant for each frame of data. The variation in the value of the posterior probabilities for given frames depends on the dimensionality of the underlying data as well as the value of M, i.e. the number of mixture components. It can be noted from Figure 4.18 that almost 50% of the component posterior probability values lie in the range $0.9 - 1.0$, while the percentage is even higher at 80% for a 4 component GMM. This implies that the major contribution towards the determination of converted speech vectors, which are obtained as a linear weighted combination of the corresponding target feature vectors, is provided by a single GMM component for a large number of source speech frames.

4.7.1 Linear Multivariate Regression Framework

In order to deal with the problem of over smoothing in GMM based conversion systems, the values of the posterior probabilities can be utilized to single out the components that have the most and least influence in the construction of converted feature vectors from the target feature vectors. Depending upon the value of the posterior probability a hybrid solution is presented in this thesis that combines the traditional GMM with Linear Multivariate Regression framework. When the contribution of a GMM component, frame based component posterior probability, exceeds the threshold α, the conversion is carried out within the highest probability component and the components exhibiting lower posterior probabilities are discarded from the estimation process. On the other hand, frames with highest component posterior probabilities less than α are converted using the GMM based conversion function.

If $p(\lambda_k|x)$ represents the highest value of the component posterior probability for a given frame $\mathbf{x}$, with $k = \text{argmax}(p(\lambda_m|x))$, the process can be represented in the mathematical form as

$$F(\mathbf{x}) = \begin{cases} W_k\mathbf{x} + b_k & if\, p(\lambda_k|x) > \alpha \\ \sum_{m=1}^{M} p(\lambda_m|\mathbf{x})\left[W_m\mathbf{x} + b_m\right] & if\, p(\lambda_k|x) \leq \alpha \end{cases}$$

The optimal value of α is determined by maximizing the value of P_{SD}, Equation 4.15, for different values of α in the interval $[0, 1]$. Starting with 0.1 an incremental step of 0.025 was used for every iteration. The value of α has to be evaluated for every source-target speaker pair. For frames with component posterior probability values exceeding α, the conversion is carried out within the highest probability component only. Frames with component posterior probability values less than the threshold are converted using the GMM components weighted by the respective posterior probabilities.

The performance index of Equation 4.15 is computed for analysis order 8,12,16,20 and 24 with the number of components varied as $M = 1, 2, 4, 8, 16, 32$ and 64. A comparison of the traditional GMM approach with the proposed approach is shown in Figure 4.19. It can be seen that the proposed method produces better objective results for the same input speech material than the traditional GMM based conversion.

4.7.2 Temporal Variations in the Converted Speech

Referring to Figure 4.18, it can be inferred that the clustered nature of the posterior probabilities can cause rapid transitions in successive frames of the converted speech. It can be seen from Figure 4.20 that a single GMM component is dominant for a given frame of data for a GMM with $M = 4$ components. The converted speech frames are obtained as a linear weighted combination of target speech vectors with the posterior probabilities determining the mixing proportions of the target feature vectors. Rapid

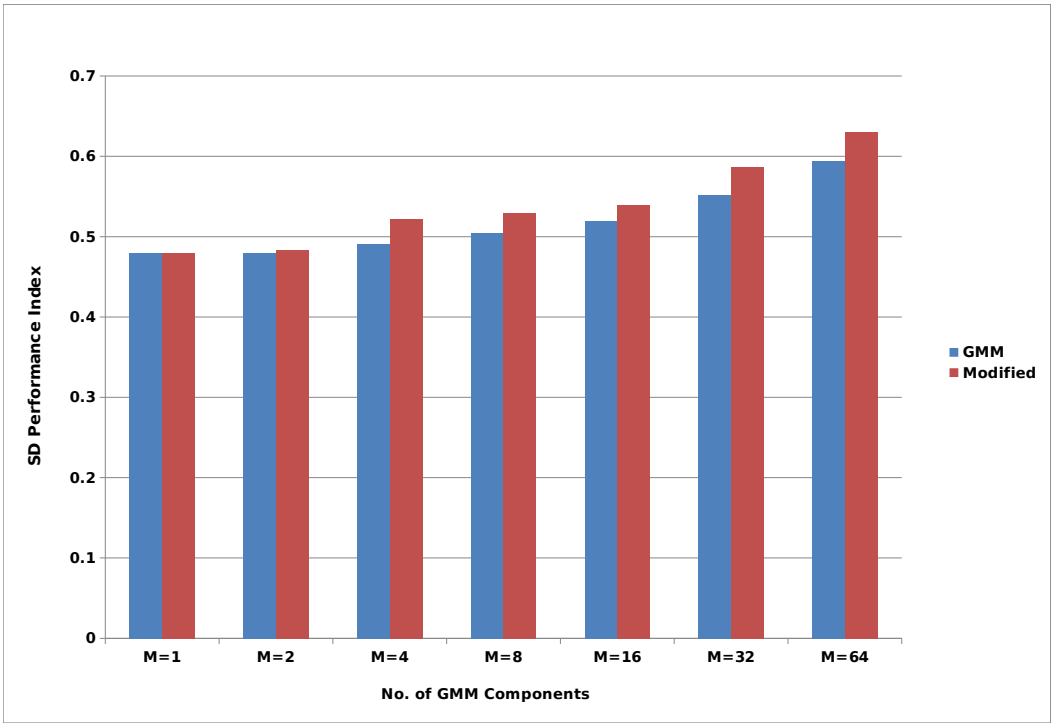


Figure 4.19: P_{SD} Comparison Plot between Conventional GMM and the Proposed Scheme

switching between different components for frames which are not far apart in time can cause audible degradations in the converted speech since different local transforms are used. The problem is compounded further if the amount of training data is limited as it would aggregate the clustered nature of the training data.

In order to deal with the temporal variations smoothing of the converted features was applied in [125]. A similar approach was presented in [126] where the use of post-filtering on the converted speech frames was suggested. However, both these techniques work on individual frames which can lead to over-smoothed features. Furthermore without taking into consideration the dependence of the features upon each other can lead to audible degradations in the converted speech.

To reduce the effects of temporal variations among successive speech frames, the mixing proportions of the constituent target feature vectors can be altered in a way to reduce the jump from one GMM component to the other in successive speech frames. In order to smooth the component posterior probabilities, a Gaussian window of length 9 has been used in this thesis with a step size of 1 and a lag of 4 samples. The use of a Gaussian filter gives maximum weight to the present values and places less emphasis on the adjacent values. The length of the kernel should not to be too short to achieve proper smoothing and it cannot be too long as it will take into account the values of the posterior probabilities that would otherwise not affect the present value. The procedure is similar to the process of computing a weighted moving average with

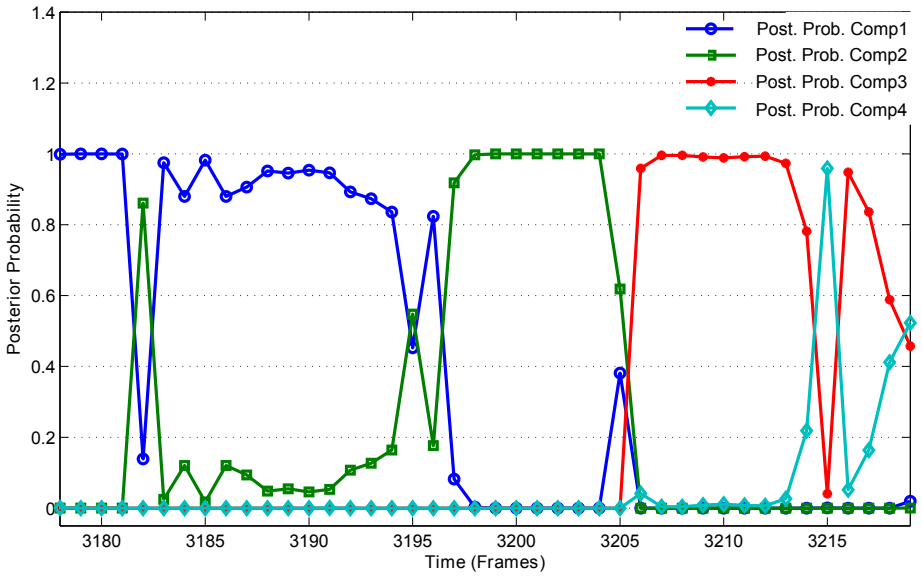


Figure 4.20: Frame-wise GMM Component Posterior Probabilities

Gaussian weights, providing a much smoother and less turbulent posterior probability plot. Figure 4.21 shows the effectiveness of the smoothing scheme by presenting the temporal derivative of the component posterior probabilities of a GMM with $M = 4$. The smoothed posterior probabilities are then updated so that their sum equals 1 and the converted source feature vectors are obtained using the updated set of smoothed posterior probabilities.

4.7.3 Subjective Assessment

In order to asses the performance of the proposal smoothing of the posterior probabilities against the traditional GMM based conversion, a subjective experiment was conducted, consisting of 12 participants. Each participant was presented with 12 sets of sentences, where each set comprised of an original target sentence, a converted sentence using the traditional GMM approach and a converted sentence using GMM-PS (GMM with Posterior Smoothing). Both inter-gender and intra-gender cases were presented in the test sets with three sets each for the male-male, male-female, female-male and female-female conversions. For each set the participants of the experiments were asked to choose the converted sentence which they found to be superior in terms of quality with reference to the original utterance. The results of these preference experiments are shown in Figure 4.22.

The listening results indicate a similar pattern of preference for the test sets presented to the participants. For both the inter-gender and intra-gender cases the GMM-PS scheme was preferred over the traditional GMM based conversion method. The scheme alters the mixing proportions of the selected GMM components in estimating the converted

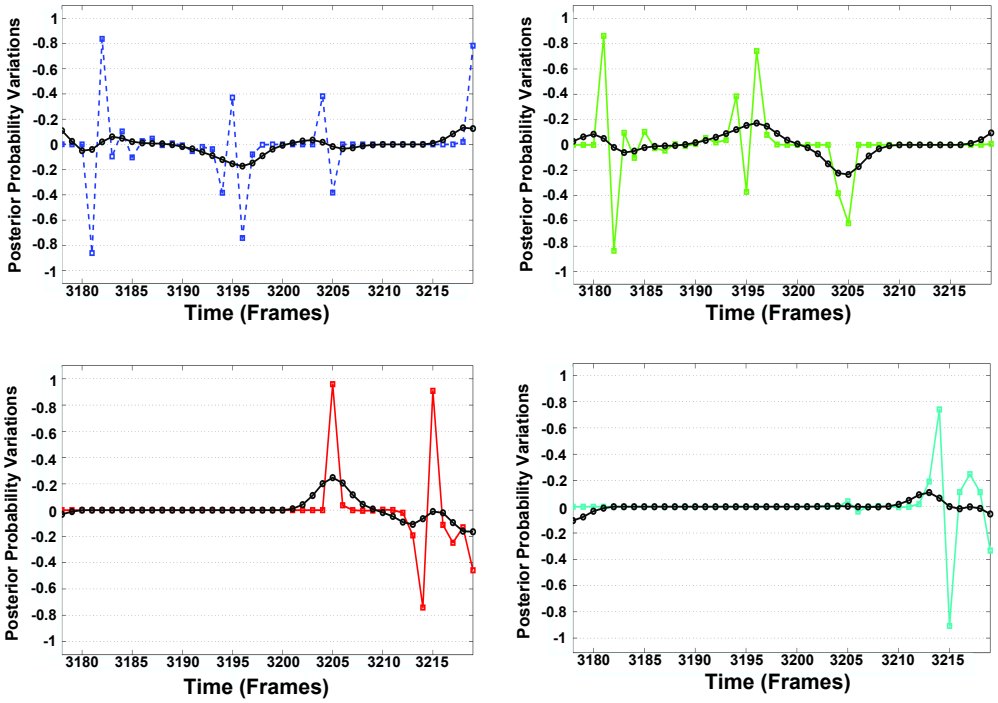


Figure 4.21: Component posterior probability temporal derivatives and their smoothed versions. Data in black represents the smoothed plot.

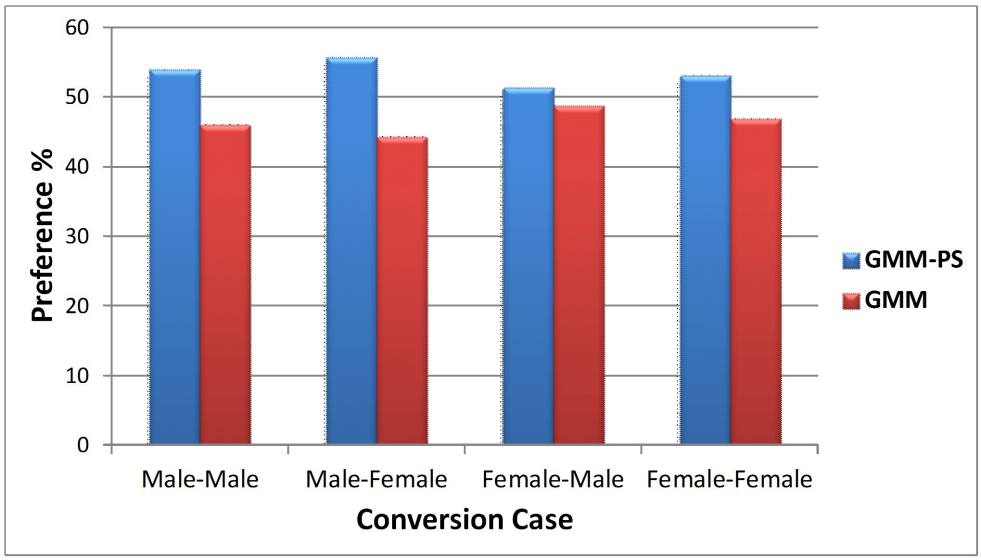


Figure 4.22: Results of the Subjective Assessments for the Converted Speech obtained using the GMM-PS method against the traditional GMM based approach

speech parameters. The subjective test indicate the effectiveness of the proposed scheme in terms of reducing the audible artefacts resulting from rapid switching among the GMM components.

4.8 Summary

This chapter presented the voice impersonation system used for converting the utterances of the source or the impostor speaker to sound like they have been spoken by the target speaker. The system utilizes speech utterances taken from the VOICES speech corpus, that comprises of 12 speakers, to train the conversion function based on the joint density estimate of the GMM. The voice conversion systems based on GMM based kernels tend to suffer from the problem of over-smoothing which is caused by the phenomenon of statistical smoothing. A solution for the over-smoothing problem was presented utilizing the linear multivariate regression by determining a posterior probability threshold. For a given frame of speech data, if the value of the component posterior probability exceeds the threshold α, linear multivariate conversion within the highest probability component is employed. For frames, where the highest component posterior probability is below the threshold, the traditional GMM technique is used. Objective evaluation using the P_{SD} demonstrated that the proposed scheme produces better conversion results compared to the traditional GMM approaches.

It was also shown that for most of the speech frame data, generally one component is dominant over the others. This leads to rapid transition in the posterior probability values among adjacent speech frames and leads to audible artefacts in the converted speech. In order to deal with the problem of rapid temporal variations, smoothing of the posterior probabilities is proposed using a Gaussian weighted moving average filter. Section 4.6.4 demonstrated that the converted feature vectors can be obtained as a weighted linear combination of the target feature vectors and by altering the mixing proportions i.e. the posterior probabilities, the rapid temporal transitions can be reduced between adjacent frames. A subjective evaluation was performed to determine the effectiveness of the proposed scheme. The subjective evaluation included both the inter-gender and the intra-gender cases. The speech obtained using GMM-PS scheme was preferred over converted speech obtained from traditional GMM method.

Having presented the speaker recognition system in Chapter 3 and the voice impersonation system in Chapter 4, the next chapter explores the effectiveness of a speaker identification system against deliberate voice impersonation attacks using computer-aided algorithms such as the ones presented in this chapter.

Chapter 5

Speaker Identification, Identity Disguise and Targeted Voice Conversion

5.1 Introduction

Speaker recognition has become a popular biometric tool for recognizing individuals from the traits of their voices in the recent years. The uniqueness of an individual's voice stems from both the differences in the physiological features of the human sound production organs e.g. difference in size and shape of the vocal tract etc. as well the variations in the sociological aspect of speech production including the use of accents and the intonation patterns. Speaker identification systems, mostly focus on the variations in the physical dimensions of the human vocal tract system. These variations are highlighted by features which are derived from the speech of an individual. Commonly used features, describing the properties of the vocal tract system, are the MFCC and LPCC and their temporal derivatives, which were presented in Chapter 3. GMM is the most widely used technique for generating speaker models based on the features representing the vocal tract characteristics and have shown to provide excellent recognition performances under clean speech environments. Present speaker recognition systems, however, make no assumption about an individual concealing his/her voice deliberately to breach the security of the speaker recognition system. The lack of focus on the possibility of identity concealment or manipulation leave the speaker recognition systems open to voice impersonation attacks both by professional voice imitators and synthetic voices generated by voice conversion algorithms.

In everyday life, the human voice impersonation can be attributed to three different aspects of human communications: in entertainment industry for impersonating a well-

known personality, acquisition of linguistic information and concealing one's identity by disguising one's voice [127]. However, human voice impersonation is not the only means of altering the properties of one's voice and concealing their identities: an automatic voice conversion system can also modify the characteristics of an impostor's voice, to match those of a target speaker. These modifications are carried out in a manner to preserve the message of the spoken text. Rodman *et al* [128] classified these different types of identity concealment by voice modification as non-electronic and electronic intentional manipulations, respectively.

Researchers have carried out studies on speaker recognition systems when dealing with identity concealment by means of imitated voices and converted synthetic voices. [129] showed the vulnerability of the speaker verification system by using different types of synthetic voices that were generated from a database of speakers enrolled in the speaker verification system. Also the impostor acceptance rates have been shown to increase in [130], when a speaker recognition system was presented with speech that was synthesized using voice conversion techniques.

A common form of voice disguise that is commonly used by speakers is the alteration of one's pitch and nasalization. A study was conducted by Kunzel *et al* [131] investigating the effect of an increased pitch, decreased pitch and the nasalization of the human voice by pinching of the nose. Their results indicate that the performance of the automatic speaker recognition system declines in these cases with the smallest degradation occurring in the case of lowered pitch voices.

Speaker and dialect imitation research have been conducted concerning the human speech perception and the automatic speaker recognition systems. It has been shown that the speaker recognition system performs better than the human listeners [132]. A comparative analysis on the automatic speaker recognition system and the human speaker perception system was carried out by [133]. The authors conducted the experiment to determine the perception of imitation by the speaker verification system with respect to the target speaker. The authors found a minimal correlation between the human listeners and the automatic speaker verification they used in their experiments.

The aim of this chapter is not to strengthen the existing experimental set up relating to the speaker identification systems but to demonstrate the apparent weaknesses in the existing speaker identification systems when dealing with computer-aided voice impersonation. The next section describes a review of professional voice imitation studies that have been reported in the literature. The later half of the chapter explores the performance of the speaker identification systems against converted voices by analysing the identification performance for identity disguise and targeted voice conversion scenarios.

5.2 Professional Voice Imitation

It is a well documented fact that several factors associated with the physical condition of the speaker such as ageing, sickness and emotional stress cause a high degree of variability in the characteristics of the human voice. Furthermore a speaker can deliberately change his/her voice by speaking in a foreign accent or talking in falsetto. These deliberate modifications on the part of the speaker vary depending on the speakers. Certain strategies employed by speakers during inter-gender voice conversions were studies by [134]. During their experiments the subjects were asked to raise or lower their fundamental frequency during recordings or use nasalization by pinching their noses. The authors observed that speakers with higher that average F_0 values were more likely to raise their fundamental frequencies. During a gender based experiment they also reported that the men are more likely to make drastic changes to their F_0 values than women who are more reluctant to vary their F_0 values.

Voice imitation can be carried out in the field of entertainment, language acquisition or for concealing one's identity by means of voice disguise [127]. For language acquisition, voice imitation is primarily used either for learning foreign and native languages, or for incorporation of various sociolect and dialects in one's speaking style for better integration into a community [127]. Language acquisition, in terms of voice imitation can be achieved in many different ways: word repetition, copying syntactic structures, reproduction of phonetic content etc. [135]. *Impersonation* is a form of voice imitation where the aim is to reproduce the characteristics of another speaker's voice [135]. Professional voice imitators, normally try to copy the most prominent features of the target speaker's voice and exaggerate them [127].

When the aim of the impersonator is to hide their identity, the changes involve modifications to the vocal tract filter settings, variation in the pitch, adaptation of a dialect or speaking in a particular accent etc. In such a scenario the goal may not be to imitate someone else but simply try and conceal their own identity. However good an impersonator is, there are certain physiological features that are difficult to modify and manipulate among speakers, and given large enough variations in these features, posing as another person by means of voice manipulation is not always possible [136]. An extreme example of the variation in these features is the differences between the female and male voices. Such a scenario involves the differences between the fundamental frequencies, shape of the glottal wave and the level of intensity in the speech waveform [137]. In order to determine the effect of gender disguise on a speaker identification system, a study was conducted by [138], in which speakers were encouraged to speak in falsetto, but an auditory analysis by the authors revealed the true gender of the speakers. Furthermore, [127] interviewed professional imitators who described their ease at imitating older voices as compared to the younger ones. In this regard it is important to determine whether having a similar voice to the target speaker is more important as compared to picking out and copying a number of features specific to the voice of the target speaker. This question was addressed by [127] by concluding that

the impersonators generally try to copy several different aspects of the target speaker's voice. A successful voice impersonation results when some of the prominent features are impersonated successfully even though the rest are not.

There are many different types of features that can be used by an impersonator for voice manipulation or identity disguise. These features can be more linguistic in nature as compared to others. For example the features can be related to a particular accent or dialect, a certain linguistic style or a selection different lexical items. Dialect disguise in speaker recognition systems was studied by [139]. Rest of the features are generally termed as more phonetic in nature such as those defining the vocal tract filter and those belonging to the voice source. The automatic speaker recognition system defined in Chapter 3 is based on the vocal tract filter parameters like others defined in the literature and will be used in analysing the effect of converted synthetic voices on the automatic speaker recognition system in the later sections. Source parameters have also been introduced recently in the state-of-the-art speaker recognition systems. These features are mostly related to the fundamental frequency and the power of the speech waveforms [140, 141]. Some of the prosodic features presented in the literature are [141]:

- Log of the number of frames per word
- Log of the number of intra-word voiced frames
- Log of the number of intra-word unvoiced frames
- Log of the mean F_0, max. F_0, min. F_0, and the F_0 range

In addition to these prosodic features, shimmer and jitter have also been proposed and used as prosodic features [142]. These features are not directly related to the prosody of an utterance but are related to the small variations in the power and frequency respectively. The use of pauses in the sentences have also been analysed by [141]. The length and rate of pauses in conversational speech depends upon the speaking rate and style of a speaker and as such are not relevant in the context of speaker recognition. After analysing the performance of the speaker recognition system against professional impersonators, various studies have concluded that the security and integrity of the speaker recognition system cannot always be breached [133, 139].

In the next section, the performance of the speaker identification system is analysed against synthetic converted voices. To analyse the robustness of the speaker identification system, it is tested against both the original and converted synthetic voices. The following section determines the performance of the speaker recognition when presented with converted synthetic voices.

5.3 Speaker Identification and Synthetic Converted Voices

Voice imitation and other form of voice disguise present a potential threat to the security of a speaker identification system. Automated voice conversion is the alteration of a speaker's voice, known as the *source speaker*, to make it sound like as if it has been uttered by a different speaker, known as the *target speaker.* A voice conversion system aims to determine a transformation function between the features extracted from the speech utterances of both the source and target speakers. The transformation function replaces the effects of the physical characteristics of the speech utterance without altering the message information present in the speech signal [143].

The vulnerability of the speaker recognition systems has been tested against the impostor and converted synthetic speech in various studies [129, 130]. In [129] the authors conducted experiments to deceive the state-of-the-art speaker verification system in accepting the speech of an impostor by the use of various converted and impostor speech utterances. In [130] the authors presented converted synthetic speech utterances, created specifically to alter the characteristics of the source speaker to match those of the target speaker, to the speaker recognition system. In this case the authors reported an increase in the impostor acceptance rates of the automatic speaker recognition system.

This section analyses the performance of the state-of-the-art speaker identification systems against the converted synthetic voices. The voice conversion system of Chapter 4 is used to generate the converted synthetic voices. The voice conversion system is based on the GMM modelling of the speaker space and was first presented by [82] and was later improved by [98]. The voice conversion system uses the linear regression between the GMMs of the source and target speaker for the transformation of the spectral properties of the source speaker and is given by Equation 4.8. The fundamental frequency of the source speaker is modified according to the F_0 values of the target speaker speech using Equation 4.14. The system has good performance when masking the identity of the source speaker and converting the characteristics of the source speaker's voice successfully to those of the target speaker.

In analysing the performance of the speaker identification system, two aspects of a voice conversion attempts are explored. In the first experiment, the ability of the speaker identification system to identify the source/impostor using the voice conversion apparatus to disguise his/her identity is analysed. Later the scenario of targeted voice conversion is considered, where the source/impostor speaker is trying to target another speaker who is enrolled in the speaker identification system.

The material and methodology used to test the robustness of the speaker recognition system is described in the following paragraphs.

Number of Speakers	4: 2 Male, M1 and M2, 2 Female, F1 and F2
Number of Sentences	50 sentences for each speaker
Amount of Data	An average of 3 sec. per sentence
Corpus Type	Parallel corpus obtained by a mimic approach

Table 5.1: General Description of the voice conversion corpus

5.3.1 Speech Material

The VOICES speech corpus, (Section 4.4.1.1) is used for testing the robustness of the state-of-the-art in speaker identification system against converted voices. The VOICES speech corpus consists of 12 speakers from the US, each reading 50 phonetically rich sentences. The sentences have been taken from the TIMIT [76] and the Harvard Psychoacoustics Sentences [115]. For each speaker, there are three sets of sentences, obtained by three different strategies, totalling 150 sentences per speaker with 50 sentences obtained per strategy. Two male and two female speakers were used for the voice conversion experiment. The speaker identification system is trained with all the 12 speakers in the speech corpus. This was done to have a more realistic evaluation of the robustness of the speaker recognition system against synthetic converted voices. Two of the speech sets per speaker are used for the training and testing of the speaker identification system, while the third set is used for the voice conversion system for the speakers enrolled in the voice conversion system. Table 5.1 shows the general dynamics of the speakers and the speech material used in the voice conversion system.

The sentences used for the voice conversion system are the same for each speaker which allows the use of parallel training corpus for the training of the voice conversion system. The sentences have been recorded by asking the participant speakers to mimic as closely as possible by listening to speech from a target speaker. Following a mimic approach negates the presence of significant prosodic differences between the speakers, since the participants were asked to imitate the person using a neutral speaking style.

5.3.2 Speaker Identification against Converted Synthetic Voices

The 12 speakers of the VOICES speech corpus are used for conducting the experiments to determine the robustness of the speaker identification system against identity disguise and voice impersonation using converted synthetic voices. All the speakers have been enrolled in the speaker identification system. These 12 speakers form a set which will be denoted as SID set. The SID set is used for analysing the speaker identification system when dealing with identity disguise. Of these 12 speaker of the SID set, 2 male and 2 female speakers are selected to be enrolled in the voice conversion system, forming a speaker set which is referred as the VC set. For each of the speakers, two of the three sets are used in the speaker identification system for training and testing purposes while for the speakers selected for the voice conversion system the third set is used.

For each of the four selected speakers of the VC set, 50 speech utterances are used for generating the GMM which are then used to convert the original voices of these speakers to the speakers in the VC set. This resulted in 12 source-target pairs: 4 sets corresponding to the case of intra-gender voice conversion i.e. M1-M2, M2-M1, F1-F2 and F2-F1, and 8 cases of inter-gender voice conversion i.e female to male and male to female. The conversion function for each of the source-target pair was trained using 10, 30 and 50 sentences from both the source and target speakers. A total of 50 sentences are used in the speaker identification system for each of the 12 speaker in the SID set along with 50 converted sentences each for the testing of the speaker identification system for M1, M2, F1 and F2 of the VC set.

The speaker identification system used in these experiments is based on the GMM as is described in detail in Section 3.8. A total of 32 GMM components are used for each of the enrolled speakers with diagonal, nodal covariance matrices using the short-term feature vectors consisting of 19 MFCC and their corresponding delta and acceleration coefficients. The features are extracted using a frame size of 20 msec with a 10 msec overlap.

The outcome of the identification experiments were classified as:

- *Source*: the converted voice is identified as belonging to the source speaker (impostor) rather than the target speaker, meaning that the voice conversion failed in its attempt to deceive the speaker identification system.
- *Target*: the converted voice is identified by the speaker identification system as belonging to the target speaker, meaning that the impostor was successful in fooling the speaker identification system.
- *Other*: the converted speech utterances have been identified as an enrolled speaker other than either the target or the source speaker. This would suggest that the impostor was unsuccessful in obtaining the desired result from his/her attempts to deceive the system but would be seen as a security breach of the speaker identification system.

To test the performance of the speaker identification system against converted synthetic voices, two different simulations are designed. In the first simulations, the assumption is that the source speaker will disguise his/her voice by means of a voice conversion algorithm to target a speaker who is not enrolled in the speaker identification system. In the second simulation, the performance of the speaker identification system is tested when an enrolled speaker is targeted by an impostor who is also enrolled in the system. This test will help to determine the true classification performance of the speaker identification system by estimating the ability of the classifier to distinguish between the original source and target models, which are both known to the speaker identification system.

To form the basis of the experiments, 50 original speech utterances from all the 12 speakers of the SID set were used to form a closed-set speaker identification system. The identification performance of the system with the SID set is shown in Table 5.2. Due to the simplistic nature of the experiment and the relatively low number of enrolled speakers, 9 of the speaker achieved 100% identification accuracy. However there are some identification discrepancies between the speaker pair Sp4 and Sp6, M1 and M2 and between F1 and F2 which indicates some degree of similarity between the voices. This leads to an overall identification performance of 99.33%.

	Sp1	Sp2	Sp3	Sp4	M1	M2	F1	F2	Sp5	Sp6	Sp7	Sp8
Sp1	100											
Sp2		100										
Sp3			100									
Sp4				98						2		
M1					98	2						
M2						100						
F1							100					
F2							4	96				
Sp5									100			
Sp6										100		
Sp7											100	
Sp8												100

Table 5.2: Identification Matrix for the speakers enrolled in the Speaker Identification System using 50 sentences from each speaker of the SID set

The performance of the speaker identification system against the identity disguise scenario and targeted voice conversion are described below.

5.3.2.1 Identity Disguise

To analyse the robustness of the speaker identification system against identity disguise, 10, 30 and 50 converted sentences belonging to each of the 12 source-target pairs of the VC set are used. During the testing of the 12 converted source-target pairs, the target speaker is excluded from the enrolment in the speaker identification system. The aim is to determine the performance of the speaker identification system against the speech of a speaker who is deliberately trying to avoid detection by targeting the speech of a speaker which is not enrolled in the speaker identification system. The corresponding identification matrices using 10, 30 and 50 sentences for each of the 12 source-target pair of the VC set are shown in Table 5.3.

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M2	1/10	2/30	1/50
F1	0/10	0/30	0/50
F2	0/10	0/30	0/50

(a) Impostor M1

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M1	1/10	1/30	0/50
F1	0/10	0/30	0/50
F2	1/10	1/30	0/50

(b) Impostor M2

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M1	4/10	3/30	0/50
M2	0/10	0/30	1/50
F2	0/10	0/30	0/50

(c) Impostor F1

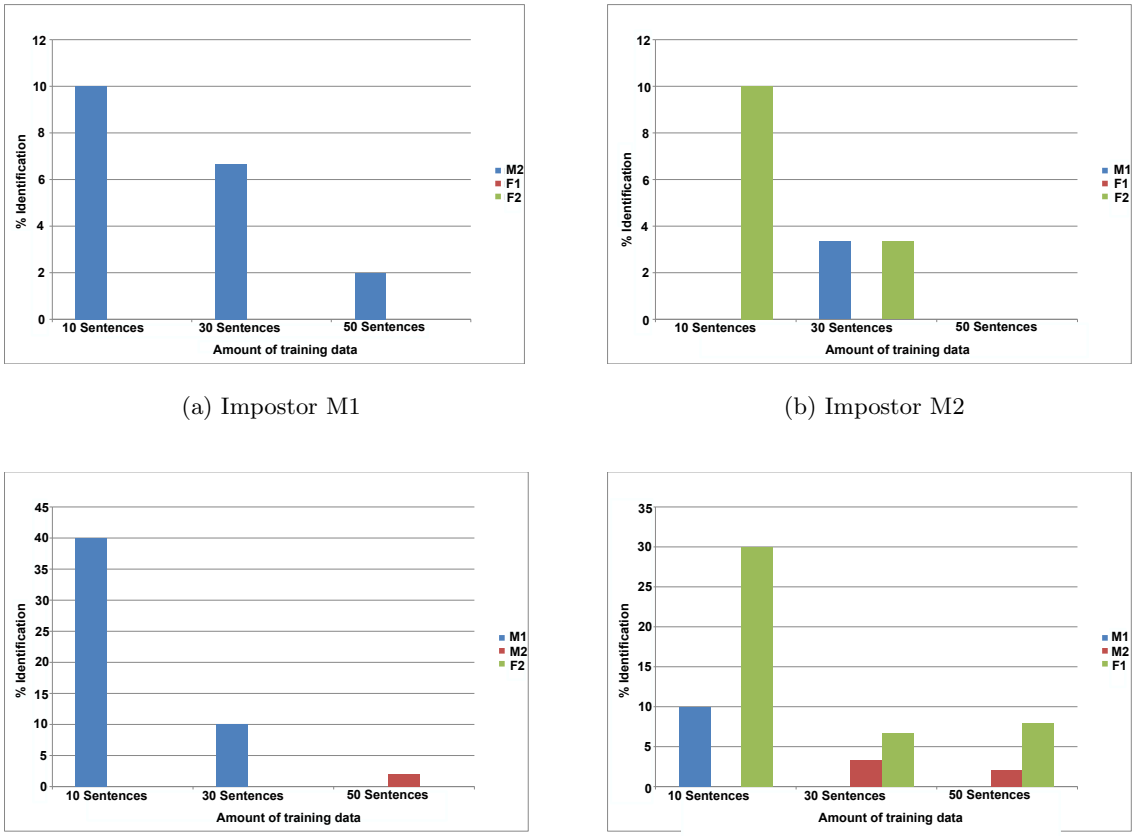
Target Speakers	10 Sentences	30 Sentences	50 Sentences
M1	1/10	0/30	0/50
M2	0/10	1/30	1/50
F1	3/10	2/30	4/50

(d) Impostor F2

Table 5.3: Results of the Identity Disguise Experiments

Table 5.3 details the identification results obtained with the SID set where the target speaker is omitted from the enrolment process and the transformation function has been trained using 10, 30 and 50 sentences for each of the source-target pair.

Figure 5.1 shows the source identification rates, where the conversion function has been trained by using 10, 30 and 50 sentences. It can be seen that the success rate of the impostor improves with an increase in the amount of data available for the training of the transformation function. In other words, the identification of the source, as the input speaker, by the speaker identification system, decreases with an increase in the data. This would indicate that the training of the conversion function has moved in the right direction, i.e. away from the source speaker space and towards the target speaker space. From Figure 5.1 it can be seen that the identification performance of the system decreases as the amount of training data used for the transformation function increases. However, there is an exception for the speaker source-target speaker pair F2-F1, where the identification of the source has increased when the transformation function has been trained with 50 sentences. This can be explained by the identification performance on the SID set from Table 5.2, which shows a strong overlap between the speaker pairs where the speaker F2 is misclassified as speaker F1. The similarity in the voice characteristics of F1 and F2 is emphasized with the increased amount of training data resulting in the increased identification of the speaker F2. The dependence of



(a) Impostor M1

(b) Impostor M2

(c) Impostor F1

(d) Impostor F2

Figure 5.1: Results of the Identification experiments on the converted voices with the target speaker omitted from the enrollment in the speaker identification system

the voice conversion success on the source-target dynamics has also been reported by [104], where the authors have shown that the voice conversion system changes with each source-target pair. This implies that the selection of the source and target pairs is important for a successful voice conversion and the intrusion of the speaker identification system.

5.3.2.2 Impersonating a Target Speaker

In the second experiment, 50 original utterances from the speakers of the VC set were used to form a closed set speaker identification system. The performance of the system against the original unmodified utterances of the speakers in the VC set is shown in the identification matrix of Table 5.4. Since the material used in the training of the speaker identification system using the SID and VC set are the same, the identification performance of the speakers M1, M2, F1 and F2 are identical in the two experiments. This experiment is aimed at analysing the performance of the speaker identification system for targeted voice conversion, where the impostor is trying to map the characteristics of his/her own voice on to the target speaker's properties. Similar to the results

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	20		
M2	10	-		10
F1	40		-	
F2	10		30	-

(a) Source Identification (%) with the 10 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	80	100	100
M2	90	-	90	80
F1	50	100	-	100
F2	90	80	60	-

(b) Target Identification (%) with the 10 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-			
M2		-		10
F1	10		-	
F2		20	10	-

(c) Other Identification with the 10 Converted Sentences on the SID Set

Table 5.5: Identification (%) of the *Source*, *Target* and *Other* identifications using 10 converted sentences

on the SID set, the speaker identification returns 100% identification performance on 2 of the 4 speakers of the VC set. However, like the SID set there is overlap between the speakers F1 and F2 and M1 and M2, suggesting similarity between the speaker of these sets, leading to an overall identification performance of 98.5%.

	M1	M2	F1	F2
M1	98	2		
M2		100		
F1			100	
F2			4	96

Table 5.4: Identification Matrix for the speakers enrolled in the Speaker Identification System using 50 sentences from each speaker of the VC set

The performance of the speaker identification system using the VC set is tested using 10, 30 and 50 converted sentences for each of the source-target pair, similar to the speaker identification experiments with voice disguise in the previous section. The identification matrices for the VC set using 10, 30 and 50 converted sentences are presented in Tables 5.5, 5.6 and 5.7. It is clear from the results of the experiments on the SID and VC sets that in most of the cases the voice conversion system succeeds in its attempt to deceive the speaker identification system. Most of the converted voice are identified as those belonging to the target speakers than the source speakers.

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	4		
M2		-		
F1	6	4	-	
F2		2	6	-

(a) Source Identification (%) with the 50 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	92	98	100
M2	98	-	100	100
F1	94	96	-	100
F2	96	98	88	-

(b) Target Identification (%) with the 50 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	4		2
M2	2	-		
F1			-	2
F2	4		6	-

(c) Other Identification with the 50 Converted Sentences on the SID Set

Table 5.7: (%) Identification of the *Source*, *Target* and *Other* identifications using 50 converted sentences

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	10		
M2	3.34	-		3.34
F1	13.34		-	
F2		3.34	10	-

(a) Source Identification (%) with the 30 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	86.67	100	100
M2	96.670	-	100	96.67
F1	83.34	100	-	100
F2	100	93.34	86.67	-

(b) Target Identification (%) with the 30 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	3.34		
M2		-		
F1		3.34	-	
F2	3.34	3.34		-

(c) Other Identification with the 30 Converted Sentences on the SID Set

Table 5.6: (%) Identification of the *Source*, *Target* and *Other* identifications using 30 converted sentences

The results of the the identification experiments on identity disguise and voice impersonation, using the SID and the VC sets respectively, indicate that most of the converted voices are identified as their respective target speakers. From the results of the speaker identification experiments it can be seen that the source and target identification rates have increased in the case of experiments on the VC set compared with the SID set.

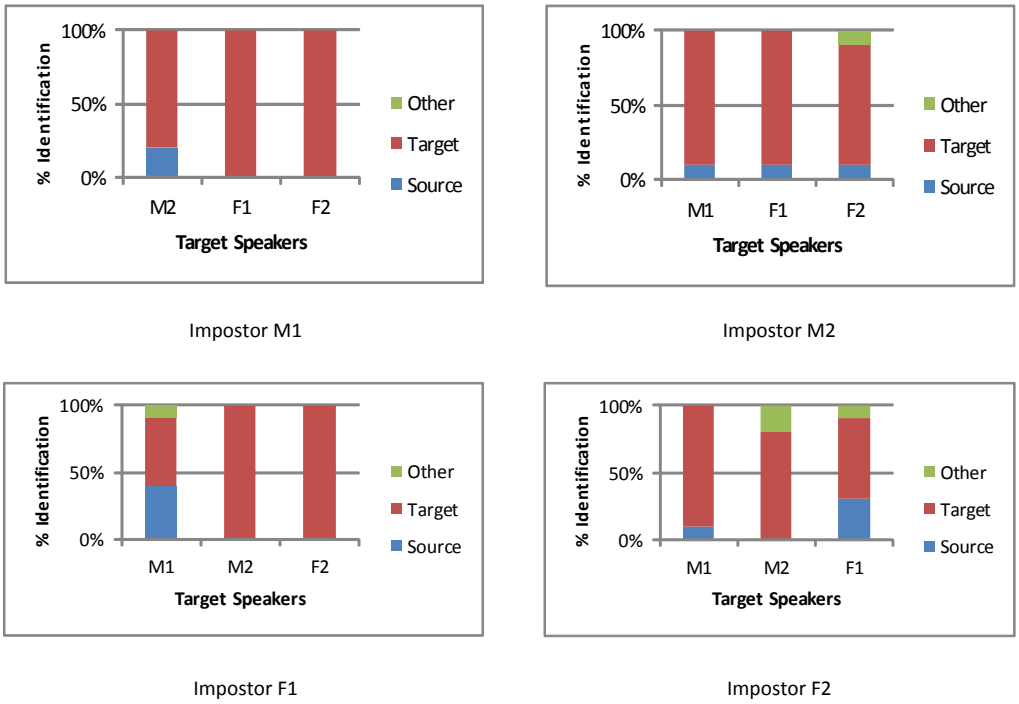


Figure 5.2: Transformation function trained with 10 sentences

This is primarily because of the lesser number of competing speaker models in the speaker identification system based on the VC set. However, this fact is accompanied by a reduction in the other identification statistics on the VC set. This suggests that the voice conversion system has performed well in its effort to deceive the speaker identification system where the relative increase in the target identification is more than the rate of source and other identification. Furthermore, it is clear that the source or the impostor speaker was highly successful in disguising their identity and impersonating a target speakers. If the aim of a voice conversion attack is to impersonate another speaker, the identification of the *other* speakers would be considered as a failure, however, if the objective was to conceal the identity of the source speaker from the speaker identification system, the identification of the *other* speaker alongside the *target* speaker, can be considered as a success accompanied by the low identification rates of the source speaker in the two sets of experiments.

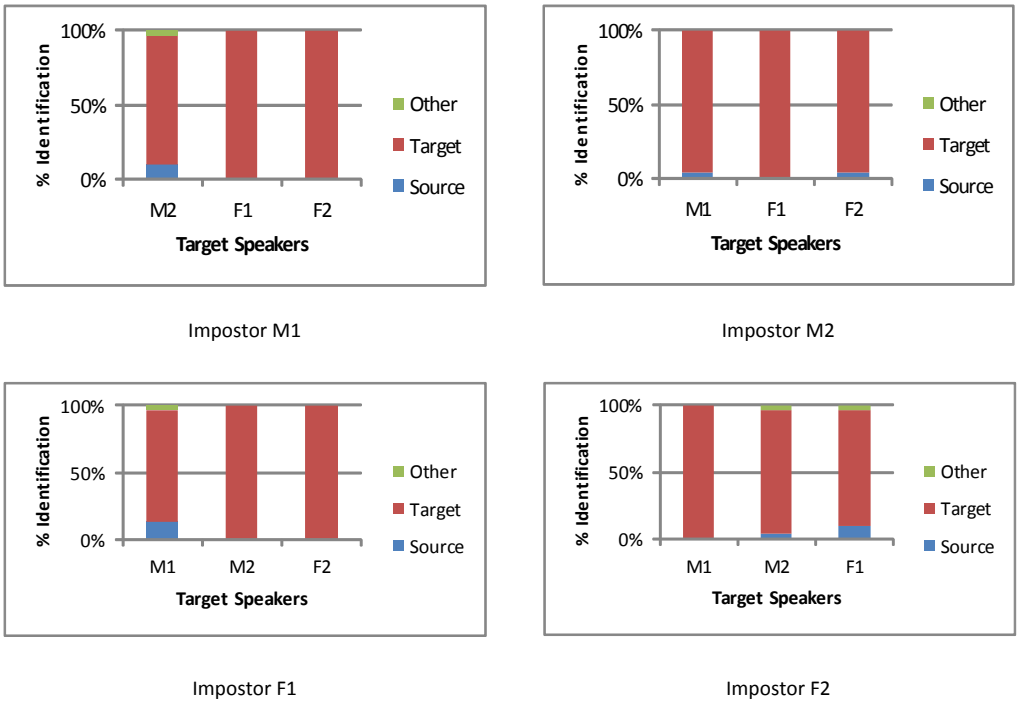


Figure 5.3: Transformation function trained with 30 sentences

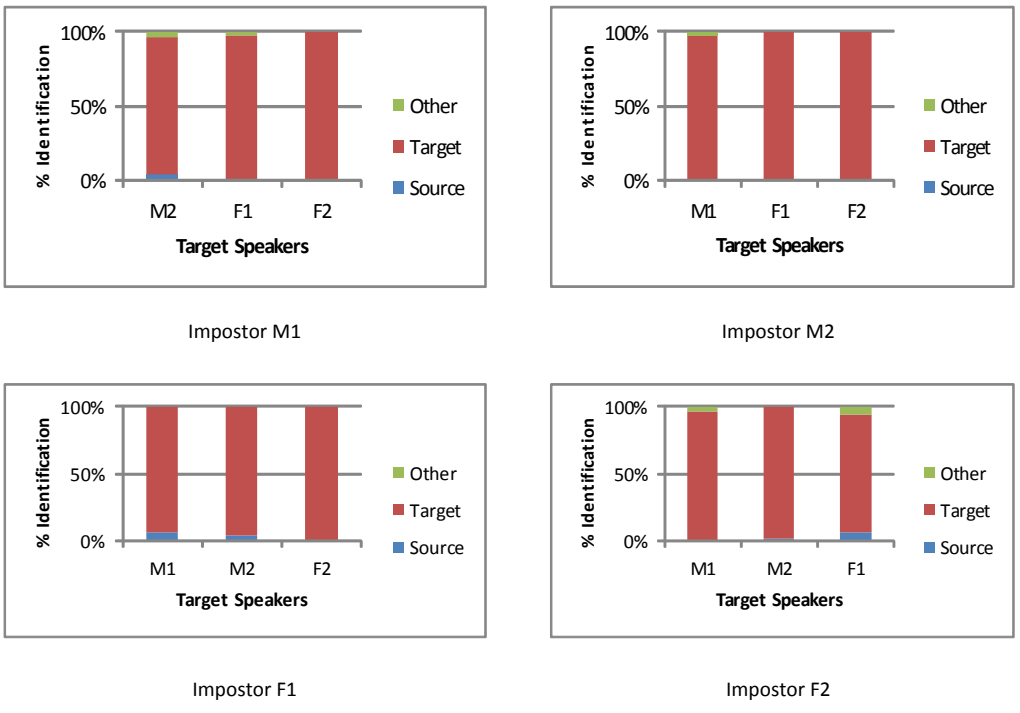


Figure 5.4: Transformation function trained with 50 sentences

5.3.2.3 Inter-gender and Intra-gender Voice Conversion

The identification experiments were also devised to test the robustness of speaker identification system against intra-gender and inter-gender voice conversion attacks. From Table 5.2, it can be observed that there exists an overlap between the speaker pair M1 and M2 and the speaker pair F1 and F2. This overlap introduces difficulties for the voice conversion system when converting M1 to M2 and F2 to F1. However, the reverse conversion discrepancy i.e. M2 to M1 and F2 to F1 is not observed in these experiments. The identification overlap between the speaker pairs M1, M2 and F1, F2 also causes the source identification percentage, which is the *correct identification*, to increase between these source-target pairs when the identification system is presented with the converted synthetic voices. The increased source identification rates in the case of these source-target pairs can be seen from Figures 5.1 ,5.2, 5.3 and 5.4 although the source identification rates drop in the case of increased training sentences used for the conversion function.

In the case of intra-gender identification, the results show that for two of the four sets of intra-gender converted voices, most of the converted voices were successfully identified as their intended target speakers, so that the identification system failed to identify the source speakers when presented with the converted voices. However, there are two cases where the source identification rates are higher in comparison to the rest i.e. M1 and M2 and F2 and F1. In other words, for these two cases the speaker identification system performs well against converted synthetic voices and the source identification rates are higher in comparison to the other two cases. This can probably be explained by the fact that the speakers M1 and F2 are highly characterized by the unvoiced segments of their voices and since the voice conversion system only converts the voiced segments of the source speech, the unmodified unvoiced segments would still be detected by the speaker identification system. It can however, be noted from the results of the intra-gender identification experiments, that the source identification rates in these two cases decrease considerably with an increase in the amount of the training data. This would suggest that in order for the conversion function to be trained properly and to achieve good results on the regression function of Equation 4.8, a large amount of training data is required by the voice conversion system.

For inter-gender voice conversion, for half of the eight sets of inter-gender source-target speaker pairs, the voice conversion system achieves a high degree of miss identification and hit conversion. This means that not only the source speakers in these cases were able to conceal their identity, which is the *miss identifications*, they were also successful in impersonating the target speakers; a successful impersonation or *hit conversion*. The other half of the inter-gender converted voices were not associated with their corresponding target speakers. One particular example is the conversion of speaker F1 to speaker M1. For this source-target pair, the percentage of source identification is relatively high as compared to the others. As in the case of intra-gender voice conversion, it is highly likely that the speaker M1 is highly characterized by his unvoiced and

since the voice conversion system does not take into account the unvoiced segments for conversion and renders the conversion function only for the voiced segments, conversion of the speaker F1 to M1 proves challenging for the voice conversion system. In this particular case the speaker identification system fares well against the converted voices and is able to achieve the a high percentage of source identification i.e. 40% using 10 training sentences. However, in all the cases of inter-gender converted voices, the miss identifications decrease with an increase in the number of sentences used for the training of the conversion function with a corresponding increase in the percentage of hit or correct conversions.

A summary of the inter-gender and intra-gender converted voices on the speaker identification system is listed in Table 5.8.

Type of Conversion	Source Identification	Target Identification	Other Identification
Inter-gender	1.67	96.67	0.84
Intra-gender	5.83	92.5	1.67

Table 5.8: Summary of the average % identification of source, target and other speakers with intra-gender and intra-gender converted voices, using 30 sentences for the conversion function training

From the results of the inter-gender and intra-gender converted voices (Tables 5.5, 5.6 and 5.7) on the speaker identification system, it can be observed that in terms of the inter-gender converted voices the percentage source identification is relatively lower than those of the intra-gender converted voices. On the other hand, in terms of correct conversion, the intra-gender converted voices, in general, achieve lower target identification results accompanied by an increase in the other identification rates, indicating an error prone conversion where the converted voices are identified as a speaker other than the source or the target speaker.

It can be concluded from these simulations that given enough speech material for the training of the conversion function an impostor would be able to deceive the speaker identification system with alarming success. Although in the case of intra-gender voice conversion, the task of voice conversion is relatively more difficult and the speaker identification system was able to identify the original author of the voice conversion attacks.

5.4 Summary

One of the main drawbacks of trying to measure the performance of automatic speaker identification systems against the converted and imitated voices is the lack of availability of databases containing converted and imitated voices. In this chapter, it was shown that increasing the amount of training data for the training of the voice conversion

function can lead to high success rates of miss identification and hit conversion. In order to test the performance of the speaker identification system, an identity disguise scenario was tested where the objective of the input speaker was to disguise his/her identity. In another scenario, the effect of impersonating a targeted speaker on the performance of the automatic speaker identification system was analysed. Ignoring the relatively small size of the speaker set used in speaker identification system, the test scenarios indicate that even with a small of training data for the voice conversion system, an impostor can easily deceive the speaker identification system.

As mentioned above the aim of this chapter was not to strengthen the existing experimental set up relating to the speaker identification system but to demonstrate the apparent weaknesses in the existing speaker identification systems when dealing with computer-aided voice impersonation.

Chapter 6

Multiple Classifier Systems and Residual based Information for Speaker Identification

6.1 Introduction

In pattern recognition applications, the systems that classify a test sample from one of the pre-specified patterns are known as *classifiers* and the pre-specified patterns are known as *classes*. In speaker identification problem, each class corresponds to a speaker. In all classifiers the input is a test sample belonging to one of the specified classes and the output of the classifier is a label describing the class associated with the pattern. Different types of classifiers exist in literature depending upon the type of the pattern classification problem at hand with each classifier carrying some advantages and disadvantages with respect to the others. Depending upon the operational conditions and the type of pattern recognition problem, the performance of a classifier is analysed on a set of test data, and the classifier is considered as a good classifier if it provides satisfactory recognition performance.

For speaker identification problems, it is difficult to develop a good classifier considering the availability of limited amount of training data, presence of noise and the high dimensionality of the feature vectors. Considering that the classifier is made up of three main components, namely the preprocessing stage, the preprocessing stage and the classification stage, a classifier is considered a good classifier if a good choice is made from a given set of possibilities, for each of these stages. Due to limitations on practical implementations, it is not always possible to have an optimal or a good classifier. For pattern recognition applications such as speaker identification, where the condition of training and testing samples and the knowledge of whether the extracted features have

been tempered with, by the use of voice conversion techniques, is not known a priori, it is difficult to select the optimality criteria for feature selection and the selection of the modeling technique. Taking into account these considerations, the performance gap between an optimal classifier and a reasonable classifier can be understood easily.

In the first half of this chapter, the use of multiple classifier systems for the task of speaker identification is analysed. The outputs of different classifiers using different feature sets are combined through various schemes. The performance of the system is analysed against synthetic converted voices in the identity disguise and targeted voice conversion scenarios.

The later half of the chapter describes the use of speaker specific information present in the LP-residual of the speech signal. The use of LP-residual based features and the spectral envelope based features are later tested against the intrusion from the voice conversion system. The performance of the system is analysed against identity disguise and targeted voice impersonation.

The next section gives a brief description of the main concepts of the multiple classifier systems.

6.2 Multiple Classifiers Systems

The main idea behind the use of multiple classifiers can be explained by considering a classifier with a given recognition performance which is less than a hundred percent, suggesting that for some test inputs the classification will be in error. Assuming that the requirement is to increase the recognition rate by building a multiple classifier system, the important question to be answered is: what type of classifier should be build in harmony with the existing ones, so that once combined the system should be able to give improved performance? It turns out that the answer to this question is not a straightforward one. However, it has been suggested that the classifiers should not make the same classification errors or in other words they should not be strongly correlated in their miss-classifications [144]. In this way, given that a classifier makes an incorrect decision about a test sample after combination, the miss classification can be compensated by the output of other classifiers in the system. In this regard, it is important that the classifiers in the system do not provide erroneous results on the same set of test samples otherwise many of the classifier combination techniques will struggle to provide the improved recognition performance required. Two classifiers are said to be *complimentary* if one classifier provides incorrect information about a test sample and the other is able to correctly classify it. Complimentary classifiers is an important subject in the context of multiple classifier systems and will be described later in the Section 6.2.5.

Figure 6.1 describes the stages in training two individual classifiers in a multiple classifier configuration. The task of combining classifiers is composed of three main parts.

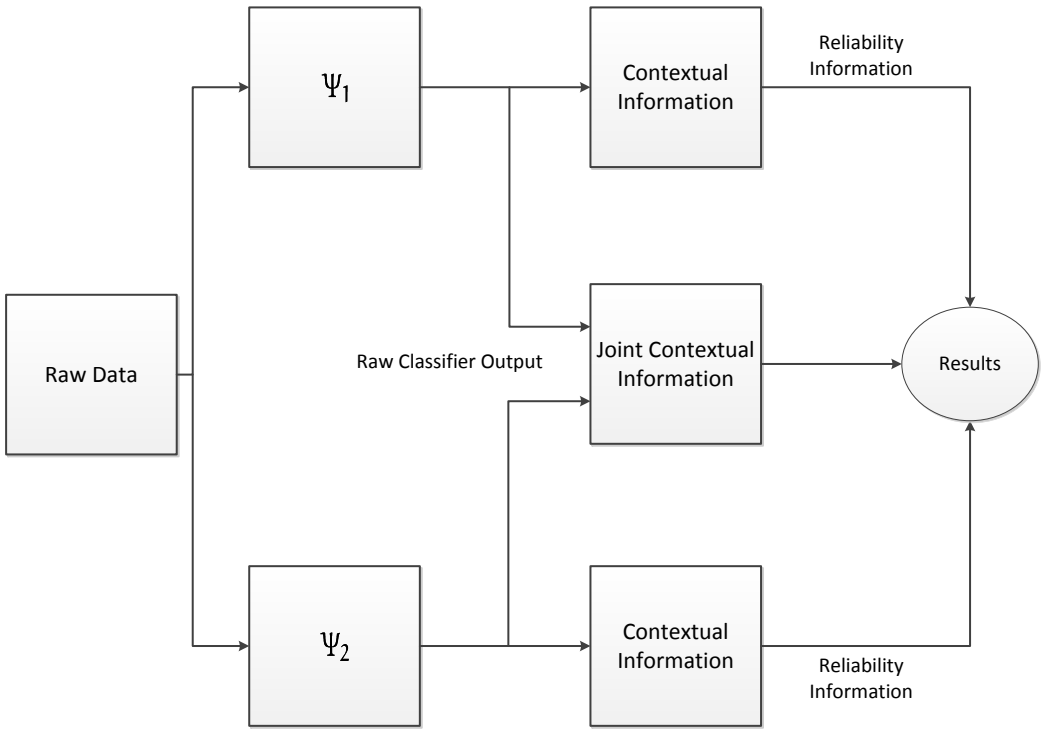


Figure 6.1: Training of two classifiers in a multiple classifier system

The first stage is the selection of the classifiers that are to be used in the combination scheme. Unfortunately this is a complex problem that demands further research. The second part is the extraction of contextual information from the individual classifiers or in other words, the determination of the ways in which the classifiers will express their opinions and some of the commonly used methods are detailed in [145]. Finally the third part is the combination of the information extracted from individual classifiers to reach a joint decision. The output of the classifier can be in the form of a label of the most likely outcome or class, ranking of the labels or posterior probability. The raw outputs of the individual classifiers may not be feasible for use in the combination schemes as determination of strengths and weakness of the individual classifiers is also necessary [146]. Focusing on the strengths and weakness of the classifiers can allow the building of better multiple classifier systems. As a result, contextual information including class dependent classifier reliability, universal classifier reliability and the conflicts among classifier should be extracted and from the classifier raw outputs. These three stages are discussed in some details in the following passages.

6.2.1 Description

Let K be the total number of classifiers in the multiple classifier system and N denote the total number of pattern classes. The classifiers in the system are denoted by Ψ_k, where $k = 1, 2, \ldots, K$. Assuming that a random variable X represents the pattern

classes which can take on the values $\mathbb{R} = \{1, 2, \ldots, N\}$ representing the labels of the pattern classes. A random variable D_k represents the labels of the decisions provided by the kth classifier, Ψ_k. A *joint decision* represents a class which is determined as the most likely class by the multiple classifier system.

6.2.2 Selection of the Classifiers

Selection of the classifier for use in the multiple classifier system is generally task dependent. Each selected classifier should have a reasonable classification performance for a particular classification task and as such these classifier should be a part of an "optimal" group of classifiers. Although, the selection of an optimal set of classifiers is a difficult problem, there are, however, some general rules that can be applied for selecting a suitable set of classifiers that have been designed previously for a specific task [42]. In order to select a suitable set of classifiers, the concepts of *complementariness* and *statistical independence* are used in the literature without suggesting a measure for the satisfaction of decisions generated by the classifiers.

Determining a useful complementariness measure to select the classifiers in the optimal selection is an open question. There are some ideas presented in the literature e.g. [144] suggested that the classifiers should not be strongly correlated in their misclassification i.e. they should not assign the same incorrect label to a test sample. It has also been stated in [147, 148] that an improvement in classification can be obtained by the multiple classifier systems if they are independent in the errors that they make. The statistical independence of the classifiers is based on assumptions that are made for theoretical purposes only but the validity of these assumptions are not well known in practice. It has also been suggested in [147] that the performance of a multiple classifier system is not totally dependent on the performance of the individual classifiers but also on the independence of the classifiers used in the combination. However, in some of the studies [149, 150] this fact is disputed where the independence of the classifiers is not taken into consideration and yet significant improvements in the performance have been reported.

There exists no established measure to quantify the complementariness of the information provided by individual classifiers in a multiple classifier system. However, simple intuitive approaches which avoid the use of classifiers that make similar errors can result in improved classification performance. The concept of complementariness is described in some detail in 6.2.5 which is used in the experiments presented in this chapter.

Selection of the classifiers is preceded by important question about how the information provided by the classifiers can be combined to reach an improved decision. To simplify matters, the output of the various classifiers can be divided into the following three categories:

- *Category 1* classifiers provide the least information about the pattern classes by

providing only a unique label associated with the most likely class to which the test sample belongs to.

- *Category 2* provide a ranking of the pattern classes by returning the labels of the most likely class, the second most likely class and similarly the least likely class.
- *Category 3* provide the likelihood or the probabilistic values of all the pattern classes. Classifiers falling into this category provide the highest amount of information about the pattern classes.

The combination schemes of the individual classifiers are named in accordance with the output information provided by the individual classifiers e.g. a category 1 combination scheme deals with the abstract level information generated by the individual classifiers. Whether the classifier belongs to category 1, category 2 or category 3, the output information is still regarded as raw output and some form of validation must be carried out through the training samples to extract reliable contextual information that would provide statistics about the strengths and weaknesses of the classifiers.

6.2.3 Contextual Information

A fundamental problem related to the extraction of reliable contextual information about the different classifiers used is the limited amount of data available for this purpose. In order to extract reliable contextual information, enough of the validation session should be used and each validation should contain a large section of the acoustic sound classes. The problem of obtaining reliable statistics can be solved by carrying out more than one validation session. To address this issue, the frames of the training session are divided into non-overlapping groups which are known as *tokens*. The tokens should be phonetically rich and contain enough training material to properly represent the different sound classes. As an example, a 20 s speech signal can be divided into 20 distinct tokens with the length of each token equalling 1 sec. If the speech signal is segmented into 10 msec frames, each token will contain 100 frames. In order to obtain tokens that are rich in phonetics, every other frame is assigned to a different token i.e. the first frame is assigned to the first token , the second frame to the second token etc., and the algorithm is repeated by putting the 21st frame in the 1st token and so on. This leads to a judicious allocation of phonetic content in all the tokens, representing different acoustical classes in the session.

The training tokens are used to validate the classifiers, resulting in the calculation of conditional probabilities which can be used for further mathematical formulations. The tokens associated with a speaker i are classified by a classifier Ψ_k, and the speakers in the top rank are counted. This information is used to fill out the ith row of a confusion

matrix, Υ_k, which is given as:

$$\Upsilon_k = \begin{bmatrix} \eta_{11}^{(k)} & \eta_{12}^{(k)} & \cdots & \eta_{1N}^{(k)} \\ \eta_{21}^{(k)} & \eta_{22}^{(k)} & \cdots & \eta_{2N}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{N1}^{(k)} & \eta_{N2}^{(k)} & \cdots & \eta_{NN}^{(k)} \end{bmatrix} \tag{6.1}$$

$\eta_{in}^{(k)}$ is the number of tokens belonging to speaker i that have been classified as speaker n by the classifier Ψ_k. If D_k is the decision of the classifier Ψ_k, the conditional probability that a token from speaker i is classified as speaker j, i.e. $P(D_k = j|X = i)$ can be computed from the values of the confusion matrix as [151]:

$$P(D_k = j|X = i) \cong \frac{\eta_{ij}^{(k)}}{\sum_{n=1}^{N} \eta_{in}^{(k)}} \tag{6.2}$$

The term in the denominator represents the total number of tokens used for the validation and is the same for all the speakers.

With the use of raw classifier outputs, information about joint or marginal classifier behaviours can be estimated. This can point to the strengths and weaknesses of the individual classifier and in turn can lead to the development of a better multiple classifier systems. Class dependent classifier reliability is also a form of contextual information i.e. the reliability of the classifier may depend upon the underlying class [152]. This has lead to the development of measures of the form *reliability*$(\Psi_k|D_k = j)$ are defined in literature to address the class dependent reliability. A more widely used reliability contextual information type is the *reliability*(Ψ_k) which is a numerical measure of reliability designated to a classifier e.g. by using a validation set the performance of a classifier can be tested and this value can be used to denote the global classifier reliability [151].

6.2.4 Classifier Combination Techniques

Based on the type output information provided by the individual classifier i.e. category 1, category 2 or category 3, some of the commonly used combination techniques used in the literature are described briefly in this below.

6.2.4.1 Category 1 Classifier Combination Techniques

For category 1 classifiers, majority and plurality voting are the two most commonly used combination techniques. As the name applies, in majority voting the class selected by more than half of the classifiers in the system is selected as the decision of the multiple

classifier system. An error is declared if no such pattern class exists [153]. A modified and relaxed version of majority voting is the plurality voting in which the final selected class is one which gets the most votes. If more than one class gets the most votes, the output is selected at random among them. Such combination techniques however, do not consider the contextual information provided by the individual classifiers.

6.2.4.2 Category 2 Classifier Combination Techniques

The three main types of combination techniques for the category 2 classifiers are the highest rank, Borda count and the logistic regression. For the highest rank techniques, each speaker is assigned a rank based score based on the testing of the input pattern. By convention the speaker that is ranked the highest receives the highest score. The combined score allocated to a speaker is the maximum of the scores given to that speaker by all the classifiers. The final decision is awarded to the speaker with the maximum score. In Borda count method, the score of a speaker is generated by determining the number of speakers ranked below that speaker. The sum of scores, assigned to that speaker by all the classifiers, represents the combined score of the system for that speaker. The speaker with the maximum score is selected as the joint decision [145]. Logistic regression [153], is a modified version of the Borda count technique, where the combined score of a speaker is the weighted linear combination of the individual scores and the weights reflect the relative significance of each classifier in the combination i.e. weights represent the contextual information provided by the individual classifiers.

6.2.4.3 Category 3 Classifier Combination Techniques

The most commonly used combination techniques for the category 3 classifier are the Bayesian probability theory [154, 155], and the consensus based combination techniques [146]. These techniques are described briefly in the following paragraphs.

Bayesian Formalism

Given K probabilistic classifiers with $P\left(X = i|D_k = j\right)$ representing the a posteriori probability that the correct pattern class is i when the output of the classifier Ψ_k is j where D_k represents the decision of the k^{th} classifier. Using Bayes' theorem and considering all the classifiers in the combination we have

$$P\left(X = i|D_1 = j_1, D_2 = j_2, \ldots, D_K = j_K\right) = \frac{P(D_1=j_1,\ldots,D_K=j_K|X=i)P(X=i)}{P(D_1=j_1,\ldots,D_K=j_K)} \quad (6.3)$$

Assuming conditional independence of the classifiers [156, 157]:

$$P\left(D_1 = j_1, \ldots, D_K = j_K|X = i\right) = \prod_{k=1}^{K} P\left(D_k = j_k|X = i\right) \quad (6.4)$$

From Equations 6.3 and 6.4,

$$P\left(X=i|D_1=j_1, D_2=j_2, \ldots, D_K=j_K\right) = \frac{P(X=i)\prod_{k=1}^{K} P(D_k=j_k|X=i)}{P(D_1=j_1,\ldots,D_K=j_K)} \quad (6.5)$$

The denominator of Equation 6.5 can be written in terms of the a priori probabilities as

$$P\left(D_1=j_1, \ldots, D_K=j_K|X=i\right) = \sum_{i=1}^{N}\left(\prod_{k=1}^{K} P\left(D_k=j_k|X=i\right)\right) P\left(X=i\right) \quad (6.6)$$

which finally leads to the a posteriori decision probability as

$$P\left(X=i|D_1=j_1, D_2=j_2, \ldots, D_K=j_K\right) = \frac{P(X=i)\prod_{k=1}^{K} P(D_k=j_k|X=i)}{\sum_{i=1}^{N}\left(\prod_{k=1}^{K} P(D_k=j_k|X=i)\right)P(X=i)} \quad (6.7)$$

The computation of Equation 6.7 is based on conditional independence of the individual classifiers, since otherwise huge amounts of data would be required to compute the joint statistics. The assumption of statistical independence is widely used in pattern recognition applications, although its validity remains largely unknown [158]. Bayesian formalism is also used along with a reject threshold θ for the combined probability, so that the class with the joint decision is accepted if and only if the combined probability is larger than θ [149].

Linear Opinion Pool

The linear opinion pool is one of the most frequently used combination techniques for category 3 classifier combination. The linear opinion pool is a linear weighted sum of the a posteriori probabilities [146]. For a given set of K classifiers, the general form of a consensus function among the individual classifiers is given by

$$\Phi\left(X=i\right) = \sum_{k=1}^{K} \omega_k P\left(X=i|D_k=j_k\right) \quad (6.8)$$

The relative magnitude of the weights ω_k determine the contribution of each individual classifier in the joint decision. Such a combination technique has been frequently studied and applied in literature to pattern recognition applications [159, 160].

Logarithm Opinion Pool

The logarithm opinion pool is derived using the Bayes' rule on the conditional independence of the individual classifiers. The consensus function in this case is given

as

$$\Phi\left(X=i\right)=\prod_{k=1}^{K}P\left(X=i|D_k=j_k\right) \tag{6.9}$$

As the interest lies in finding the class which would maximize the consensus function Φ, Equation 6.9 can be modified with monotonic logarithm function without altering the decision reached through the Bayesian formalism as

$$\Phi\left(X=i\right)=\sum_{k=1}^{K}\omega_k \log P\left(X=i|D_k=j_k\right) \tag{6.10}$$

The above equation represent the sum of the logarithms of the a posteriori probabilities where ω_k represents the weights of the classifiers which reflect the relative significance of the information provided by the individual classifiers.

The classifier combination techniques described so far are applicable to a set of already existing classifiers. The recognition performance of the multiple classifier systems depends upon the joint performance of the classifiers. In order to improve the classification performance of the multiple classifier systems, the concepts of complementariness should be discussed in the design of such systems.

6.2.5 Complementariness

The main aim of using an additional classifier in combination with an existing classifier Ψ_1, is to obtain a classification error which is much smaller in magnitude to the classification error $p_{err}\left(\Psi_1\right)$ of Ψ_1. If $p_{err}\left(\Psi_1,\Psi_2\right)$ represents the error probability of a multiple classifier system with Ψ_1 and Ψ_2 as the classifiers, then the simplest form of complementariness measure would be of the form

$$\mathrm{cmp}\left(\Psi_1,\Psi_2\right)=p_{err}\left(\Psi_1\right)-p_{err}\left(\Psi_1,\Psi_2\right)$$

or

$$\mathrm{cmp}\left(\Psi_1,\Psi_2\right)=p_{correct}\left(\Psi_1,\Psi_2\right)-p_{correct}\left(\Psi_1\right) \tag{6.11}$$

where $p_{correct}\left(\Psi_1\right)=1-p_{err}\left(\Psi_1\right)$ represent the correct classification probability of Ψ_1. In a multiple classifier system, among K other classifiers, the best choice for accompanying Ψ_1 would be a classifier that would maximize the above equation.

For a plurality voting scheme, a symmetric complimentary measure proposed by [161] is given as

$$\mathrm{cmp}\left(\Psi_1,\Psi_2\right)=\sum_{i=1}^{N}\max\left\{P\left(D_1=i|X=i\right),P\left(D_2=i|X=i\right)\right\}$$

which states that for two classifier Ψ_1 and Ψ_2 to provide complimentary information to each other, at least one of the classifiers should provide correct information about the pattern class being tested. This measure is used in the experiments that are conducted to measure the performance of a speaker identification system based on multiple classifiers against the converted synthetic voices. The details of the experiments are described in the next section.

6.3 Combining Classifiers for Speaker Identification against Voice Conversion

For speaker identification, different types of features can be used along with various different types of classifiers. In section 2.4 different feature sets were presented, which highlight the speaker properties from different perspectives and in the same fashion different classifiers can postulate different models for the speakers. As a result of using different classification strategies, the speakers which are misclassified may not essentially overlap. In such a case it is reasonable to use multiple classifiers at the same time instead of a single classifier to avoid the miss classifications of a particular classifier. The assumptions that the classification shortcomings of different classifiers do not overlap, is the foundation of using multiple classifier systems for pattern recognition problems such as speaker identification. This approach aims to take advantage of the strengths of individual classifiers while avoiding their weaknesses to improve the recognition performance.

In Section 2.4 various features that are commonly used in speech processing systems were introduced. Each feature vector representation addresses the properties of the speech signal through a different viewpoint: e.g. LPC based feature vectors are derived from the solution of an all-pole model fit 2.23, emphasizing the formants of the speech signal and the cepstral features [6], which are obtained by the application of the logarithm to the magnitude of the Fourier transform of the speech signal. The cepstral features have been widely used in literature and have shown to outperform other representations in speaker recognition applications [162]. The selection of features is task dependent and is determined by their stability, linearisation and interpolation properties.

For voice conversion algorithms LPC, LSF, LPCC and MFCC features have been selected in literature for conversion and synthesis [103, 98, 114]. LSF are the feature vectors of choice for altering the characteristics of human voice by means of voice conversion algorithm because of their ease of computation, good interpolation properties and good inter-frame and intra-frame correlation values. In Chapter 4 the task of voice conversion was carried out with success by the use of LSF as feature vectors representing the spectral properties of the source and target speakers. The success of the converted voices against the speaker identification system was demonstrated in Chapter 5 where

the performance of the speaker identification system was tested for the identity disguise and voice conversion scenarios. Unlike recognition tasks where MFCC are the preferred acoustic features, for speech synthesis it is impossible to recover the original spectrum from the MFCC representation as the filter-banks operate on a non-invertible integration of the spectral samples.

The performance of the speaker identification systems has been shown to improve by the use of multiple classifier system [163, 144, 164]. However, to date there are very few studies which determine the performance of multiple classifiers based speaker identification systems against converted voices. e.g. the performance of the speaker identification system was improved on telephone speech in [165, 161] by fusing the outputs of two classifiers where one of the classifiers employed channel compensation and the other did not. It was reported that the performance of the speaker identification system is very sensitive to the signal processing done in the extraction of the feature vectors from the speech signal [161].

The modifications in the speech signal of a source speaker, who may want to hide their identity or target a particular speaker, can be viewed as a deliberate degradation of the speech signal of source speaker. These degradations, unlike the alterations caused by the channel or mismatch conditions, cannot be quantified since the statistics of the impostor (source) speaker may or may not be known to the system. Also, these modifications are heavily dependent on the source target pair, as has been highlighted in previous work by researchers [106] as well as from the results presented in Chapter 5, and determining the spectral characteristics of the source speaker in the case when they have been masked is an extremely difficult task. Statistics related to this so called noise, whose characteristics are dependent on the speaker pair involved, cannot be obtained by the present noise estimation techniques. In such a scenario, nullifying an intrusion into the speaker identification system becomes a difficult endeavour.

Finding a deterministic model for such degradations, which can subsequently be used in speaker identifications tasks, is hypothetical at best. There have been some speaker identification studies on deliberately modified voices e.g. in [166] the authors have conducted experiments on speech signals obtained with altered pitch values i.e. raised and lowered, speech generated by placing hand on the mouth and whispering. These studies focus only on the prosodic modifications and not on the intentional modifications to the vocal tract characteristics by the voice conversion techniques. Also, accurate detection of the pitch is a challenging task and dependence on the pitch values can allow the impostors to gain access to the speaker identification systems by changing their own pitch values. Computation of the pitch values for specific speech sounds, such as nasals and consonants, is a difficult task. As such the different front-end processors of speaker identification system do not seek to use pitch as a speaker specific feature but try to find the speaker specific properties in other parts of the speech signals. Furthermore, the issue of discriminating synthetic voices from the converted ones is important, and is a research topic that is still in its infancy. Although it is believed that the voice

conversion techniques will not be able to convert all the features of the source speaker, there is a lack of research in determining the effect of these transformation and their differences from the synthetic voices.

It was mentioned before that the different feature sets highlight the properties of the speech spectrum from different angles, where some feature sets are more suited for the task of voice conversion while others perform better in speaker identification systems. The performance of speaker identification systems using various feature vectors has been well studied in the speaker recognition community, where cepstral features clearly outperform LPC and its variants. For speaker identification systems, the amount of information a particular feature set or a particular classifier is able to extract cannot be quantified, and the performance is dependent on the dynamics of the enrolled population. In view of the above, it is plausible to use a multiple classifier system for the task of speaker identification utilizing different features.

In this section we propose the use of multiple classifier systems employing different features to analyse the performance of s speaker identification system against deliberately altered voices. The details of the experimental set up are described below.

6.3.1 Features and Speaker Modelling

A speaker identification system consists of the feature extraction stage followed by generating the speaker models and the classification engine. A block diagram of the system used in the simulations is shown in Figure 6.2 . The front-end of the speaker identification system is the feature generation stage.

Current state-of-the-art speaker identification system employ MFCC as the feature vectors and GMM for modelling the speakers and classification engine. Ideally the front-end of the speaker identification system, should be able to extract all speaker specific information from the input speech of the enrolled speakers, without focusing on the issue of what is being said. It is important to point out that MFCC feature vectors are employed in both speaker identification and speech recognition applications. For a speaker independent speech recognition task any speaker specific information is considered as noise by the speech recognition system, but this speaker specific information is exactly the sort of information required by the speaker identification system. The use of MFCC as feature vectors in the two systems seeking different kind of information suggests that the MFCC contain both speaker level information and the linguistic information. The extraction of the MFCC from the speech signal was described in detail in Section 2.4.4.3.

LPCCs have also seen used in speaker and speech recognition applications. The LPCC parameters are derived from the corresponding LPC of the speech signal using simple recursive equations. The process of LPCC computation was described in detail in Section 2.4.4.2. With the emergence of better computation performance MFCC has replaced LPCC as the front-end of most of the speech and speaker recognition systems.

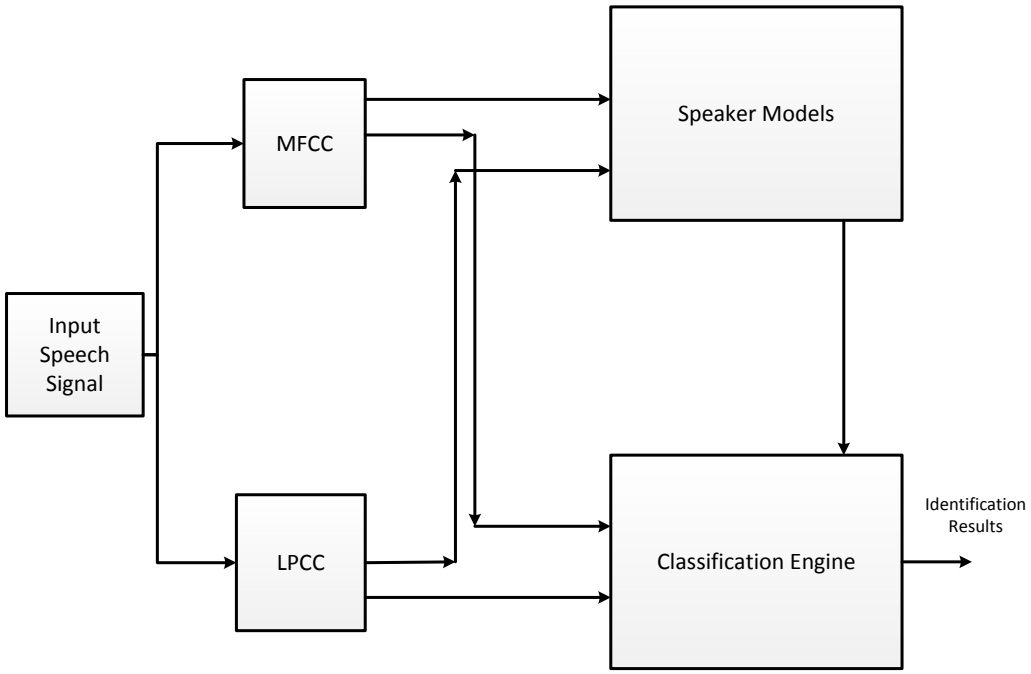


Figure 6.2: Speaker identification system using MFCC and LPCC in the feature extraction stage

In a voice conversion system on the other hand (see Chapter 4), the LSF are used in generating speaker models, prediction of the target speaker characteristics and in the synthesis of the converted speech signals. It was mentioned before that the success of a voice conversion system depends not only upon the source-target pair but also on the properties of the feature vectors used. And as such the voice conversion system will struggle to transform the source speaker's characteristics to match those of the target speaker properties in a scenario where the properties of the source and the target pair are not conformable for conversion. The use of LPC as feature vectors for speaker identification applications have been proposed in literature but compared to cepstral features they possess poor performance. The LPCC feature vectors are derived from the LPC feature vectors and as such they possess the same disadvantages of the LPC representation. The LPCC feature vectors have been chosen to accompany the MFCC in the proposed multiple classifier system because unlike the MFCC, LPCC only represent the speaker information present in the spectral envelope of the speech spectrum by removing the pitch information after the application of a low-pass lifter. This allows for a judicious choice of feature vectors targeting the characteristics of the speech spectrum from different angles.

The transitional feature vectors are also found to be useful in speaker identification task apart from the instantaneous spectral feature vectors such as MFCC and LPCC [30]. For a feature vector representing the instantaneous spectral information i.e. MFCC and

Label	Feature Vectors	Speaker Model
Ψ_1	$\{19 - MFCC, 19\Delta MFCC\}$	32-GMM (Nodal Covariance)
Ψ_2	$\{14 - LPCC, 14\Delta LPCC\}$	32-GMM (Nodal Covariance)

Table 6.1: Summary of the classifiers used in the system, feature vectors and speaker models

LPCC the transitional or the dynamic feature vectors are calculated as the difference of two successive frames.

Selection of the features for use in the speaker identification system is followed by the generation of speaker specific models, utilizing features. Different types of modeling methods have been proposed in literature with some of the most commonly used techniques described in Chapter 3. For the purpose of building the multiple classifier based speaker identification system Gaussian Mixture Models or GMM were used. GMM aims to model each acoustical speech sound with a different uni-model Gaussian or a Gaussian component. Given a sequence of feature vectors $X = \{x_1, x_2, \ldots, x_T\}$ with a total of T frames, which are assumed to be independent, the log-likelihood of a speaker model λ_s is computed using

$$\hat{S} = \sum_{t=1}^{T} \log p\left(\mathbf{x}_t | \lambda_s\right) \tag{6.12}$$

The value of $\hat{S}$ is computed for all the speaker models λ_s enrolled in the system and the speaker model that generates the highest value is returned as the identified speaker. The speaker modelling using GMM was described in details in Section 3.5.1. In the experiments presented in this section, the various speaker models were trained using 32-component GMM with nodal covariance matrices [58].

These features and the speaker modeling technique is widely used in literature for speaker identification tasks [43, 8, 58]. In these simulations, 19th order MFCC vectors are obtained using 24 mel-scale filter-banks and similarly 14th order LPCC vectors were obtained from 16th order LPC are extracted from a 20 msec speech frame with a 10 msec overlap. These feature vectors were appended with their corresponding delta feature vectors, giving 38th order MFCC and 28th order LPCC feature vectors. Each feature vector stream was used to train a 32-component speaker specific GMM with diagonal nodal covariances. A description of the feature vectors and the speaker models used in these experiments is summarized in Table 6.1.

6.3.2 Simulation Set up

To test the performance of the individual classifiers, speech material from the Dialect Region 1 (DR1) of the NTIMIT [68] corpus was used. NTIMIT was collected by transmitting all the TIMIT (3.8.1.1) recordings through a telephone handset and over various channels. The NTIMIT waveforms are aligned with the TIMIT waveforms so

System	Speaker Identification Performance (%)
Reynolds [167]	60.7
Mashao and Baloyi [168]	69.2
Lerato [169]	71.1

Table 6.2: Identification performance on NTIMIT database in literature

Type of Features	Speaker Identification Performance (%)
MFCC	71.3
LPCC	65.7

Table 6.3: Identification Performance of baseline classifiers Ψ_1 and Ψ_2

that the TIMIT transcriptions can be used with the NTIMIT corpus. The DR1 region of the NTIMIT (or the TIMIT corpus) contains 47 speakers (16 females and 31 males). The sentence structure for each of the speakers in the NTIMIT corpus is the same as TIMIT i.e. each speaker utters 10 sentences each. Two sentences with the prefix sa (sa1 and sa2), these two sentences although different, are common to all the speakers in the database. There are three si sentences and five sx sentences. These si and sx sentences are different from each other and different across speakers. All the data has been recorded at 16KHz at a resolution of 16-bits.

For the utterances in the NTIMIT database, first eight sentences including the sa1 and sa2 sentences are used for model training and the last two sentences are used for the testing of the speaker identification system. The same configuration is used in the simulations with the classifiers Ψ_1 and Ψ_2. The performance of the speaker identification system using the NTIMIT database has been widely reported in the literature and Table 6.2 gives the identification performance found in literature for the NTIMIT database. Table 6.3 details the identification performance of the baseline classifiers Ψ_1 and Ψ_2.

From Tables 6.2 and 6.3, it can be seen that the identification performance of the classifier Ψ_1 using MFCC as the feature vector is slightly better than the values reported in literature. This is because the classifiers Ψ_1 and Ψ_2 are operating on a reduced set of the NTIMIT database i.e. DR4. The Cepstral Mean Normalization (CMN) is not used in the experiments presented here. Although CMN approach has shown to provide improvement in the identification performance when used on telephone speech, the identification performance has been shown to decrease when used with clean speech [170]. This would suggest that the CMN process removes some of the speaker specific information from the extracted features. Since the aim of the experiments presented in this chapter is to analyse the performance of speaker identification system when presented with converted synthetic voices, removal of speaker specific information would not be beneficial.

Table 6.3 shows that the baseline systems provide reasonable identification performance when dealing with noisy speech signals. To test the performance of the systems against

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M2	0/10	1/30	1/50
F1	0/10	0/30	0/50
F2	0/10	0/30	0/50

(a) Impostor M1

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M1	0/10	1/30	0/50
F1	0/10	0/30	0/50
F2	1/10	0/30	0/50

(b) Impostor M2

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M1	2/10	2/30	0/50
M2	0/10	0/30	0/50
F2	0/10	0/30	0/50

(c) Impostor F1

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M1	1/10	0/30	0/50
M2	0/10	1/30	0/50
F1	3/10	1/30	3/50

(d) Impostor F2

Table 6.4: Results of the Identity Disguise Experiments on Ψ_2

converted synthetic voices, the SID set and the VC set of the VOICES speech corpus, which were used for identity disguise and targeted voice conversion and introduced in Chapter 5, are used along with Ψ_1 and Ψ_2. A brief description of the experimental apparatus using the two sets is described in the following passages.

Identity Disguise

To test the performance of the classifiers Ψ_1 and Ψ_2 against identity disguise using voice conversion techniques, the simulation set up described in Section 5.3.2.1 is used. The results for the identity disguise using the MFCC based classifier Ψ_1 are listed in Table 5.3. Table 6.4, however, lists the source identification performance of classifier Ψ_2 with LPCC as the feature vector, on the identity disguised test.

From the comparison of Tables 5.3 and 6.4, it can be seen that the identification of the source speaker, when he/she is deliberately trying to deceive the speaker identification system, decreases in the cases of Ψ_2 using LPCC as the feature vectors. LPCC have been shown to have inferior performance compared to MFCC since they are derived from the LPC and as such inherit the same problems. The main reason for using the LPCC as the feature vectors for Ψ_2 is that the voice conversion system uses LSF as features representing the vocal tract characteristics. Selection of LPCC for Ψ_2 allows

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	4		
M2		-		
F1	6	4	-	
F2		2	6	-

(a) Source Identification (%) of Ψ_2 with the 50 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	92	98	100
M2	98	-	100	100
F1	94	96	-	100
F2	96	98	88	-

(b) Target Identification (%) of Ψ_2 with the 50 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	4		2
M2	2	-		
F1			-	2
F2	4		6	-

(c) Other Identification (%) of Ψ_2 with the 50 Converted Sentences on the SID Set

Table 6.5: (%) Identification of the *Source*, *Target* and *Other* identifications using 50 converted sentences

to determine the effectiveness of the voice conversion system when it modifies the vocal tract characteristics only. The decrease in the performance of Ψ_2 compared to Ψ_1 indicate the fact the converting speech sounds among speaker using features that describe the same information can lead to reduced identification performance. This test would indicate that the use of LPCC as feature vector does not necessarily increase the performance of the speaker identification system. However, for any given source-target speaker pair, the source identification rates have not decreased by a huge margin. The performance of the voice conversion system is dependent on the source-target pair and the system is limited in its ability to overcome these dependencies.

Voice Impersonation

The voice impersonation experiments were performed on the VC set of Chapter 5. The same experimental set up was used in the voice impersonation testing. Each of the four selected speakers M1, M2, F1 and F2 were used as both source and target speakers, giving a total of 12 source-target speaker pairs. Also, unlike the identity disguise experiments, the target speakers were enrolled in the speaker identification system. The objective of this test to determine the performance of the voice conversion system when dealing with converted voices specifically targeting a speaker who is a part of the speaker identification system. The identification performance of Ψ_2 using LPCC as feature vectors, with 50 sentences used in the training of the transformation function are detailed in Table 6.5.

The performance of Ψ_2 for the VC set and SID set indicate the source identification performance loss, suggesting an improvement in the success rate of the impersonated voices. In the following section the performance of the two classifiers Ψ_1 and Ψ_2 in a combination scenarios is analysed.

6.3.3 Classifier Combination

A lot of empirical evidence exists which reveals that the use of multiple classifiers can improve the recognition performance in many pattern recognition tasks [163, 144, 165, 164]. e.g. Doddington *et al.* [163] showed that the use of simple combination of scores obtained from different classifiers, improved the performance of the baseline recognition system on the NIST 1998 Speaker Recognition evaluations. Chen and Chi [171] combined multiple probabilistic classifiers using different feature sets obtained from the same speech data for the task of speaker identification. They demonstrated that the robustness of the speaker identification system can be improved by a combination of different classifiers using features representing different spectral characteristics. Ramachandran *et al.* [172], gave a description of the different forms of redundancy, diversity and fusion that can be employed to improve of the performance of speaker recognition system. In their experiments, they reported an improvement in the performance of the speaker verification system using different classifiers trained from the same set of features extracted from the front-end.

As was mentioned before, the multiple classifier systems can generally be divided into three main categories depending upon their structure, the types of outputs produced by the individual classifiers and the different types of combination techniques used for obtaining the final decision. The classifiers in combination can be either serial or parallel or hybrid i.e. containing both parallel and serial architectures. In parallel combination techniques, each of the classifiers in combination are activated at the same time and the fusion output is obtained using a single combination function. On the other hand, in a serial combination the output of one classifier reduces the set of pattern classes to the next classifier in combination [173]. The outputs of the individual classifiers in a multiple classifier system is generally divided into three categories abstract, rank and measurement level [151]. A description of these levels was listed in Section 6.2.4. The methods used to combine these different output levels are are generally classified either fixed or trained rules. As the name suggest, the fixed rules are stationary in the sense that the states and parameters do not change as a consequence of the change in the output of the individual classifiers. These combination techniques are well suited for the group of classifiers which make uncorrelated errors and exhibit similar performances. The trained rules, however, adapt their parameters and form in accordance with the alterations in the outputs of the constituent classifiers of the multiple classifier systems. The trained rule classifiers are more more suited to the classifiers which produce different types of outputs and make correlated errors on the same test material [173].

In the context of multiple classifier systems, there are however, very few studies that provide a sound theoretical basis for understanding the improvements obtained in the multiple classifier systems. One such study was performed by Kittler et al. [144], which provided a theoretical framework for combining classifiers, using different feature sets, to obtain an estimate of posterior probabilities for the given patterns. They presented a number of rules based on Bayesian theory under the assumption of conditional independence and the difference between the estimated posterior probabilities and the prior probabilities is negligible.

Assuming that X_1 refers to the feature vectors related to Ψ_1 using MFCC as the feature vectors obtained from the front-end processors and X_2 are the feature vectors corresponding to the LPCC based classifier Ψ_1, we applied the rules defined by [144] to our problem of speaker identification against computer-aided voice conversion as follows:

Sum Rule

$$\hat{S}_{sum} = \arg\ \max_{n=1}^{N} \left[\sum_{i=1}^{2} S_n\left(X_i\right) \right] \tag{6.13}$$

Product Rule

$$\hat{S}_{prod} = \arg\max_{n=1}^{N} \left[\prod_{i=1}^{2} S_n\left(X_i\right) \right] \tag{6.14}$$

Maximum Rule

$$\hat{S}_{max} = \arg\ \max_{n=1}^{N} \left[\max_{i=1}^{2} |S_n\left(X_i\right)| \right] \tag{6.15}$$

and finally the Minimum Rule

$$\hat{S}_{min} = \arg\ \max_{n=1}^{N} \left[\min_{i=1}^{2} |S_n\left(X_i\right)| \right] \tag{6.16}$$

where N is the total number of speakers enrolled in the SID system. The four sets of rules are used in these experiments to determine the identification performance of the multiple classifier system using the classifiers Ψ_1 and Ψ_2.

6.3.4 Results

The main argument in these simulations is that the differences in the signal processing used for the extraction of the MFCC and LPCC can lead to the extraction of different spectral information from the same speech sample. Consequently these differences can

Combination Rule	Identification Performance (%)
Sum	78.0
Product	78.0
Maximum	71.3
Minimum	71.7

Table 6.6: Sum, Product, Maximum and Minimum Rule Combinations on the NTIMIT Corpus

cause the base classifiers Ψ_1 and Ψ_2 to misclassify the different speakers in a speaker identification system when presented with converted voices.

Before proceeding to the use of classifiers Ψ_1 and Ψ_2 in a multiple classifier system, the combination rules of Equation 6.13, 6.14, 6.15 and 6.16 are validated on the NTIMIT corpus. Table 6.6 lists the identification performance of the system for the NTIMIT corpus.

For Table 6.6, the sum and the product rule outperform the maximum and the minimum rule with the maximum rule providing the least performance. The sum rule outperforms the others since it is more robust to the estimation errors [144]. The identification results obtained so far, using the combination rules of Equations 6.13, 6.14, 6.15 and 6.16, assume that each class is equally likely. However using a linear weighted combination, a further improvement of 80.3% in the identification performance has been obtained. The weights were estimated by using a simple search for the best weights. The weight $\alpha_1 = 0.70$ for classifier Ψ_1 and $\alpha_2 = 0.30$ for Ψ_2 were used. The values of the two weights suggests that the classifier Ψ_1 using MFCC as the feature vectors is more reliable than classifier Ψ_2 using LPCC as the feature vectors.

Given the identification performance of the linear weighted combination it can be concluded that the use of classifiers in combination can improve the performance of the speaker identification systems in the case of noisy speech even though the two classifiers used in these experiments extract different types of features, which represent different characteristics of the speech spectrum.

After establishing that the performance of the speaker identification system can be improved by the use of a multiple classifier system, the same performance measure i.e. weights are used in the scenario where the speaker identification system is presented with converted voices. Although the weights are estimated for the NTIMIT corpus, the argument from the previous sections, that a converted speech signal can be viewed as a noisy speech signal where the characteristics of the noise depends on the dynamics of the source-target pair, holds true. The determination of the optimum weights for classifier combination for the converted voices is a difficult task where the characteristics of the source and the target speaker may or may not be known to the speaker identification system.

The identification performance of the multiple classifier system against the converted voices in the identity disguise and target voice impersonation scenarios is described in

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M2	10	6.67	4
F1	0	0	0
F2	10	0	0

(a) Impostor M1

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M1	10	3.34	2
F1	0	0	0
F2	10	6.67	0

(b) Impostor M2

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M1	50	13.34	0
M2	0	0	2
F2	0	0	0

(c) Impostor F1

Target Speakers	10 Sentences	30 Sentences	50 Sentences
M1	10	3.34	0
M2	0	3.34	2
F1	40	6.67	8

(d) Impostor F2

Table 6.7: Results of the Identity Disguise Experiments

Tables 6.7 and 6.8.

A comparison of Tables 6.7 and 6.8 reveals an improvement in the performance of the classifier system against the converted voices. For the identity disguise scenario, where the speaker is deliberately disguising his/her own voice, the percentage identification of the source speaker has increased in all the cases. Also for targeted voice impersonation experiments, the identification of the target speaker has decreased with a corresponding increase in the source identification with relative decrease in the other identification rates. The combination of classifiers using features representing the different spectral characteristics can lead to an improvement in the performance of the system against converted voices. These identification rates, however, are still high due to the limited nature of the speaker data set used which results in the lesser number of competing models in the decision making process resulting in the higher number of successful intrusions.

The voice conversion system and the speaker identification systems explicitly use the feature vectors representing the vocal tract characteristics. From the identification performances listed in this section against converted voices, it is clear that in order to reduce the impersonation success rates and increase the identification performance against converted voices, more venue for feature extraction must be explored. In this regard, the use of speaker specific information in the LP-residual of the speech signal

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	12	4	10
M2	10	-	8	16
F1	10	14	-	4
F2	14	2	6	-

(a) Source Identification (%) with the 50 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	86	96	86
M2	86	-	92	82
F1	88	70	-	86
F2	82	98	88	-

(b) Target Identification (%) with the 50 Converted Sentences on the SID Set

	Identified Speakers			
Source Speakers	M1	M2	F1	F2
M1	-	2		4
M2	4	-		2
F1	2	16	-	10
F2	4		6	-

(c) Other Identification with the 50 Converted Sentences on the SID Set

Table 6.8: (%) Identification of the *Source*, *Target* and *Other* identifications using 50 converted sentences

is explored, in the context of multiple classifier systems.

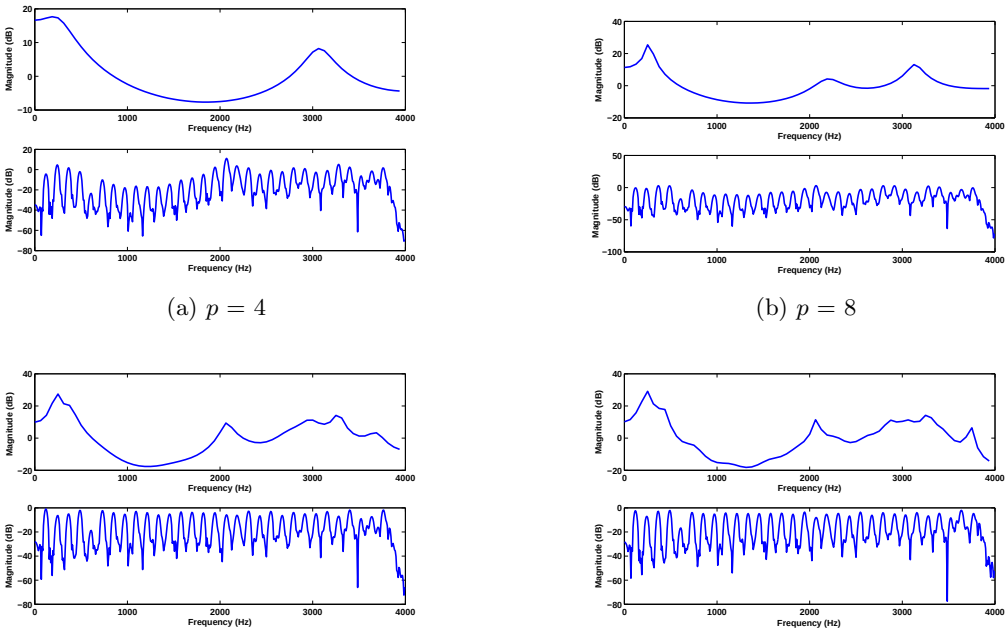
6.4 Speaker Specific Information in the LP-Residual

Following the source-filter model (Section 2.2.1), various researchers have attempted to derive the features from the LP-residual that contain speaker specific information such as glottal information [174]. The potential of auto-associative neural network was explored using the sub-segmental and segmental features extracted from the linear predictive analysis [175]. The authors presented promising results on the use of these features in the speaker identification applications. The state-of-the-art speaker identification systems, however, prefer the use of features representing the vocal tract spectral characteristics only. Features such as MFCC and LPCC have been used extensively for speaker modeling using GMM.

Under the framework of the source-filter theory, the vocal tract is associated with the filter and the excitation with the residual in the context of linear prediction. The linear prediction analysis estimates the LPC by minimization of the prediction error. i.e. the predicted samples from a linear combination of the past p samples is given by [1]

$$\hat{s}(n) = -\sum_{k=1}^{p} \alpha_k s(n-k) \tag{6.17}$$

The LPC coefficients α_k are related to the vocal tract characteristics and may also contains speaker-dependent information. The LP-residual is obtained as a difference



(a) $p = 4$ (b) $p = 8$

(c) $p = 20$ (d) $p = 30$

Figure 6.3: LP Spectrum and the Spectral Envelope for different values of the predictor variable p

between the current and the predicted samples i.e.

$$e(n) = s(n) - \hat{s}(n) \tag{6.18}$$

The predictor order p plays an vital role in speech processing systems. As the value of p increases, the LP spectrum provides a better estimation of the speech spectrum. The envelope of spectrum estimates the frequency response of the vocal tract filter. Typically, in the 0-4 kHz band, the vocal tract filter contains a maximum of five resonances. Hence a value of p in the range of 8-14 is normally used for speech signal sampled at 8 kHz [1]. Figure 6.3 shows different spectral envelopes extracted from a segment of voiced speech for different values of the predictor variable p.

For lower values of p e.g. 4, the LP spectrum may focus on the significant peaks of the spectra only. In such a case the LP residual will still contain significant information about the vocal tract filter. in Figure 6.3a, the spectrum of the residual signal contain significant information about the spectral envelope. If a large value of p e.g. 30 is used, the spectral envelope will contain many spurious peaks, containing information about the harmonic structure of the speech spectrum, which may not reflect the true spectral envelope (Figure 6.3d). These peaks can affect the residual signal when is processed with the corresponding inverse filter.

With a proper use of the predictor variable p, the residual will mostly contain the excitation information only. Among the different kinds of excitation sources, the voiced segment may contain significant speaker-specific information, as the glottal vibrations

may vary for different speakers [176]. The speaker specific characteristics can be attributed to the difference in the rate of global variations, strength of the excitation and the shape of the glottal pulse. The excitation strengths is dependent on the rate of glottal closure. This is indicated by a large residual error around the instant of heavy excitation in each of the pitch periods [177].

Several studies have been carried out to use the LP-residual for the betterment of the SID systems [175, 178]. In [179] it is proposed to exploit the orthogonality between the vocal tract filter and the residual. The results suggest the complementary nature of these representations for speaker verification tasks. The use of NN have also been proposed for the characterization of the LP residual [180]. In [105] Auto-Associative Neural Networks (AANN) are used for modelling the speaker specific information present in the LP-residual. The authors conclude that SID systems can attain adequate rates by the use of residual features alone. The nature of the residual signal should also be taken into consideration when designing effective and efficient systems. For the original speech signal, many investigations have been conducted [181, 182]. During speech production, the physiological behaviour can cause turbulence in the output speech signal. This can result in the presence of non-linearities in the speech signal. The non-linearity of the speech signal can determined by statistical methods such as higher-order statistics [181, 183]. Due to the lack of efficiency in the residual estimation process, e.g. due to the analysis order, presence of noise, short-comings of the algorithm etc., it can be suggested that the residual can be modelled by second-order statistics as well as higher order statistics. Non-linear modelling has been proposed as a possible solutions in different applications [105, 184, 185] due to the non-linearity of the residual. The results show the potential and confirm the presence of non-linearity. Thyssen et al. [185] suggested the presence of non-linearity in the residual after performing multiple linear prediction analysis to remove all linear information from the residual. This approach, however, demands caution since adaptive methods can result in approximately Gaussian residual signals [181]. In this thesis, it is proposed to explore the fact that the residual signal conveys all information that are not accounted for by the LPC filter. The proposed representation of the LP-residual is based on the spectral models. These investigations aim to show the potential of residual speech signal processing for speaker identification task against converted synthetic voices. The features extracted from the residual can provide complementary information along with the LPCC or even the MFCC.

6.4.1 Representation of the LP-residual

In this section, the processing of the residual signal, based on the residual spectrum is described. The approach was first presented by [186] and was termed as the Power Difference of Spectra in Sub-bands (PDSS). The representation in their work was used in a speaker identification problem. In our simulations the R-PDSS features provides an identification rate of 66.9% and in combination with LPCC features the identification

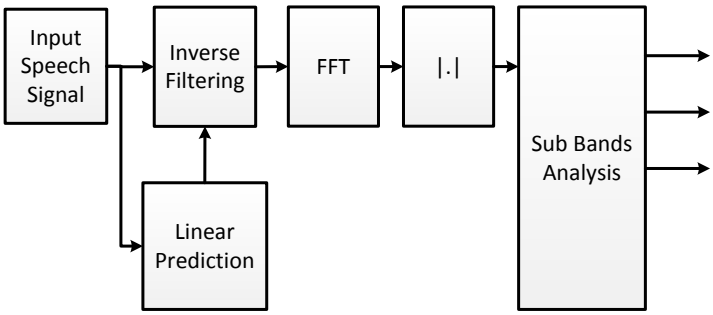


Figure 6.4: PDSS feature Extraction Process

rate jumps to 99%. The processing steps involved in the extraction of R-PDSS features are described in the following steps.

- Estimate the LP-residual
- FFT by zero padding to increase the frequency resolution
- Divide the spectrum into sub-bands
- Calculate the ratio of the geometric mean to arithmetic mean for each sub-band and subtract from 1

$$\text{R-PDSS}(j) = 1 - \frac{\left(\prod_{k=L_j}^{U_j} S(k)\right)^{1/N_i}}{\frac{1}{N}\sum_{k=L_j}^{U_j} S(k)} \tag{6.19}$$

where $N_i = U_j - L_j + 1$ is the number of sub bands used with L_i and U_i representing the lower and upper frequency limits in the jth sub band. Figure 6.4 shows a graphical description of the feature extraction process. The speaker specific models were trained using GMM with 32-components and diagonal covariance matrices.

6.4.2 Score Fusion

In this section, the aim is to evaluate and to compare the performance of the features in a speaker identification problem when dealing with converted voices. We combine the output of the classifiers Ψ_1, Ψ_2 and Ψ_3 using the features MFCC, LPCC and R-PDSS respectively. The output of the classifiers in this case is combined using the opinion fusion after [144], using Equation 6.13. The results of the data fusion using

Classifiers	Combined Identification Performance (%) against Identity Disguise
$\Psi_1 + \Psi_3$	40.7
$\Psi_2 + \Psi_3$	29.3
$\Psi_1 + \Psi_2$	21.0

Table 6.9: Source Identification Performance against identity disguise using spectral envelope and LP-residual features

combinations of classifiers Ψ_1 Ψ_3 and Ψ_2 and Ψ_3 for the identity disguise scenario using 50 sentences for the training of the transformation function are described in Table 6.9

From Table 6.9, it can be seen that the use of spectral envelope features along with features representing the properties of the LP-residual, perform much better in detecting the source or the impostor speaker than the combination of classifiers using only the spectral envelope features, namely MFCC and LPCC. The performance obtained with MFCC + R-PDSS based features i.e. 40.7% clearly outperforms the other two combinations indicating that the LP-residual indeed contains some speaker specific properties which should be utilized in a speaker identification system to improve the identification performance.

6.5 Summary

In this chapter the use of multiple classifier systems was presented for the speaker identification task. The concept of multiple classifier systems in the context of speaker identification system was explained, for the different types of classifiers, extraction of contextual information and the concepts of complementariness were presented. Different combination schemes depending on the type of the classifiers were also presented. The use of different feature vectors, namely MFCC and LPCC, representing the spectral properties of the speech spectrum from different viewpoints was used in the training of the individual classifiers in the multiple classifier environment. The classifier combination was tested against two form of voice conversion attacks: identity disguise and targeted voice impersonation. The results showed that in both the cases, where the objectives of source/impostor speaker were to disguise their identity and to appear to the system as the target speaker, indicate a decrease in the success rate of the intrusion in the speaker identification system.

The LP-residual was investigated for speaker specific information and R-PDSS based features were proposed for use with the traditional spectral envelope based features. The combination of information extracted from the spectral envelope and the LP-residual indicate that the source identification performance increase by nearly 50% as compared to the case where only the spectral envelope parameters have been used.

Chapter 7

Conclusion and Future Work

This thesis presented different techniques, models and experiments to analyse the performance of Speaker Identification and Voice Conversion Systems. The first part of this chapter describes a summary of the contributions made in this thesis. The second part of the chapter is dedicated to describe future work that can be carried out to further the contributions in this thesis.

7.1 Conclusions

The main research outcomes of the work done in this thesis can be summarized as follows:

- *Baseline GMM based Speaker Identification and Speaker Verification Systems* have been developed and analysed for clean speech. The performance of the systems on the TIMIT database is comparable to the results presented in literature.
- *Voice Conversion System* based on spectral envelope transformation was implemented. The baseline systems transforms the spectral envelope as represented by the LPC spectrum. The transformation function was implemented as a regressive, joint density Gaussian Mixture model, trained on time aligned line spectral frequencies of the source and target speakers. The problem of over smoothing in voice conversion systems based on probabilistic approaches was analysed. A hybrid GMM and Linear Multivariate Regression adaptation technique was implemented to reduce the audible effect of over smoothing.
- The limited amount of training data for the transformation function results in discontinuities between the successive frames of the synthesized speech. To reduce the effect of the discontinuities, a posterior probability smoothing technique was proposed. Since the posterior probabilities are used as weights for the linear

combination of target feature vectors, smoothing the posterior probabilities results in a reduction of audible artefacts in the converted speech. Subjective evaluations also favour the converted speech that has been obtained as a result of posterior probability smoothing.

- The performance of the speaker identification system was analysed. Two different scenarios of deliberate modifications of the speech signal were proposed namely; identity disguise and targeted voice impersonation. The performance of the system was analysed in terms of the ability of the speaker identification system to identify the source and the target speakers from the converted voices. The performance of the speaker identification system was also analysed in terms of intra-gender and cross-gender voice conversions, with the results suggesting that for the conversion of intra-gender voices, the voice conversion system has inferior performance compared to the case of cross-gender.

- The use of multiple classifiers was investigated for the task of speaker identification. The use of GMM based classifiers using MFCC and LPCC as feature vectors are used in the framework of multiple classifier system against converted synthetic voices. Also the linear prediction residual of the speech signal is analysed for speaker specific information and the PDSS is used for the extraction of speaker specific information from the LP residual. Different combination of MFCC, LPCC and R-PDSS are explored in improving the performance of the speaker identification system against the identity disguise and targeted voice impersonation. The identification performance of the system using both spectral envelope features and features representing the LP residual outperform the traditional spectral envelope based classification techniques in the case of computer aided converted voices.

7.2 Future Work

Based on the research carried out in this thesis, this section proposes some research areas for future work.

- Speaker identification and speaker verification have been studied quite extensively in the field of pattern recognition. The performance of these systems has been analysed in clean, noisy and mismatch conditions thoroughly. However, with the emergence of voice conversion techniques the absence of suitable databases for testing the performance of these systems impedes the robustness of speaker identification and speaker verification systems against computer aided voice conversion attacks. Future work in this case, should involve development of speaker databases that are specifically designed with the threat of voice conversion to speaker recognition systems in view.

- The current state of the art voice conversion systems employ parallel speech corpus for the training of the transformation system and are known as text-dependent systems. The parallel text also needs to be time aligned to extract the correspondences between the source and target feature spaces. In order to obtain satisfactory performance, the training of the transformation function usually requires huge amounts of training data. A possible future work in the case of voice conversion systems would be the evolution of text-independent system, where the speech of the source and the target speaker need not be parallel and aligned in time.

- The speaker recognition systems generally involve features which represent the spectral envelope characterization of the speech signal only. It has been demonstrated in this thesis that such system are highly vulnerable to voice impersonation attacks where the author of the attack is able to modify his/her voice properties to match those of a target speaker to deceive the system. In this thesis the use of features representing the LP-residual have been used in combination with the traditional spectral envelope based features. The performance of the speaker identification system using these feature in a multiple classifier environment have shown to improve the robustness against converted voices. Future work may include the search for new features that may or may not be utilized by speakers, but which may contain useful speaker specific information. This will trigger a search for better modeling techniques since linguistic modeling is a difficult task in process.

7.3 Summary

In this chapter the main contributions of thesis are summarized and some suggestions for future work have been discussed. In summary, the thesis discussed the theoretical framework of the speaker identification and speaker verification systems. Baseline speaker identification and speaker verification systems are developed and analysed. Another major contribution of this thesis is the development of a voice conversion system. A hybrid solution is proposed to address the problem of over-smoothing in the GMM based voice conversion systems. Furthermore, a novel technique to alleviate the audible degradations in the converted speech signals is proposed based on the smoothing of the component posterior probabilities. The performance of the speaker identification systems is shown to deteriorate when presented with converted synthetic voices. In this regard, the performance evaluations are carried out in the scenarios of identity disguise and targeted speaker impersonation. The final contribution of the thesis proposes the use of multiple information sources for speaker identification problem. The use of speaker information present in the LP residual signal is also suggested for use with the traditional spectral envelope features. The use of multiple classifiers is shown to improve the robustness of the speaker identification system against synthetic converted voices. It is expected that the findings of this thesis will have an important bearing on

the robustness of the speaker identification systems when dealing with computer aided voice impersonation attacks.

References

[1] A. M. Kondoz. *Digital Speech: Coding for Low Bit Rate Communication Systems*. John Wiley & Sons, 2004.

[2] S. S. Stevens abd J. Volkmann and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of Acoustic Society of America*, 8:185–190, 1937.

[3] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines and Other Kernal-Based Learning Methods*. Cambridge University Press, 2000.

[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 655 –658 vol.1, apr 1988.

[5] T. F. Quatieri. *Discrete-Time Speech Signal Processing: principles and Practice*. Prentice Hall Inc., 2002.

[6] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall Inc., 1993.

[7] K. Fukunaga. *Introduction to Statistical Pattern Classification*. Academic Press Inc, 1990.

[8] J. R. Deller, J. G. Proakis, and J. H. Hansen. *Discrete-Time Processing of Speech Signals*. MacMillan Publishing Company, 1993.

[9] D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison Wiley Publishing Company, 1987.

[10] D. H. Klatt and L.C. Klatt. Analysis, synthesis and perception of voice quality variations female and male talkers. *Journal of Acoustical Society of America*, 87-2:820–857, 1990.

[11] H. Kawabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: Control and conversion. *Speech Communications*, 16:165–173, 1995.

[12] J. W. Picone. Signal modeling techniques in speech recognition. *Proc. IEEE*, 81, September 1993.

[13] Wai C. Chu. *Speech Coding Algorithms.* John Wiley & Sons, 2003.

[14] D. A Reynolds. *A Gaussian Mixture Modeling Approach to text-Independent Speaker Identification.* PhD thesis, Georgia Institute of Technology, 1992.

[15] D. A. Reynolds, R. C. Rose, and M. J. T. Smith. Pc-based tms320c30 implementation of the gaussian mixture model text-independent speaker recognition systems. *Proc. International Conference on Signal Processing Application and Technology*, pages 967–973, November 1992.

[16] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals.* Prentice, 1978.

[17] John Markhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE Transactions on Acoustics, Speech and Signal Processing*, 3:283–296, June 1975.

[18] J. P. Norton. *An Introduction to Identification.* Academic Press, 1986.

[19] F. Itakura. Line spectrum representation of linear prediction coefficients of speech signals. *Jour. Acoust. Soc. America*, 57:S35, April 1975.

[20] J. D. Markel and A. H. Gray. *Linear Prediction of Speech.* Springer-Verlag, 1976.

[21] U. Vishwanathan and J. Makhoul. Quantization properties of transmission parameters in linear predictive systems. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23:309–321, 1975.

[22] T. Eriksson, J. Linden, and J. Skoglund. Interframe lsf quantization for noisy channels. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 7, no. 5:495–509, September 1999.

[23] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing.* Prentice Hall Inc., 1989.

[24] B.P. Bogert, M.J. R. Healy, and J.W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, psuedo-autocovariance, cross-cepstrum and saphe cracking. *Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed)*, 15:209–243, 1963.

[25] R. W. Schafer. Echo removal by distance genaralized linear filtering. Technical report, Research Lab of Electronics, MIT, February 1969.

[26] B. S Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Acoust. Soc. of America*, 55(6):1304–1312, June 1974.

[27] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:357–366, August 1980.

[28] R. W. Schafer and L. R. Rabiner. Digital representation of the speech signals. *Proc. IEEE*, 63:662–677, April 1977.

[29] F. Bimbot, I. Margin-Chagnollean, and L. Mathan. Second-order statistical measures for text-independent speaker identification. *Speech Communications*, 81 no. 9:177–192, August 1995.

[30] F. K. Soong and A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36:871–879, June 1988.

[31] B. S. Atal. Automatic recogniton of speakers from their voices. *Proc. IEEE*, 64:460–475, APril 1977.

[32] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29:254–272, April 1981.

[33] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds. Intergrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 2:245–257, April 1994.

[34] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. magrin Chagnolleau, S. meigneir, Merlin T, Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied SIgnal Processing*, 4:430–451, April 2004.

[35] R. L. Klevans and R. D. Rodman. *Voice Recognition.* Artech House Inc., 1997.

[36] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang. A vector quantization approach to speaker recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 387–390, March 1985.

[37] j. P. Campbell. Speaker recognition a tutorial. *Proc. IEEE*, 85:1437–1462, 1997.

[38] J. Ashbourn. *Practical Biometrics: from Aspirations to Implementations.* Springer-Verlag, 2004.

[39] R. D. Rodman. *Computer Speech Technology.* Artech House Inc., 1990.

[40] A. Martin and M. Przybocki. The nist 1999 speaker recognition evaluation - an overview. *Digital Signal Processing*, 10:1–18, 2000.

[41] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assesment of detection task performance. *Proc. Eurospeech'97*, 4:1895–1898, 1997.

[42] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons Inc., 1973.

[43] H. Gish and M. Schmidt. Text-independent speaker identification. *IEEE SIgnal Processing Magazine*, 11:18–32, October 1994.

[44] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 2000.

[45] V. Wan. *Speaker Verification Using Support Vector Machines.* PhD thesis, University of Sheffield, 2003.

[46] S. Abe. *Pattern Classification - Neurao-fuzzy Methods and their Comparison.* Springer, 2001.

[47] Y. Gu and T. Thomas. A text-independent speaker verification system using support vector machines classifier. *Proc. Eurospeech'01*, pages 1765–1769, 2001.

[48] M. Schmidt and H. Gish. Speaker independent via support vector classifiers. *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1:105–108, 1996.

[49] V. Wan and S. Renals. Svmsvm: Support vector machine speaker verification methodology. *IEEE International Conference on Acoustics Speech and Signal Processing*, 13 no.2:221–224, 2005.

[50] J. Kharroubi, D. Petrovska-Delacretaz, and G. Chollet. Combining gmm's with support vector machines for text-independent speaker verification. *Proc. Eurospeech'01*, pages 1757–1761, 2001.

[51] S. Fine, J. Navratil, and R. A. Gopinath. Enhancing gmm scores using svm "hints". *Proc. Eurospeech'01*, pages 1761–1765, 2001.

[52] S. Haykin. *Neural Networks: A Comprehensive Foundation.* McMillan, 1994.

[53] J. Hertz, A. Krogh, and R. J. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley, Santa Fe Institute Studies in the Sciences of Complexity, 1991.

[54] J. Oglesby and J. S. Mason. Optimization of neural models for speaker identification. *Proc. IEEE Conference on Acoustic and Speech Signal Processing*, pages 261–264, April 1990.

[55] D. A. Reynolds. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3:72–83, January 1995.

[56] A. Gresho and R. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Press, 1992.

[57] Y. Linde, A. Buzo, and R. Gray. An algorithm for vactor quantization design. *IEEE Transactions on Communications*, 28 no. 1:84–95, October 1980.

[58] D. A. Reynolds. Speaker identification using gaussian mixture speaker models. *Speech Communications*, 17:91–108, 1995.

[59] A. Dampster, N. Liard, and D. Rubin. Maximum liklihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[60] J. N. Holmes and John Nicholas. *Speech Synthesis and Recognition.* Taylor & Francis, 2001.

[61] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 4–16, January 1986.

[62] A. E. Rosenburg, J. DeLong, C. H. Lee, B. H. Juang, and F. K. Soong. The use of cohort normalized scores for speaker verification. *International Conference on Speech and Language Processing*, pages 599–602, 1992.

[63] T. Matsui and S. Furui. Similarity normalization methods for speaker verification based on a-posteriori probability. *Proc. of the ESCA Workshop on AUtomatic Speaker Recognition, Identification and Verification*, pages 59–62, 1994.

[64] D. A Reynolds. Comparison of background normalization methods for text-independant speaker verification. *Proc. Eurospeech'97*, pages 963–966, 1997.

[65] D. A. Reynolds and T. F. Quatieri. Speaker verification using adapted gaussian mixture models. *Digital Speech Processing*, 10:19–41, 2000.

[66] 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*, 2:291–298, 1994.

[67] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The dapra speech recognition research database: Specifications and status. *Proc. DAPRA Workshop on Speech Recognition*, pages 93–99, February 1986.

[68] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. Ntimit: A phonetically balanced, continuous speech telephone bandwidth speech database. *Proc. International Conference of Acoustic, Speech and Signal Processing*, pages 109–112, April 1990.

[69] J. P. Jr. Campbell. *Features and Measures for Speaker Recognition.* PhD thesis, Oklahoma State University, 1992.

[70] A. Higgins, L. Bahler, and J. Porter. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1:89–106, 1991.

[71] J. J. Godfrey, E. C. Holliman, and J. MacDaniel. Switchboard: Telephone speech corpus for research and development. *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1:517–520, 1992.

[72] Dr. Alan Higgins and Dave Vermilyea. King speaker verification. Technical report, Linguistic Data Consortium, Philadelphia, 1995.

[73] www.nist.gov. www.nist.gov.

[74] The nist year 2005 speaker recognition evaluation plan.

[75] www.ldc.upenn.edu.

[76] J. S. Garofolo, L. F. lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic-phonetic continuous speech corpus. *U.S. Department of Commerce, National Institure of Standards and Tecnology*, February 1993.

[77] C. Ergun. *Speech Coder Detector Based Speaker Verification System in a Multi-Coder Environment.* PhD thesis, Eastern Mediterranean University, May 2004.

[78] L. C. Schwardt and J. A. Perez. Voice conversion based on static speaker characteristics. *Proc. IEEE*, 1998.

[79] A. Kain, J. Hosom, X. Niu, J. van Santen, M. Fried-Oken, and J. Staehely. Improving the intelligibility of dysarthic speech. *Speech Communication*, 49:743–759, 2007.

[80] A. Kain, X. Niu, J. Hosom, Q. Miao, and J. van Santen. Formant re-synthesis of dysarthric speech. *Proceedings of 5th ISCA Workshop on Speech Synthesis*, June 2004.

[81] M. Abe. A segment-based approach to voice conversion. *Proc. IEEE*, 2:765–768, April 1991.

[82] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6-2:131–142, March 1998.

[83] S. Furui. Research on individual features in speech waves and automatic speaker recognition techniques. *Speech Communications*, 5:183–197, 1986.

[84] H. Hollien. *The Acoustics of Crime-The New Science of Forensics.* New York: Plenum, 1990.

[85] C. C. Johnson, H. Hollien, and J. W. Hicks. Speaker identification utilizing selected temporal speech features. *Journal of Phonetics*, 12:319–326, 1984.

[86] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory.* Englewood Cliffs, NJ: Prentice Hall, 1984.

[87] K. Itoh and S. Saito. Effects of acoustical features parameters on perceptual speaker identity. *Rev. Electronics Communications Lab*, 36:135–141, 1988.

[88] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communications*, 56:12–40, 2010.

[89] I.. Karlsson. Glottal waveform parameters for different speaker types. *Proc. Speech '88 7th False Symposium*, 1:225–231, 1988.

[90] A. L. Lalwani and D. G. Childers. Modeling vocal disorders via formant synthesis. *Proc. International Conference of Acoustic, Speech and Signal Processing*, 1:505–508, April 1991.

[91] E. Moulines and Y. Sagisaka. Voice conversion: State of the art and perspectives. *Speech Communications (Special Issue) Elsevier*, 16, Feb 1995.

[92] Alexander Kain. *High Resolution Voice Transformation*. PhD thesis, OGI School of science and engineering, 2001.

[93] V. Fischer and S. Kunzmann. From pre-recorded prompts to corporate voices. *Proc. of the Interspeech*, 2006.

[94] E. M. Eide and M. A. Pichney. Towards pooled-speaker concatenative text-to-speech. *International Conference on Audio, Speech and Signal Processing*, pages 73–76, 2006.

[95] O. Turk and L. M. Arslan. Subband based voice conversion. *Proc. of the ICSLP*, pages 289–292, 2002.

[96] J. Hosom, A. Kain, T. Mishra, J. van santen, M. Fried-Oken, and J. Stachely. Intelligibility of modifications to dysarthic speech. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 1:924–927, 2003.

[97] H. Boøil, P. Fousek, D. Sündermann, P. Èreva, and J. Zdansky. Lombard speech recognition: A comparative study. *Proceedings of the 16th Czech-German Workshop on Speech Processing*, pages 141–148, 2006.

[98] A. Kain and M. Macon. Spectral voice conversion for text-to-speech synthesis. *Proc. of International Conference on Speech Audio and Signal Processing*, 1:285–288, May 1998.

[99] K. S. Lee, D. H. Youn, and I. W. Cha. A new voice transformation method using both linear and nonlinear prediction analysis. *Proceedings of ICSLP '96*, 3:1401–1404, 1996.

[100] L. M. Arslan. Speaker transformation using segmental codebooks (stasc). *Speech Communications*, 28:211–226, 1999.

[101] M. Narendranath, H. A Murthy, S. Rajendran, and B. Yegnanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech Communications*, 16-2:207–216, 1995.

[102] D. G. Childers. Glottal source modeling for voice conversion. *Speech Communications*, 16-2:127–138, 1995.

[103] H. Valbret, E. Moulines, and J. P. Tubach. Voice transformation using psola technique. *Speech Communications*, 11-2:175–187, 1992.

[104] L. M. Arslan and D. Talkin. Spealer transformation using sentence hmm based alignment and detailed prosody modification. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 289–292, 1998.

[105] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana. Extraction of speaker specific information from linear prediction residual of speech. *Speech Communication*, 28:1243–1261, 2006.

[106] L. M. Arslan and D. Talkin. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. *Proceedings of EUROSPEECH '97*, 3:1347–1350, 1997.

[107] C. Shih, W. Gu, and J. Van Santen. Efficient adaptation of tts duration models for new speakers. *Proceedings of ICSLP '98*, 1998.

[108] W. Gu, C. Shih, and P. H. Van Santen. An efficient speaker adaptation method for tts duration model. *Proceedings of EUROSPEECH '99*, September 1999.

[109] K. J. Kohler, J. P. H. Van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg. Parametric control of prosodic variables by symbolic input in tts synthesis. *Progress in Speech Synthesis*, 37:459–476, 1996.

[110] H. Mizuno and M. Abe. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. *Speech Communications*, 16-2:153–164, 1995.

[111] Y. Stylianou, O. Cappe, and E. Moulines. Statistical methods for voice quality transformation. *Proceedings of EUROSPEECH '95*, 9-1:21–29, 1995.

[112] A. Kain and M. Macon. Personalizig a speech synthesizer by voice adaptation. *International Spech Synthesis Workshop*, pages 225–230, November 1998.

[113] W. A. Ainsworth, K. K. Paliwal, and H. M. Foster. Problems with dynamic frequency warping as a technique for speaker-independent vowel classification. *Proc. Institute of Acoustics*, 6-4:303–306, 1984.

[114] G. Baudoin and Y. Stylianou. On the transformation of speech spectrum for voice conversion. *Proceedings of ICSLP '96*, 2:1405–1408, 1996.

[115] J. P. Egan. Articulation testing methods. *Laryngoscope*, 58:955–991, 1948.

[116] B. C. J. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, 1997.

[117] J. Wouters. Control of spectral dynamics in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9-1:30–38, January 1991.

[118] M. Unser, A. Adroubi, and M. Eden. B-spline signal processing. *IEEE Transactions on Speech and Signal Processing*, 41-2:821–833, Februray 1993.

[119] S. G. Kang and L. J. Fransen. Applications of line-spectrum pairs to low-bit rate speech encoders. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 244–247, 1985.

[120] K. K. Paliwal and B. S. Atal. Efficient vector quantization of lpc parameters at 24 bits/frame. *IEEE Transactions on Speech and Signal Processing*, 1:3–14, January 1993.

[121] F. K. Soong and B. H. Huang. Line spectrum pair and speech data compression. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 1.10.1–1.10.4, 1983.

[122] N. Kambhatla. *Local Models and Gaussian Mixture Models for Statistical Data Processing*. PhD thesis, Oregon Graduate Institute of Science and Technology, January 1996.

[123] T. F. Quatieri and R. J. McAulay. Shape invariant time-scale and pitch modification of speech. *IEEE Transaction on Signal Processing*, 40-3:497–510, March 1992.

[124] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Communications*, 4:1–58, 1992.

[125] Y. Chen, M. Chang, J. Liu, and R. Liu. Vocie conversion with smoothed gmm and map adaptation. *EUROSPEECH'03*, pages 2413–2416, 2003.

[126] K. Koishida, K. Tokuda, T. Kobayashi, , and S. Imai. Celp coding based on mel-cepstral analysis. *International Conference on Audio, Speech and Signal Processing*, 1:33–36, 1995.

[127] E. Zotterholm. *Vocie Immitation: A Phonetic study of perceptual illusions and acoustic success*. PhD thesis, Travaus de l'institut de linguistic de Lund 44. Lund University , Department of Linguistics and Phonetics. Lund, 2003.

[128] R. D. Rodman. Speaker recognition of disguised vocies. *proceedings of the consortium on speech technology on speaker recognition by man and machines: directions for forensic application, Ankara, Turkey*, pages 9–22, 1998.

[129] J. Lindberg and M. Bloomberg. Vulnerability in speaker verification. a study of technical impostor techniques. *In proceedings of interspeech, Budapest, Hungary*, September 1999.

[130] T. Masuka, K. Tokuda, and T. Tobayashi. Impostor using synthetic speech against speaker verification based on spectrum and pitch. *In proceedings of ICSLP. Beijing, China*, 2, October 2000.

[131] H. J. Kunzel, J. Gonzalez-Rodriguez, and J. Ortega-Garcia. Effect of voice disguise on the performance of a forensic automatic speaker recognition system. *in proceedings of the speech and language recognition workshop*, pages 153–156, May 2003.

[132] K. P. H. Sullivan and J. Pelecanos. Revisiting carl bildt's impostor: would a speaker verification system fioil him? *In proceedings of 3rd international conference on Audio and Video based biometric person authentication. Halmstad Sweden*, 2091:144–149, 2001.

[133] E. Zotterholm, M. Bloomberg, and D. Elenius. A comparison between human perception and a speaker verification system score of a vocie immitation. *In procedings of the 10th australian international conference on speech science and technology, Sydney Australia*, pages 393–397, 2004.

[134] Hermann J. Künzel. Effects of voice disguise on speaking fundamental frequency. *International Journal of Speech Language and the Law*, 7 No. 2, 2000.

[135] D. Markham. *Phonetic Immitation, Accent and Learner*. PhD thesis, Department of Linguistics and Phonetics, Lund University, 1997.

[136] J Laver. *Principles of Phonetics*. Cambridge University Press, 1994.

[137] J. Pittam. *Voice in social interaction: an interdisciplinary approach*. Thousand Oaks: SAGE publications, 1994.

[138] N. J. Lass, D. S. Trapp, K. A. Scherbick M. K. Baldwin, and D. L. Wright. Effect of vocal disguise on judgments of speakers' sex and race. *perceptual and motor skills*, 3-2:1235–1240, 1982.

[139] R. W. Shuy. Dialect as evidence in law cases. *Journal of English Linguistics*, 23-1:195–208, 1995.

[140] D. A. Reynolds, W. Andrews, J. campbell, J. Navratill, B. Peskin, A. Adami, Q. Jin, D. Kulaseck, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang. The supersid project: exploiting high level information for high accuracy speaker recognition. *In proceedings of ICASSP, China*, 2003.

[141] M. Farrus, J. Hernando, J. Luque, and P. Ejarque. On teh fusion of prosody, voice spectrum and face features for multimodal person verification. *In proceedings of ICSLP*, pages 2106–2109, September 2006.

[142] M. Farrus, J. Hernando, and P. Ejarque. Jitter and shimmer measurements in speaker recognition. *In proceedings of Interspeech, Antwerp Belgium*, pages 778–781, 2007.

[143] H. Duxans. *Vocie conversion applied to text-to-speech systems*. PhD thesis, universitat politecnica de catalunya, department of signalm theory and communications. barcelona, 2006.

[144] J. Kittler. On combining classifiers. *IEEE Transaction on Pattern Matching and Machine Intelligence*, 20-3:226–239, March 1998.

[145] T. Ho, J. Hull, and S. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Parallel and Distributed Systems*, 16:66–75, 1994.

[146] J. A, Benediktsson, and P. H Swain. Consensus theoretic classifications methods. *IEEE Transactions on Systems Man. and Cybernatics*, 22:688–704, July 1992.

[147] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7-5:777–781, 1994.

[148] R. Battiti and A. M. Colla. Democracy in neural nets: Voting schemes in clasification. *Neural Networks*, 7-4:691–707, 1994.

[149] L. Lam and C. Y. Suen. Optimal combinations of pattern classifiers. *Pattern Recognition Letters*, 16:945–954, 1995.

[150] A. Saranh. *A unifying theory for rank-based multiple classifier systems, with applications in speaker identification and speech recognition.* PhD thesis, Middle East Technical University, 2000.

[151] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernatics*, 22:418–435, 1992.

[152] I. Bloch. Information combination operators for data fusion: A comparative review with classification. *IEEE Transaction on systems Man and Cybernatics*, 26-1:52–67, January 1996.

[153] K. Al-Ghoneim and B. V. Kumar. Unified decision combination framework. *Pattern Recognition*, 7-4:691–707, 1998.

[154] R. K. Bhatnagar and L. N. Kanal. *Handling uncertain information: A review of numeric and non-numeric methods.* Elseiver Science Publishers, 1986.

[155] N. S. Lee, Y. L. Grize, and K. Dehand. Quantitative models for reasoning under uncertainity in knowledge-based expert systems. *International Journal of Intelligent Systems*, 2:15–38, 1987.

[156] G. Shafer. The art of casual conjecture. *MIT Press*, 1996.

[157] Y. Viniotis. *Probability and stochastic processes for electrical engineers.* McGraw-Hill Inc, 1998.

[158] J. Pearl. *Probabilistic Resoning in Intelligent Systems.* Morgan Kaufmann Publishers Inc, 1988.

[159] X. Lin, X. Ding, M. Chen, R. Zhang, and Y. Wu. Adaptive confidence transform based classifier combination for chinese character recognition. *Pattern Recognition Letters*, 19:975–988, 1998.

[160] J. Benediktsson. Parallel consensual neural networks. *IEEE Transactions on Neural Networks*, 8-1:54–64, January 1997.

[161] H. Altincay. *Improving the performance of speaker identification system by classifier combination techniques.* PhD thesis, Middle East Technical University, 2000.

[162] D. A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and audio processing*, 4-2:639–643, 1994.

[163] G. Doddington, M. Przybocki, A. Martin, and D. Reynolds. The nist speaker recognition evaluation overview, methodology, systems, results, perspectives. *Speech Communications*, pages 225–254, 2000.

[164] A. Jain, R. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Transaction on pattern Analysis, machines and Intelligence*, 22:4–37, 2000.

[165] H.Altincay and M. Demirekler. Speaker identification by combining multiple classifiers using dempster-shafer theory of evidence. *Speech Communication*, 41:531–547, 2003.

[166] P. Perrot and G. Chollet. the question of disguised voice. *Acoustics*, pages 5682–5685, 2008.

[167] D. Reynolds. Large population speaker identification using clean and telephone speech. *IEEE Signal Processing Letters*, 2:46–48, 1995.

[168] D. Mashao and N. Baloyi. Improvements in speaker identification rate using feature sets on large population database. *Proceedings of EUROSPEECH '2001*, 4:2833–2836, 2001.

[169] L. Lerato. Hierarchical methods for large population speaker identification using telephone speech. Master's thesis, University of Capr Town, Cape Town, South Africa, 2003.

[170] D Reynolds. Speaker identification and verification using gaussian mixture speech models. *speech communications*, 17:91–108, 1995.

[171] K. Chen and H. Chi. A method of combining multiple probabilistic classifiers through soft competition on different feature sets. *Neurocomputing*, 20:227–252, 1998.

[172] R. Ramachandran, K. Ferrel, and R. Mammone. Speaker recognition general classifiers approaches and data fusion methods. *Pattern Recognition*, pages 2801–2821, 2002.

[173] F. Roli and. Fusion of multiple pattern classifiers. *English National conference of the italian association on artificial intelligence*, 2003.

[174] R.E. Slyh, E.G. Hasnen, and T. R Anderson. Glottal modelling and closed-phase analysis for speaker recognition. *Proceedings on ISCA Tutorial andResearch Workshop on Speaker and Language Recognition*, pages 315–322, 2004.

[175] L. Mary, K. Sri Rama Murty, S. R.M. Prasanna, and B. Yegnanarayana. Features and speaker and language identification. *Proceedings of ISCA Tutorial and Research Workshop on Speaker and Language Recognition*, pages 323–328, 2004.

[176] M. D Plumpe, T. F. Quatieri, and D. A Reynolds. Modelling of the glottal flow derivative waveform with applications to speaker identification. *IEEE Transactions on Speech Audio Processing*, 1:569–586, 1999.

[177] R. Smits and Yegnanarayana B. Determination of instants of significants excitation in speech usinf group delay function. *IEEE Transaction on Audio, Speech and Signal Processing*, 3:325–333, 1995.

[178] N. Zheng, T. Lee, and P.C. Ching. Integration of complimentary acoustic features for speaker recognition. *IEEE Signal Processing Letters*, 2006.

[179] P. Thevenaz and H. Hugli. Usefulness of lp-residual in text-independant speaker verification. *IEEE Signal Processing Letters*, 17:557–610, 1995.

[180] B. Yegnanarayana, K.S. Reddy, and S. P. Kishore. Source and system features for speaker recognition using aann models. *IEEE Transaction on Audio, Speech and Signal Processing*, pages 409–412, 2001.

[181] G. Kubin. Non-linear processing of speech signals. *In W. B. Kleijn, K. K. Paliwal Speech Coding and Synthesis*, pages 557–610, 1995.

[182] A. Espsito and M. Marinaro. Some notes on non-linearities in speech. *G. Chollet et al., Nonlinear speech modeling, lecture notes in artificial intelligence*, 3445:1–4, 2005.

[183] S. Gazor and W. Zhang. Speech probability distribution. *IEEE Signal Processing Letters*, 7:204–207, 2003.

[184] E. Rank and G. Kubin. Non-linear synthesis of vowels in the lp residual domain with a regularized rbf network. *Proceedings of IWANN*, 2085:746–753, 2001.

[185] H. Thyssen, H. Nielson, and S. D. Hansen. Non-linearities short-term prediction in speech coding. *IEEE Transaction on Audio, Speech and Signal Processing*, 1:185–188, 1994.

[186] S. Hayakawa, K. Takeda, and F. Itakura. Speaker identification using harmonic structure of lp-residual spectrum. *Audio Video Biometric Personal Authentification, Lecture Notes in Computer Science*, 1206:253–260, 1997.