Robust Speaker Recognition in Reverberant Condition-Toward Greater Biometric Security

Khamis Ahmed Yousif AL-Karawi

School of Computing, Science and Engineering

University of Salford, Salford, UK

Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy, April 2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ASR	Automatic Speaker Recognition
AGMM	Adaptive Gaussian mixture model
AIR	Acoustic Impulse Response
ANN	Artificial Neural Networks
AR	Auto-Regressive
CASA	Computational Auditory Scene Aanalysis
CMS	Cepstral Mean Subtraction
DCF	Decision Cost Function
DCT	Discrete cosine Transform
DFT	Discrete Fourier Transform
DET	Detection Error Trade-off
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation Maximization
FAR	False Acceptance Rate
FRR	False Rejection Rate
FFT	Fast Fourier Transform
FSC	Frame Score Competition
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
JFA	Joint Factor Analysis
ISM	Image Source Method
LFCC	Linear Frequency Cepstral Coefficients

LLR	Log-Likelihood Ratio
LR	Likelihood Ratio
LPC	Linear Prediction Coefficients
LPCC	Linear Prediction Cepstral Coefficients
MAP	Maximum Posteriori Estimation
MAP	Maximum A posteriori
MBCM	Multiple Binary Classifier Model
MFCC	Mel Frequency Cepstral Coefficients
GFCC	Gammatone Frequency Cepstral Coefficients
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MSR	Microsoft Speaker Recognition
NIST	National Institute of Standards and Technologies
PCA	Principal Component Analysis
PDF	Probability Density Function
PLP	Perceptual Linear Predictive
PNN	Probabilistic Neural Network
RBM	Reverberation Background Model
RCC	Real Cepstral Coefficients
RCM	Reverberation Classification Model
RIR	Room Impulse Response
RT	Reverberation Times
DRR	Direct to Reverberation Ratio
SAD	Speech Activity Detection

SNR	Signal to Noise Ratio
SVM	Support Vector Machine
SV	Speaker Verification
SID	Speaker Identification
TMIT	Texas Instruments (TI) and Massachusetts Institute of Technology
UBM	Universal Background Model
VQ	Vector Quantization
ZCR	Zero Crossing Rate
ACF	Autocorrelation Function

LIST OF SYMBOLS

$\sum$	Summation
F	Frequency
Ω	Angular Frequency
s(n)	Audio Signal
Fs	Sampling Rate
h(t)	Impulse Response
Log	Logarithm
Exp	Exponential
SNR	Signal to Noise Ratio
Ln	Natural Logarithm
y(t)	Received signal
Y(f)	Received Signal Fourier Transform
sign(x)	Sign of x
V	Volume
$\otimes$	Convolution Operator
θ	Model Parameter Vector
n(t)	Noise Signal
FFT	Fast Fourier Transform
IFFT	Inverse Fast Fourier Transform
α	Absorption Coefficient
W(t)	White Noise
H	Hilbert Transform

ACKNOWLEDGEMENT

I am very grateful to ALLAH who is the all-powerful and all-knowing creator of this opportunity to learn a lot and accomplish this thesis. I ask sincerity in all my actions from Almighty Allah, and I quote the verse from the Holy Quran

(مَا يَفْتَحِ اللَّهُ لِلنَّاسِ مِن رَّحْمَةٍ فَلَا مُمْسِكَ لَهَا وَمَا يُمْسِكْ فَلَا مُرْسِلَ لَهُ مِن بَعْدِهِ وَهُوَ الْعَزِيزُ الْحَكِيمُ) (فاطر- اية 2)

“Whatever Allah grants to people of mercy none can withhold it; and whatever He withholds none can release it thereafter. And He is the Exalted in Mighty, the Wise.” (Chapter Fatir, verse 2).

IAM GREATLY INDEBTED to Dr. Francis Li, who supervised the work in this thesis. He gives me a lot of advice, practical help and mostly the time for our important weekly meeting. He is leading me at best and putting me in the optimal working conditions. I sincerely thank him. During the previous four years, he has been the most important source of academic inspiration, technical advisor and mentor of research methodology for this study. This thesis could not have been completed without his support and encouragement. I am so grateful to the Dr. Paul Kendrick for his kind, interesting, generous support and constant advice throughout the past three years.

Also indigent of acknowledgement that I am very grateful to my PhD studentship sponsored by the Iraqi government represented by “the Ministry of Higher Education and Scientific Research” which makes this study in this field possible and helped provide such an educational environment to work in. Furthermore, I would also like to recognise with much thankfulness the role of all those employees involved in Iraqi Cultural Attaché in London throughout my PhD studies, notably the former and present Counsellors Attaché

I, of course, want to thank my wonderful family, especially my wife and my children *‘‘**Sara, Hiba and Abdullah**’’*, who have come to my life and have turned the whole thing

in my life beautiful. Their support and care have helped me and made it possible to finish this work. I know that I have been so selfish in spending time on this thesis, but you always supported me. Indeed, without your prayers, love, sacrifices and support, I would not have reached this point in my life, and this PhD research work would not have been possible. Thank you for your love, support and care.

Gratitude also extends to the soul of, my parents, brother, sister and father-in-law, thanks for your support in my previous life.

I also wish to express sincere thanks to my home University, Diyala University and colleagues for their help and support.

Many thanks are also due to my colleagues and friends at the University of Salford (Duraid, Naif Al-otabi, Anugrah (Nano), Josh, James, Usman, Will Bailey, Alex, and Nikilash) and for other university students for the discussion, the joy and the participation in collection our database. Thank you to Henry from the hatch, which always helps me, prepare the equipment for recording database. Gratitude is also extended to Catriona Nardone, and to those members of staff within the School of Computing, Science and Engineering who gave me tremendous support, advice and help.

Finally, I would like to thank the members of my viva committee, Dr Omar Alani, and Dr. Mahmoud Shafik for their constructive comments, which have greatly assisted me to improve the quality of the thesis.

ABSTRACT

Automatic speaker recognition systems have developed into an increasingly relevant technology for security applications in modern times. The primary challenge for automatic speaker recognition is to deal with the variability of the environments and channels from where the speech was obtained. In previous work, good results have been achieved for clean, high-quality speech with the matching of training and test acoustic conditions. However, under mismatched conditions and reverberant environments, often expected in the real world, system performance degrades significantly.“ The main aim of this study is to improve the robustness of speaker recognition systems for real-world applications in reverberant conditions by developing methods that can reduce the detrimental effects of reverberation on the single microphone speech signal”.

The collection of suitable speech data sets is of crucial importance for testing the performance in the development of speaker recognition techniques. Therefore, a data set of anechoic speech recordings was generated and used to conduct the study regarding the suggested methods in this thesis. Furthermore, a typical speaker recognition system was implemented and then evaluated based on the current state of the art technique using Gaussian Mixture Models with two standard features. The effect of “reverberation time” and the “distance from the source to a receiver” on the system performance have also been examined, and the result confirms that whilst both parameters could affect the system accuracy.

A “maximum likelihood algorithm” is used for blind-estimate reverberation time from speech signals submitted for verification. The estimated values are used to choose a matched acoustic impulse response for inclusion in the retraining or fine-tuning of the pattern recognition model.

To endeavour more improvement, the “autocorrelation function” has been used to estimate the early reflections sound value for the submitted signal. The estimated early reflections sound value has convolved with the anechoic signal, and then used for training the pattern recognition model. Furthermore, both of the early to late ratio and RT have identified for the submitted sample and practically used to determine a matched channel for the training on the fly to improve the system performance.

The principal findings are that “reverberation time”, “early reflections” and “early to late ratio” can be estimated and then used with “training on the fly methods” to improve the speaker verification performance. The system is an improvement, which is demonstrated by comparing the performance of speaker recognition using “conventional methods” with the performance of the proposed “re-training method”.

CHAPTER ONE: INTRODUCTION

This chapter introduces speaker recognition system followed by the research question of the thesis, the research motivation, the thesis aim and objectivse, followed by the research methodology and the structure of each chapter. Finally, the publications resulting from the research are listed.

1.1 Introduction

Speaker recognition is defined as the process that helps to identify a talker from his\her voice. Speaker recognition has been a research topic for at least thirty years in universities and institutes around the world (Mohn, 1970). The initial studies of speaker recognition were published in 1971 (Bricker et al., 1971; Doddington, 1971). A number of reviews and tutorial papers confirm the wide spectrum of the studies published since then (Bimbot et al., 2004; J. P. Campbell, 1997; Tomi Kinnunen & Li, 2010). The research is still ongoing, and an increasing number of commercial applications are appearing; studies related to the voice have thus both scientific and economic relevance. Some of these applications are only physiological and cannot be altered by the individual; examples can be found in the patterns in the fingerprint, iris, or even DNA. Other measurements are a combination of physiological and behavioural cues; the voice is included in this category. Usually, the non-behavioral features are more robust; a perpetrator has more difficulty in modifying his/her fingerprints than in, for instance, trying to mimic a given voice. Non-behavioural features are also more reliable because they have minimal variability for a given individual while showing significant differences between people. The voice has a degree of variability between different speakers, but also exhibits a wide expressive range for a single given speaker; there are, in fact, two variability sources for a speaker's voice: voluntary and involuntary. The former can be a problem because a speaker can use this variability to hide his/her identity. The latter is due, for instance, to pathologies (like the flu or ageing), and is problematic when a target speaker risks not being recognised by the system. “The voice as a biometric measure, however, is still an interesting topic because it has one main advantage compared with other strategies: samples are readily available from an individual, with minimal human and technical effort”. Unlike some other biometrics, such as fingerprint analysis, speaker recognition systems typically have insufficient control over the equipment used to gather samples due to the physical separation of the claimant and the system for most applications. Due to their ubiquity and familiarity for users, telephones and telephone networks are a natural choice of a sampling device for many applications. Figure 1.1 illustrates a typical speaker recognition system.

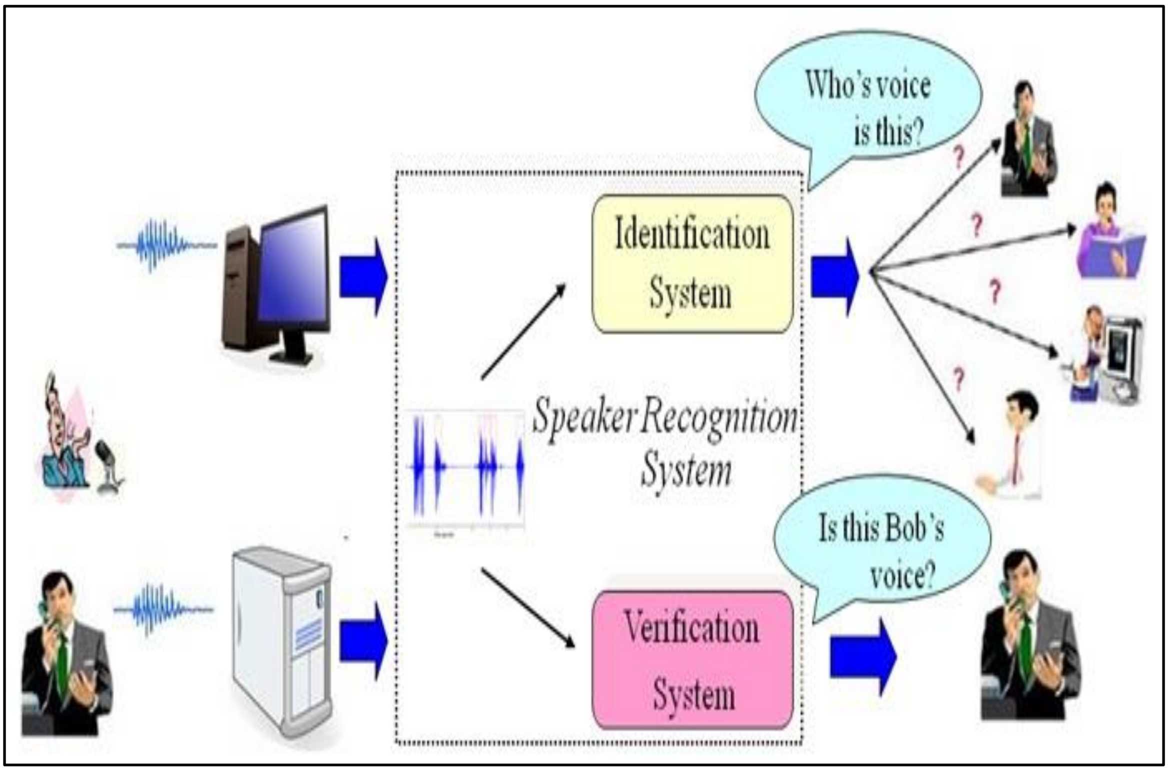


Figure 1.1 Traditional Speaker Recognition systems (C-DAC)

A speaker recognition system, performing either speaker identification (SID) or speaker verification (SV) tasks (J. P. Campbell, 1997), generally includes three processes: Feature extraction, speaker modelling, and decision-making (J. P. Campbell, 1997; Furui, 2009). Speaker features are encoded speaker-specific characteristics and are extracted from time domain signals. Usually used speaker features comprise short-time spectral/cepstral features, spectro-temporal features, prosodic features, etc. Short-time features are derived from short-time Fourier transform (STFT). Especially, time domain signals are fragmented into frames with around 20 ms duration. STFT is applied to the frames to acquire a magnitude spectrum. The spectral envelope reflects the resonance property of the vocal tract, which is closely related to the concept of formants. Typically, extracted speaker features are short-time spectral/cepstral features such as short-time Fourier transform spectral features and “mel-frequency cepstral coefficients (MFCC)”, or long-term features like prosodic features (Shriberg, 2007; Shriberg, Ferrer, Kajarekar, Venkataraman, & Stolcke, 2005). “Short-time features aim to capture vocal tract information, while long-term features mostly extract the different speaking styles”. As for speaker modelling, Gaussian Mixture Models (GMM) are usually utilised to model speaker feature distributions (D. A. Reynolds, 1995). “The literature strongly specified that, to date, the MFCC feature extraction joint with the GMM modelling and classification procedures are extensively recognised as the state of art techniques providing the best speaker recognition results(Memon, 2010). However, recent state-of-the-art speaker verification systems typically employ i-vector, or joint factor analysis (JFA) on super-vectors (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011; Kenny, 2005b; Kenny, Boulianne, Ouellet, & Dumouchel, 2007) which are high-dimensional feature vectors, to explicitly model both channel and speaker characteristics. “The experiments described in this thesis use MFCCs, GFCC features and GMM classifiers as the baseline method. For speaker identification (SID), recognition decisions are typically made based on likelihoods of observing data given applicant speaker models. The decision process of SV usually compares the score of the claimed speaker with a threshold to either accept or reject the claimed speaker”.

Automatic speaker recognition systems can realise a high level of performance in well-matched conditions. However, the performance drops significantly as speech is distorted by interference (Gong, 2002; Shao & Wang, 2006). To address this issue, several efforts have been made, including microphone arrays (González-Rodríguez, Ortega-García, Martín, & Hernández, 1996), feature normalization (Ganapathy, Pelecanos, & Omar, 2011), and alternative feature spaces and specially tailored training strategies (Krishnamoorthy & Prasanna, 2009; Ming, Hazen, Glass, & Reynolds, 2007), with limited success. Speech enhancement approaches and robust speaker features have been explored to realise noise robustness (May, van de Par, & Kohlrausch, 2012; Pullella, Kuhne, & Togneri, 2008; Shao, Srinivasan, & Wang, 2007; N. Wang, Ching, Zheng, & Lee, 2011). Research in the speaker recognition field moved on to the removal of channel effects, i.e. dereverberation methods, or in more general terms, speech cleaning, or blind channel equalization (Ning, Ching, Nengheng, & Tan, 2011; Sadjadi & Hansen, 2012; Zhang, Wang, & Kai, 2014).

“Blind dereverberation algorithms have been utilised to restore the anechoic signal or the early part of reverberant speech (Sadjadi & Hansen, 2014)”. Alternatively, one can present reverberation to speaker models to decrease the mismatch caused by reverberation (Akula, Apsingekar, & De Leon, 2009). All speech-cleaning methods employ estimation methods and thus impose distortions on the speech signals, while they attempt to remove the reverberation. For cosmetic improvement of perceived quality, speech cleaning has been found useful, but for speaker recognition, its effectiveness is insufficient, since the features enabling the discrimination of speakers are vulnerable to de-reverberation methods (Francis F Li, 2016).

1.2 Speaker Recognition System Challenges

The performance of speaker recognition drops significantly due to different factors such as acoustic environments, additive noise, room reverberation, and speaker health and channel/handset variations (Beigi, 2011; Kelly, 2014). These factors conspire to pose considerable challenges to such systems. Although research is still ongoing, there has already been a dramatic improvement and an increasing number of commercial applications (Beigi, 2011). However, the field still poses open issues and many active research centres around the world are working towards more reliable and better-performing systems (Beigi, 2012). The main drawbacks of speaker recognition are that the voice depends on the health condition of the subject (Beigi, 2012), that the voice varies throughout life (Beigi, 2009), and that the microphone or the channel used to transmit the voice has a significant effect on the speech signal (Jin, Schultz, & Waibel, 2007). Another drawback is the ‘time-lapse effect’, which occurs because of changes in speaker phonation due to changes in the environment (Beigi, 2009). Illness, whispering, speech under stress, ageing and colds can also result in a hindrance to voice production and consequently alter the natural speech of a person. These can be classified as short-term (stress, illness, cold, and whispering) and long-term (ageing) (Beigi, 2009; Kelly, 2014).

Furthermore, the microphone used in the training stage might be different from the one employed in testing. This is called channel mismatch and refers to the variation between the reference model of the speech signal and a given recognition speech signal for the same individual due to physical factors (e.g. change of handset) or “environmental factors (additive background noise and reverberation)” (Akula & de Leon, 2008; Bimbot et al., 2004; Castellano, Sradharan, & Cole, 1996; Gammal, 2004; Jin et al., 2007; Tomi Kinnunen & Li, 2010; Mammone, Zhang, & Ramachandran, 1996; Peer, Rafaely, & Zigel, 2008). “Previously it was thought that the mismatch was solely related to a change of handset. However, in addition to the handset mismatch, changes in environmental noise, acoustic properties of ambience (e.g. reverberation), microphone distance and angle (far-field), as well as many other factors, can cause this kind of mismatch (Beigi, 2011)". So far, different research efforts have been concerned with improving techniques for dealing with these challenges.

That has led to methods that have the ability to explicitly model the channel variability of the input speech utterances (Avila, Sarria-Paja, Fraga, O'Shaughnessy, & Falk, 2014; Pillay, Ariyaeeinia, Sivakumaran, & Pawlewski, 2009). "Lastly, another fundamental challenge related to speaker recognition is the unwanted changes in speech features due to environmental factors". This kind of variation could cause a mismatch between the corresponding test and the enrolment material of the same speaker, which would adversely affect the performance of the speaker recognition regarding accuracy.

1.3 Research Motivation

As discussed in previous sections speaker recognition technology has a broad range of applications and many potential applications require hands-free sound captures, such as automatic teller machine authentication, the production of video conference transcripts, and security access to buildings or vehicles, etc. Speaker recognition has achieved good performance under controlled conditions. However, real-world conditions differ from laboratory conditions. Mismatches exist between training and testing phases, such as background noise and reverberation. These factors consequently induce performance degradation in automatic speaker recognition systems.

The degradation becomes more prominent as the microphone is positioned more distant from the speaker (Jin et al., 2007). The degradation of automatic speaker recognition systems because of the mismatch motivates us to:

- achieve high accuracy in automated speaker recognition in reverberant conditions.
- enable the use of automatic speaker recognition for critical applications such as security and forensics.
- find the optimal solution to the problem that adverse acoustic conditions in the real world mitigate performance.

1.4 Problem Definition

Speaker recognition has been deployed for different applications in various acoustic conditions, and the reliability of such systems is becoming a concern. Although incremental improvements are being made every day in all branches of speaker recognition, the “channel and audio type mismatch” still seems to be the biggest hurdle in achieving perfect results in speaker recognition. It should be noted that accurate results are asymptotes and will probably never be achieved. “Reverberation represents one of the most significant challenges for speaker recognition (especially for speaker verification) in matched and mismatched conditions (Jin, 2007). The quality of speech passed through a speaker recognition system will affect the overall system performance. The degradation of this speech quality is apparent in many forms of additive noise and reverberation (Gong, 2002; Francis F Li, 2016; Shao & Wang, 2006; Zhao, Wang, & Wang, 2014). In addition, mismatched acoustic transmission channels are responsible for the degradation as many authors have identified (Akula et al., 2009; Akula & de Leon, 2008; Bimbot et al., 2004; Castellano et al., 1996; Gammal, 2004; Jin et al., 2007; Tomi Kinnunen & Li, 2010; Mammone et al., 1996; Peer et al., 2008). Blind channel equalisation and the removal or reduction of channel effects, as suggested by many authors, approximately equalises the channels and, to some extent, mitigates the mismatching issue at the cost of added distortions to the vulnerable speech signals themselves, but this simultaneously distorts the transmitted signals that are crucial for speaker recognition. Furthermore, speech enhancement methods and dereverberation methods do not handle the reverberation issue well. In this case, therefore, effectiveness is limited (Francis F Li, 2016). "This thesis investigates different methods to improve the robustness of speaker recognition, specifically speaker verification in reverberant environments, as well as examining the degradation in several algorithms, resulting from a mismatch in training and testing due to reverberation".

1.5 Research Questions

In response to the concerns mentioned above, this thesis seeks to answer the following research questions:

1) How can the robustness of speaker recognition system for real-world applications in the presence of reverberation be developed and improved?
2) What suitable features can be extracted from the reverberant signal?
3) To what extent can the estimation of some parameters of the acoustic room improve the performance of speaker recognition?

1.6 Research Aim and Objectives

1.6.1 Aim

The overall aim of this work is to reduce the effect of reverberation on the speech signals obtained from a single microphone and then improve the usefulness of a speaker recognition system for practical applications in the presence of reverberant conditions.

1.6.2 Objectives

To achieve this aim, the specific research objectives are summarised as follows:

- To produce and published a data set of anechoic speech to conduct an experimental study of the suggested methods in this thesis. This data set could be essential or useful for this study and many other disciplinary studies such as speech processing and speaker recognition and with a different condition.

- Implement and evaluate a speaker recognition system based on state of the art Gaussian Mixture Model–Universal Background Model (GMM-UBM) and Mel frequency cepstral coefficient (MFCC) feature, and then examine the impact of reverberant speech on the performance of this system. Therefore, MSR toolbox was employed as speaker recognition system, and the performance of this system is examined using clean and reverberant speech.
- To achieve a good understanding of the impact of reverberation on speech feature vectors and then classify the usefulness, sensitivity, and increased robustness of features utilised in the novel speaker feature, gemmation frequency cepstral coefficients (GFCC) system proposed by Shao (Shao et al., 2007), which has been claimed to be more robust than the traditional mel frequency cepstral coefficients (MFCC) features in noisy conditions. The objective is to examine the reverberant-robustness of MFCC and GFCC features. Therefore, the robust feature will be used in this thesis.
- To investigate the effect of the reverberation time and the source-receiver distance on the system performance.
- Improve the robustness of speaker recognition that has been done through different methods such as:
  - ✓ Improve system robustness by adding various reverberation times in the enrolment phase.

  - ✓ Use a maximum likelihood (ML) framework to estimate the reverberation time from the received reverberant speech signal in the first stage. The Training on the Fly Schema is then used to mitigate the reverberation effect in the second stage.
  - ✓ Use the estimated early reflections from the received a reverberant speech signal to improve the robustness of speaker recognition.
  - ✓ Use the reverberation time and early to late ratio with training on the fly method to improve the robustness of speaker recognition.
- To evaluate the empirical results of the proposed methods in this study.
- Finally, to identify the limitations of the method developed through this study and suggest future work.

1.7 Research Steps

An experimental methodology is adopted for this research to generate data by hypotheses and experiments followed by extensive simulations and tests. This approach is used in positivist research studies as proposed in (Charoenruk, 2012). This methodology enables the researchers to follow steps including the definition outline, implementing, processing and evaluating the results. Additionally, constructive comments of the supervisor and conferences and journal reviewers have been considered to steer this research. Accordingly, the main phases of our research methodology are shown in Figure 1.2.

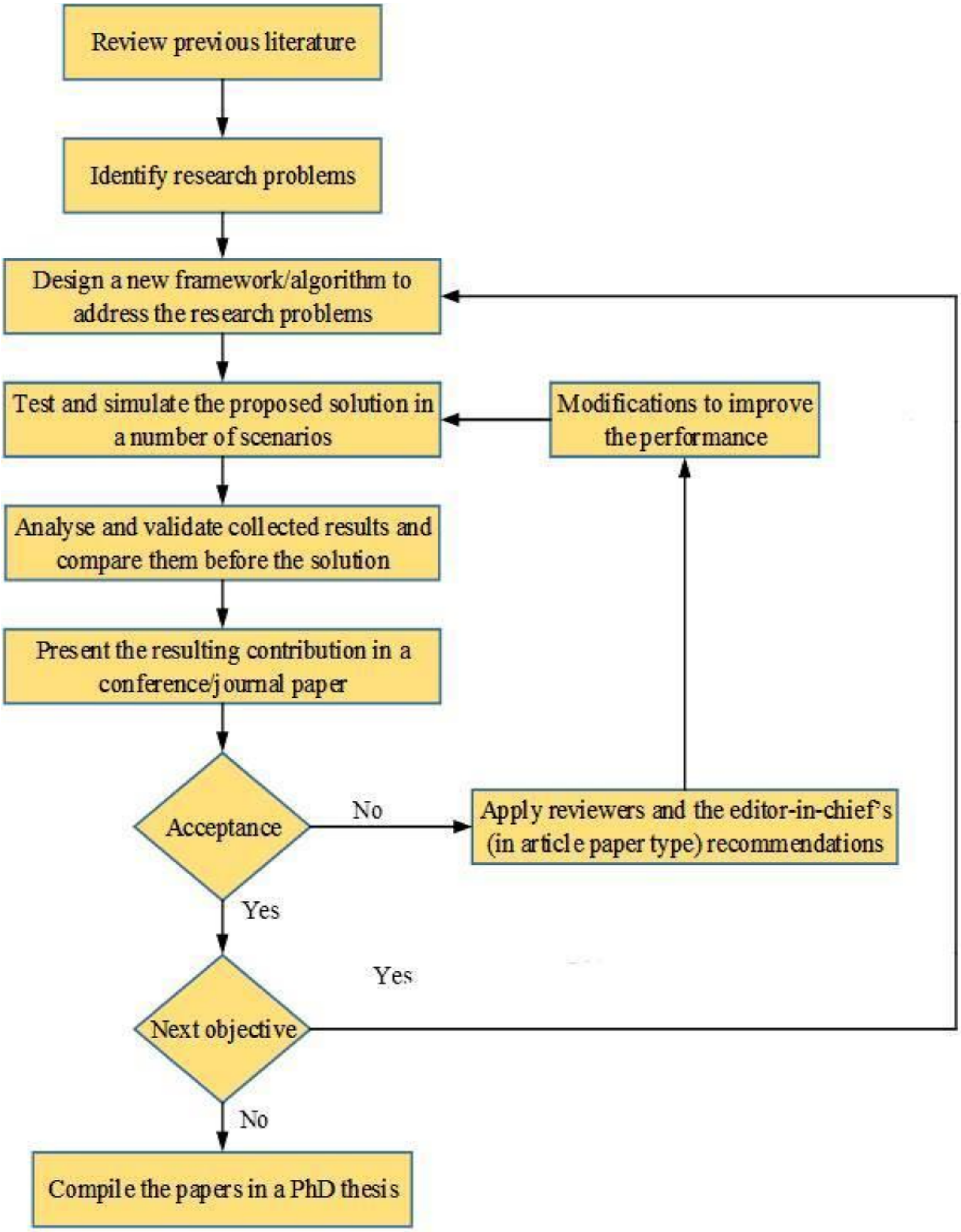

Review previous literature
Identify research problems
Design a new framework/algorithm to address the research problems
Test and simulate the proposed solution in a number of scenarios
Modifications to improve the performance
Analyse and validate collected results and compare them before the solution
Present the resulting contribution in a conference/journal paper
Acceptance
No
Apply reviewers and the editor-in-chief's (in article paper type) recommendations
Yes
Yes
Next objective
No
Compile the papers in a PhD thesis

Figure 1.2 Main phases of research methodology

1.8 Thesis Structure

Chapter 1: This chapter gives a general overview of the study.

Chapter 2: This chapter defines human speech production and defines the speaker recognition task, taxonomy of speaker recognition, the general framework of speaker recognition, and the main applications of speaker recognition. An extensive review of the state of the art of speaker recognition systems has been done by outlining the well-established audio features and" machine-learning techniques" that have been applied in a broad range of speaker recognition areas. A statistical model is introduced, particularly discussing the Gaussian Mixture Model structure and its training through the Expectation Maximization (EM) algorithm. Specific modelling techniques for recognition, such as Universal Background Model, and techniques for evaluating the speaker recognition performance are described. In addition, this chapter provides a general review of reverberation phenomenon and the effect of reverberation on speech intelligibility. Finally, it discusses the main parameters of the indoor acoustic environment and the previous work of speaker recognition in reverberation conditions.

Chapter 3: This chapter provides a review of the anechoic chamber at Salford University, which is used to record the speech data set employed in this work. In addition, these chapter reviews are provided using the simulation methods for impulse response such as the image source method and CATT-acoustic software, and some impulse response databases such as the Aachen Impulse Response (AIR) database.

Chapter 4: In this section, experiments are piloted to evaluate the MSR toolbox with different reverberation times. Additionally, the most common features used in the speaker recognition field were investigated. The effect of two acoustic parameters (reverberation time RT and the distance between the source and receiver) was also investigated and examined in this chapter.

Chapter 5: The primary aim of this chapter is to investigate and improve the robustness of speaker recognition in reverberant environments via training. Therefore, an experiment was conducted to improve the robustness of speaker recognition using speech contaminated with a reverberation time from the same or different rooms.

Chapter 6: In this section, a detailed introduction of the MLE based RT estimation method is provided, and the optimisation method used in this work is described. Furthermore, In this chapter, the “maximum likelihood estimation method” was used to estimate the reverberation time of the speech signal. Training on the fly approach was then used to select the closest set from the training sets, depending on the Euclidean distance between the estimated reverberation time and the reverberation time in the training set. Furthermore, in this chapter, the early reflections sound was estimated and convoluted with an anechoic signal to train the system to improve the performance of a speaker verification system.

Finally, the “estimated reverberation time” and “early to late ratio” was used with training on the fly schema to improve the robustness of a speaker verification system.

Chapter 7: This chapter provides a summary, concluding remarks and associated future work for this research.

1.9 Publications Resulting from Research

The work in this thesis has led to the following publications.

1.9.1 Refereed Journal and Conference Papers

1. K. Alkarawi, A. Alnoori, and F. Li,’’Automatic Speaker Recognition System in Adverse Conditions-Implication of Noise and Reverberation on System Performance’’. International Journal of Information and Electronics Engineering, Vol.5, No.6, November 2015

2. K. Alkarawi, A. Alnoori, and F. Li, ‘Automatic Speaker Recognition System in Adverse Conditions-Implication of Noise and Reverberation on System Performance’ The 7th International Conference on Computer Engineering and Technology (ICCET 2015**.**
3. A. Alnoori, K. Alkarawi and F. Li ‘Improve Robustness of Speaker Recognition in Noisy and Reverberant Conditions via Training’ IEEE- European Intelligence and Security Informatics Conference (EISIC 2015)
4. K. Alkarawi, and F. Li ‘Robust speaker verification in reverberant conditions using estimated acoustic parameters-A maximum likelihood estimation and training on the fly approach’ IEEE-the seventh international conference on Innovative Computing Technology (INTECH 2017).
5. K. Alkarawi ‘Autocorrelation Detection for Early Reflection to Improve Robustness of Speaker Verification in Reverberant Conditions’ IEEE- International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC 2018).

1.9.2 Posters at Conference in Proceedings

1. A poster on the implementation of the MSR toolbox as a speaker recognition system, *Salford Postgraduate Annual Research Conference 2015 (SPARC 2015).*

1.9.3 Abstracts in Conference Proceedings

2. K. Alkarawi and F. Li ‘Evaluate the performance of a speaker recognition system in reverberation condition’ *Salford Postgraduate Annual Research Conference 2015 (SPARC* June *2015)*

3. K. Alkarawi and F. Li ‘Evaluate the effect of reverberation time and source to receiver distance on the performance of speaker recognition’ in Proceedings of the CSE 2016 Annual PGR Symposium 2016 (CSE-PGSym16), April 2016.
4. K. Alkarawi, and F. Li ‘Evaluated the robustness of MFCC, and GFCC features in reverberation environment’ *Salford Postgraduate Annual Research Conference 2016 (SPARC 2016)*
5. K. Alkarawi, and F. Li ‘Speaker recognition in reverberation environments using multi-condition training’ *Salford Postgraduate Annual Research Conference 2017 (SPARC 2017),* June 2017

CHAPTER TWO: BASICS AND BACKGROUND OF SPEAKER RECOGNITION

In the present chapter, the current state of the art in the field of speaker recognition is highlighted. There are many papers related to this area. An exhaustive review of all the related articles is not the intention, only some major milestones and those upon which subsequent work in the thesis is built are listed. The existing approaches concerned with the robustness of speaker recognition are reviewed. These methods comprise, feature extraction techniques, speaker-modelling techniques, classification, decision-making strategies and techniques of evaluating the speaker recognition performance.

2.1 Human Speech Production

Front-end processing, the first component in the basic structure of speaker recognition system is a key element of the recognition process. The main task of front-end processing in speaker recognition system is to find the relevant information from speech, which could represent speaker's voice characters and help achieve good classification results. However, in order to get desired features for speaker recognition task, it is crucial to understand the mechanism of speech production, the properties of human speech production model, and the articulators, which have speaker-dependent characters. Figure 2.1 depicts a sagittal section of the human speech production system. The description of the speech production system is summarised from Quatieri (2002) and Juang and Rabiner (1993). The main parts of the system are the lungs, larynx (organ of speech production), the pharyngeal cavity (throat), oral cavity (mouth), and nasal cavity (nose) (Fant, 1971; Y. Lu, 2010; Rabiner & Juang, 1993). The pharyngeal and oral cavities are usually grouped into one unit referred to as the vocal tract, and the nasal cavity is often called the nasal tract. The vocal track begins at the output of the larynx (vocal cords, or glottis) and terminates at the input to the lips, which forms a resonant space shaped by various articulators such as the tongue, jaw, lips, soft palate and teeth. The nasal tract begins at the velum and ends at the nostrils. When the velum (a trapdoor-like mechanism at the back of the oral cavity) is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech

2.1.1 Lungs

The work of the lungs is similar to a power supply that pushes the air to the larynx. The lungs are utilised for the vital function of inhalation and exhalation of air. In the speech production model, they are the power source that supplies energy to the remaining blocks in the system. Inhalation is realised by decreasing the lung air pressure. This is possible thanks to the rib cage and the diaphragm. The rib cage is expanded through this process. The diaphragm, which is placed underneath the lungs, is lowered, so the lungs are expanded. This pressure lowering causes air to rush in through the vocal tract and down the trachea into the lungs. Exhalation is opposite to inhalation. It is caused by an air pressure increase in the lungs. The volume of the chest cavity is reduced by contracting the muscles in the rib cage and lifting the diaphragm. That produces airflow from the lungs to the larynx through the trachea. Inhalation and exhalation always rhythmically follow each other when breathing. However, during speaking short spurts of air are taken and then released steadily by controlling the muscles around the rib cage. Thus, the rhythmic breathing is overridden, since expiration takes one sentence or phrase time. During this time, the air pressure remains almost constantly above atmospheric pressure. However, as will be seen later, the time-varying properties of the larynx and the vocal tract cause this constant pressure to become time varying. This airflow produced by the lungs has the shape of white Gaussian noise. The only speaker-dependent information that is introduced by the lungs is the energy of this noise. However, this is not discriminating enough, and other features have to be found.

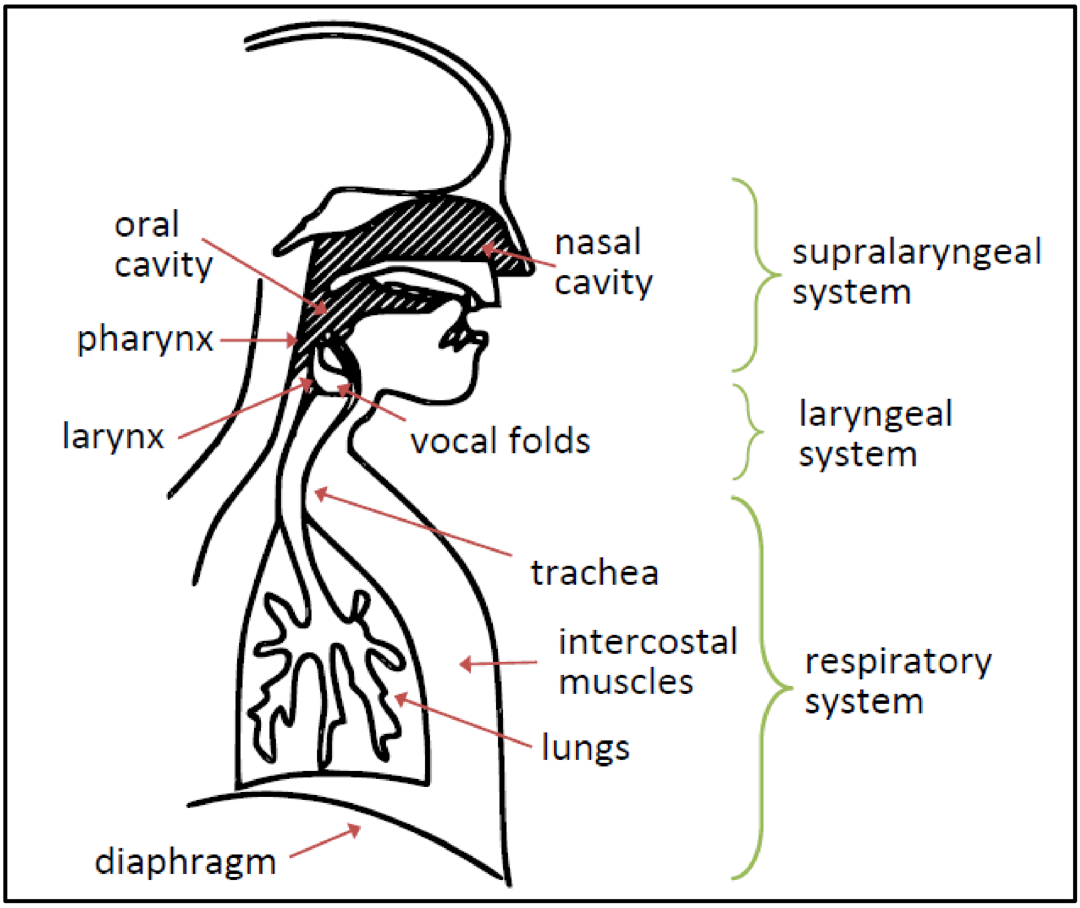


Figure 2.1 Human speech production (Kelly, 2014)

2.1.2 Larynx

The larynx also called the "voice box" is a complicated system of cartilage, muscles, and ligaments whose primary purpose, in the context of speech production, is to control the vocal cords or voicing. It has various other functions such as closing the entrance to the lower respiratory system during swallowing. Since these functions are not relevant to the speech production models, they will not be analysed in this section. From the voice production point of view, the most important parts of the larynx are the vocal folds and the glottis (Flanagan, 1972). The vocal folds are twin masses of flesh, ligament and muscle, which stretch between the front and the back of the larynx. Their size varies from one person to another; on average, they are around 15 mm long in men and 13 mm long in women. They can remain open to creating unvoiced sounds, or they can vibrate to produce voiced sounds during the speech. During breathing, they stay open, allowing the air to flow into the lungs. Figure 2.2 shows on the left, a full schematic of the larynx and the right, a view of the glottis. Figure 2.3 sketches a downward-looking view of the human larynx: (a) voicing; (b) breathing

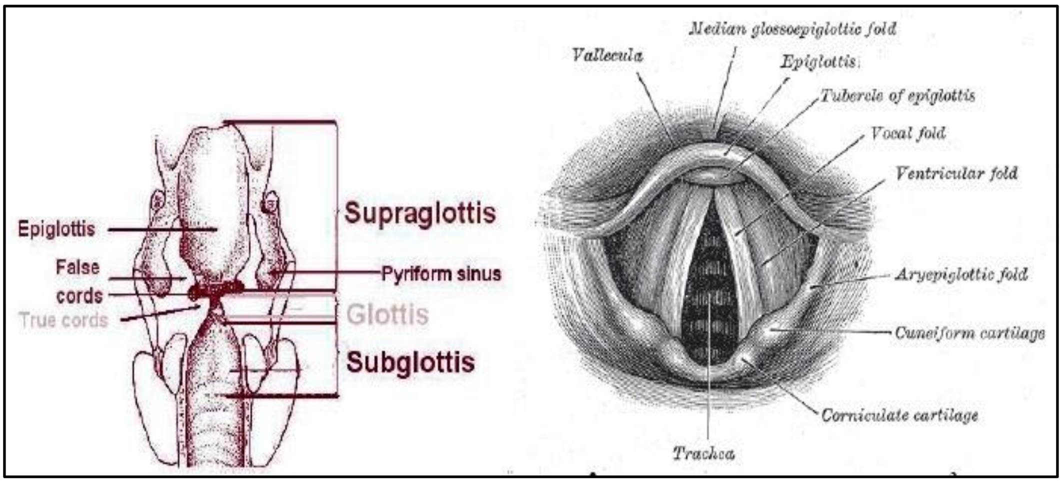


Figure 2.2 on the left, full schematic of the larynx, on the right, view of the glottis(Beigi, 2011)

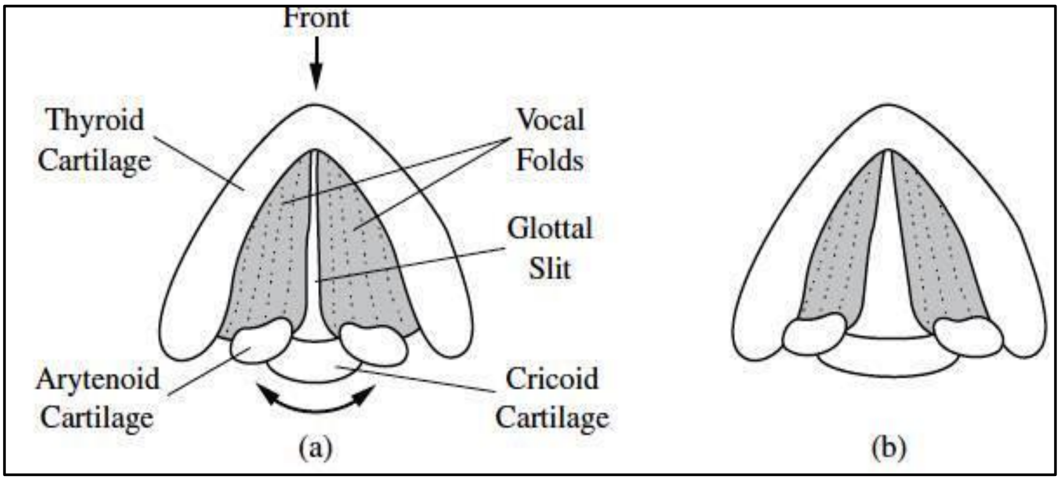


Figure 2.3 sketches of downward-looking view of the human larynx: (a) voicing; (b) breathing(K. N. Stevens, 2000)

2.2 Speaker Recognition Task

Speaker recognition methods, together with facial image recognition, fingerprints and retina scan recognition, denote some of the primary biometric tools for identification of a person. Research has considered automatic computer-based speaker recognition since early 1970, taking benefit of advances in the fields related to speech recognition. This section defines the main parts of speaker recognition, describes the general framework of speaker recognition, and summarises its applications.

2.2.1 Taxonomy of Speaker Recognition Systems

Speaker recognition systems can be categorised into three different categories. Firstly, they can be classified as speaker identification or speaker verification. Secondly, they can be text-dependent or text-independent systems. Lastly, it is possible to sort them into an open-set or closed-set system (D. A. Reynolds & Heck, 2000). In this section, all these classifications will be described.

In a speaker verification system, the speaker will first provide his identity, and then the system will check whether this identity is correct by analysing the speaker's voice. For instance, in the case of automated teller machines (ATM), the user will insert a credit card. The credit card represents the identity of the user. If the ATM includes an automatic speaker recognition (ASR) module, it can check whether the card is being utilised by its legitimate holder or by an impostor by asking the user to input some speech. In this case, notice that since the user provides his identity to the system, only a yes/no decision has to be made. The ASR only has to compare the voice input to the model attached to the identity provided by the user. That means that the ASR will perform a single comparison and then a single decision based on the result of that comparison. However, in a speaker identification system, the speaker will not provide his identity to the system. Instead, the speaker will input his speech, and the system will decide which speaker model best matches the speech input. In this case, the system has to make *N* comparisons; *N* represents the number of speakers in the system database. Each comparison will produce a likelihood score so that the system can select the identity attached to the most likely speaker model. That means this kind of decision will be “Speaker i” with i=1...N. Figure 2.4 gives a practical implementation of the systems.

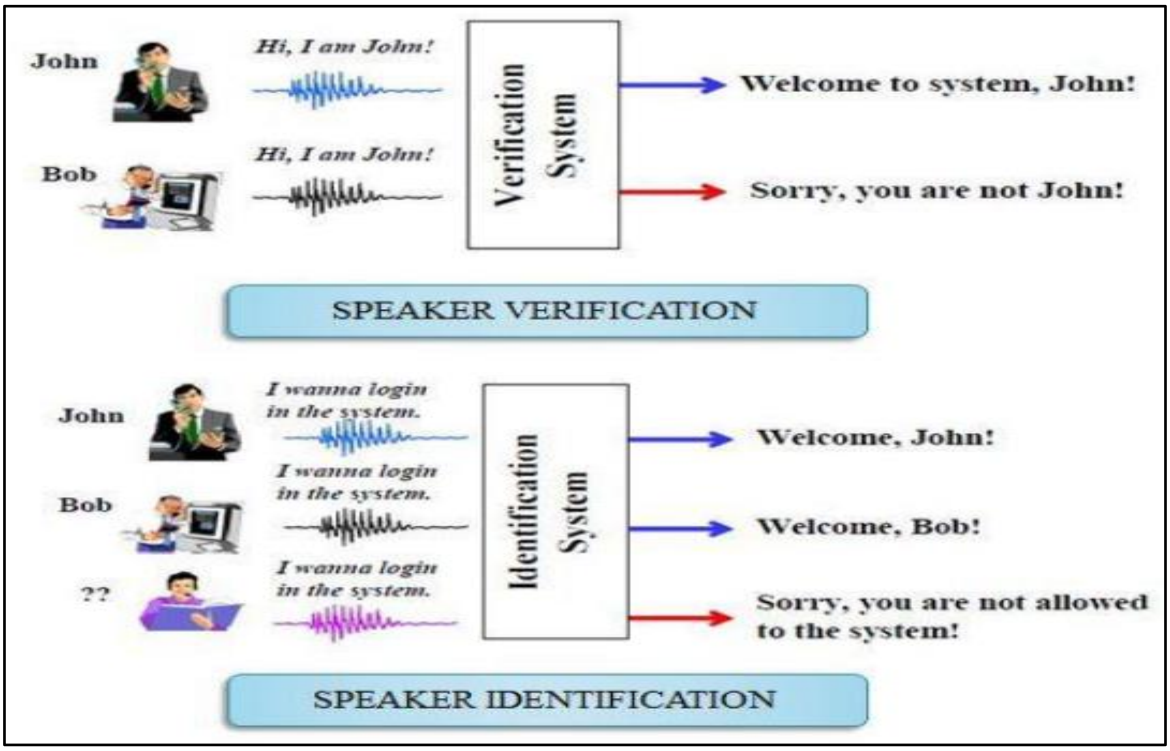


Figure 2.4 Practical examples of Identification and Verification Systems (C-DAC)

Speaker recognition may be classified into closed-set or open-set recognition, and this is dependent on whether the recognition task assumes the possibility that the speaker being identified may not be on the list of potential applicants (Memon, 2010). A closed-set recognition will assume that the speaker who is attempting to enter the system belongs to a set of known speakers. In this case, if a speaker identification system is being utilised, the decision taken by the system will simply be to match the most likely speaker to the speech input. However, an open-set system will consider the possibility that the person who is trying to access the system may be unknown. That means that there is no model related to this person and therefore the decision taken by the system is that the person is unidentified or an impostor. There are three main approaches to dealing with this issue; the first one comprises a model for unknown speakers acquired from an extensive voice database containing several speakers. “The second is setting a likelihood threshold so that if any of the speaker model scores cross this threshold, the decision will be that this is an "unidentified speaker". The last option is a combination of both approaches. It is clear that open-set systems are more complex and better match conditions in real life, where a listener does not have to be familiar with all the voices he hears every day (D. A. Reynolds & Heck, 2000). That is the reason why this thesis will consider this case".

A further possible classification method for ASRs uses a constrained sentence or text for recognition. In a text-dependent case, a given sentence or word (most commonly known as a 'password'), will be utilised for training and recognition. In this scenario, a two-level security system is acquired. Firstly, the voice has to be produced by an authorised speaker and secondly, the user has to provide the proper password. It is usual for this mode of operation to utilise a user PIN number or password. Another type of system can consist of the text-dependent recognition of a group of texts. These are text-prompted systems. In this scenario, the system will ask the speaker to pronounce a given code (e.g. a sequence of numbers) to avoid a phishing attack. In this scenario, an attacker would need voice registries of the authorised speaker, containing all the possible combinations the system may ask for. The larger the number of keywords used, the more robust a system can be built. On the other hand, text-independent systems are not sensitive to the message contained in the speech input. They only attend to speaker-dependent features of speech and do not depend on the sequence of words spoken by the user. The recognition and training will be based on any speech utterance produced by the user (D. A. Reynolds & Heck, 2000). Although text-dependent recognition, to date, delivers better performance (D. A. Reynolds, Quatieri, & Dunn, 2000), text-independent recognition is a more attractive technology due to its comprehensive scope of possible applications. Due to the unobtrusive nature of text-independent recognition, it can be integrated into many applications without imposing any additional requirements on a user.

2.2.2 General Framework of the Speaker Recognition System

A traditional speaker recognition system involves two main stages: the enrolment or training stage and the verification or testing stage. These stages commonly include three processes: feature extraction, speaker modelling, and decision making (J. P. Campbell, 1997; Furui, 2009). Figure 2.5 represents the essential elements in the enrolment stage. Through the enrolment (or training) stage, speech samples from known speakers are utilised to compute the vectors of parameters called the typical features (J. P. Campbell, 1997; Sethuraman & Gowdy, 1989). This stage aims to obtain a registry of the voice properties of a given speaker. This registry is called a model. At this stage, the speaker's identity is well known by means other than speaker recognition. Therefore, once the speaker's identity has been proven, he is asked to read out a given text or simply produce a few seconds of speech. A microphone then captures this speech signal. As shown in Figure 2.5, the output of the microphone is entered into a pre-processing block. This block performs the digitisation of the speech signal, splits the signal into smaller frames and prepares these frames for the next step. The second step is called "feature extraction". Feature extraction is a process used to transform a set of speech signals into a form that the pattern classification engine can understand. In this step, the dimensionality of the speech frames is reduced considerably. This operation is necessary for two reasons; firstly, the pattern-matching block needs to operate with low-dimensional vectors to work in real-time mode. Secondly, the feature extraction block removes unnecessary information, which is being carried in the speech frames and emphasises speaker dependent aspects of speech. Usually, extracted speaker features are short-time spectral/cepstral features such as short-time Fourier transform spectral features and mel-frequency cepstral coefficients (MFCC), or long-term features like prosodic features (Shriberg, 2007; Shriberg et al., 2005).

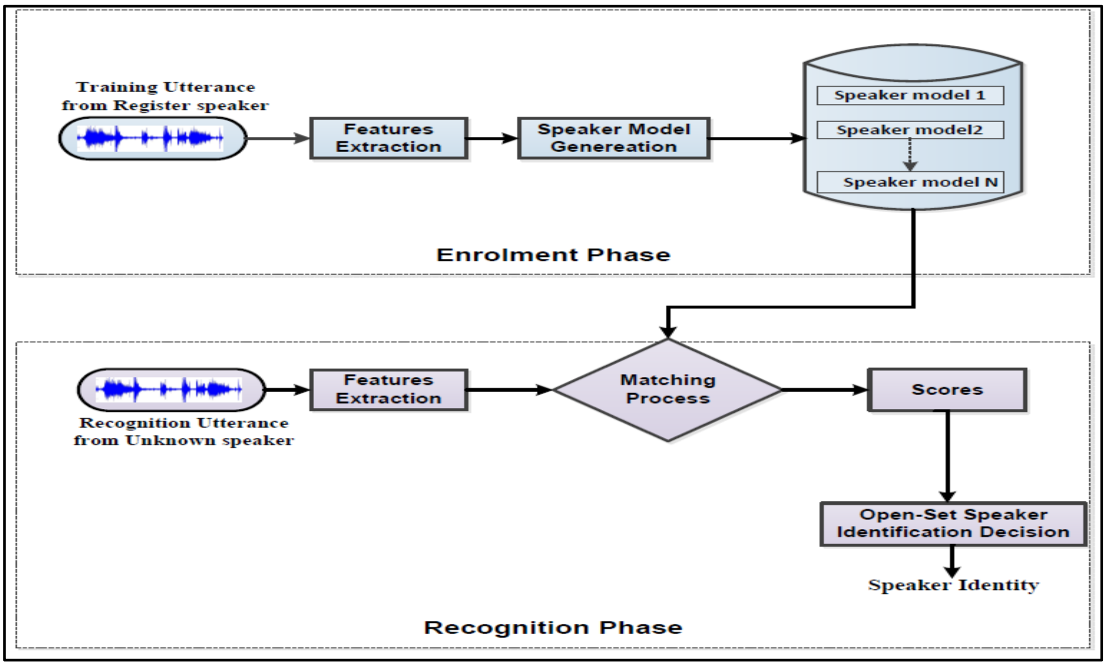


Figure 2.5 Basic structure of Speaker Identification

The third block is the pattern-matching algorithm. This algorithm can work in two different modes, training and testing. In the training mode, sequences of feature vectors produced by a known speaker are used to obtain accurate models of that speaker's voice. In the testing mode, the pattern-matching algorithm to obtain a similarity measure with feature vectors produced by the a priori unknown speaker uses the models previously created in the enrolment stage. The pattern-matching algorithm is also an essential block in speaker recognition systems. In fact, it is the "brain" of a speaker recognition system, and a good choice will produce an excellent result in the final performance of the system. For that reason, the enrolment process is usually achieved offline and repeated if the models are no longer valid. For speaker modelling, GMM is commonly used to model speaker feature distributions. In speaker identification, recognition decisions are usually made based on likelihoods of observing data-given applicant speaker models. The testing stage blocks shown in Figure 2.6 are very similar to the enrolment ones. The two main differences are that, firstly, the system is trying to identify a speaker, so the speaker who is producing speech is a priori unknown. The other difference can be seen in the pattern-matching algorithm, where during the testing (or recognition) phase, the speaker recognition system is exposed to speech data not seen through the training phase (J. P. Campbell, 1997; Naik, 1990). Speech samples from an unknown speaker or a claimant are used to calculate feature vectors using the same methodology as in the enrolment process. This algorithm will compare the input feature vectors to the speaker models stored in a database. Finally, the likelihood score for each model is analysed by a decision-making block. According to these scores, the decision-making algorithm will estimate who is the most likely speaker to have produced the speech signal (Furui, 1997; Higgins, Bahler, & Porter, 1991; T Kinnunen, 2005; K.-P. Li & Porter, 1988; D. A. Reynolds et al., 2000; Sivakumaran, Fortuna, & Ariyaeeinia, 2003). Furthermore, in the case of an open set algorithm, an “impostor” or “unknown speaker” decision can be made. In fact, the enrolment stage is completely the same, but a number of things change in the testing stage. Figure 2.6 shows the basic structure of Speaker Verification.

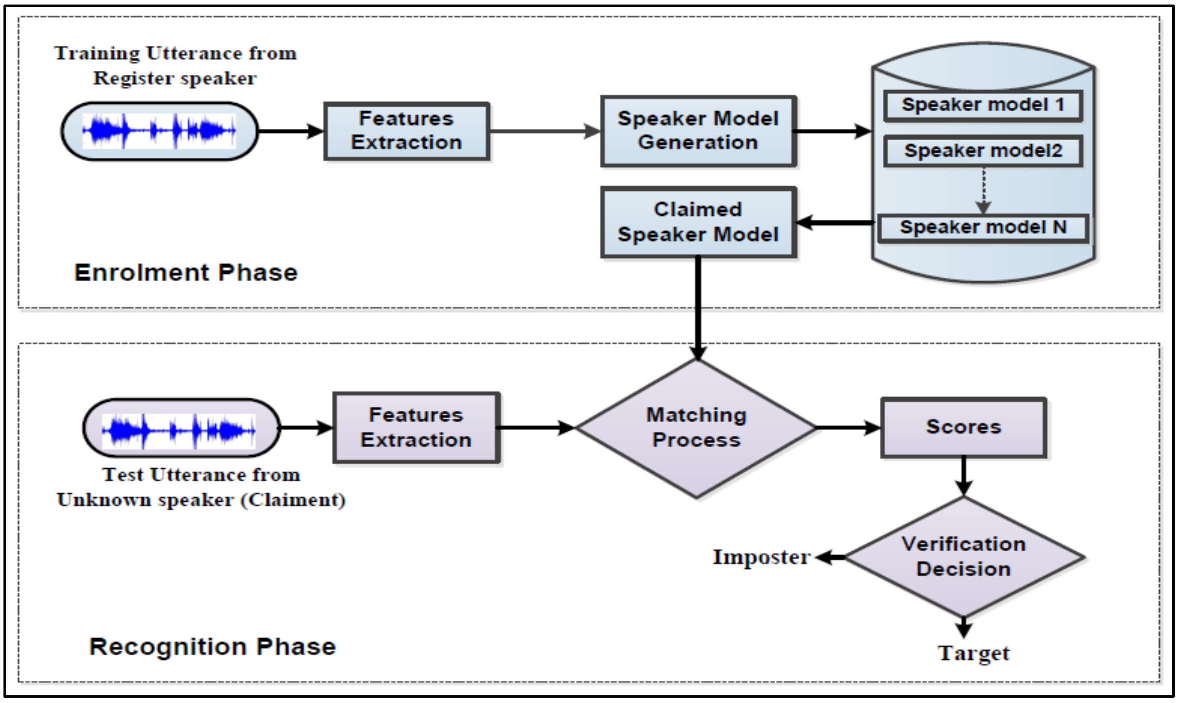


Figure 2.6 Basic structure of Speaker Verification

In the above figure, First, it is only necessary to compare the input feature vectors with the model for the identity claimed by the speaker. Secondly, the decision is either “Yes” or “No”. The testing phase is usually relatively fast and can be done online in real-time conditions.

2.2.3 Speaker Recognition Applications

From a commercial viewpoint, many markets may arise for speaker recognition or verification technologies; even limiting the discussion to speaker recognition only, some applications can be conceived. A list of possible applications is given (Beigi, 2011):

1) **Surveillance applications:** Not so widely recognised, but may significantly improve the video-only systems used today; an unknown voice detected in a critical area may prompt the cameras to record the face of an intruder.
2) **Forensic applications:** Speaker recognition can be applied to an audio signal to investigate activities by the authorities, in which an automatic and fast system is asked to skim through a recording to find any given suspect’s voice (Becker, Jessen, & Grigoras, 2008).
3) **Authentication:** In automatic answering systems, a typical example is a home banking application or other customer care service. In these cases, the recognition is usually text-dependent and asks the customer to say, for instance, a personal identification number (PIN).
4) **Security and access applications:** Instead of entering passwords or answering some security questions, one can use his/her voice as a key, and a built-in speaker recognition system automatically identifies the speaker or verifies that the speaker is the claimed one. Speaker recognition systems can offer transaction verification,

facilities such as computer access control, monitoring, telephone voice verification for long distance calling or banking access, etc. (Sokolov, 1997).

5) **Content Indexing:** In this situation, speaker recognition is utilised automatically to index a multimedia collection, such as broadcast news, audiobook archives, movies, etc. to ease searching of, and access to content.

6) **Games and entertainment:** These represent an expanding business area; the recognition of a speaker's voice could be useful in such applications, adding a new way to interact with the players.

2.3 Features Spaces used in Speaker Recognition

A fundamental part of any voice recognition system is thus the feature extraction stage, which is responsible for obtaining the most meaningful information about the speaker from the speech signal and ignoring unwanted information. More accurately, the extraction of features from a speech signal could be defined as a blend of two tasks: The first task is a detailed and efficient measurement of the speech signal, while the second task is the minimization of redundancies in the measures. Moreover, it is worth stressing the final aim of any feature extraction process is that it should best characterise underlying physical phenomena, through an optimal description of a signal produced by the phenomena itself. Historically, the following spectrum-related speech features have dominated the speech and speaker recognition areas: Real Cepstral Coefficients (RCC) introduced by (Oppenheim, 1969), pitch contours (Bishnu Saroop Atal, 1972), Linear Prediction Coefficients (LPC) proposed by (Bishnu S Atal & Hanauer, 1971), Linear Predictive Cepstral Coefficients (LPCC) derived by (Bishnu S Atal, 1974), and MFCC (Davis & Mermelstein, 1980). Other speech features such as Perceptual Linear Prediction (PLP) coefficients (Hermansky, 1990), Adaptive Component Weighting (ACW) cepstral coefficients (Assaleh & Mammone, 1994a, 1994b), and various wavelet-based features, although presenting reasonable solutions for the same tasks, did not gain widespread practical use. In light of the above discussion, any feature should possess the desirable properties listed below (T Kinnunen, 2005; Rose, 2003; Wolf, 1972).

- Efficient representation of speaker-specific information (i.e., small within-speaker variability and large between-speaker variability)
- Easy to compute
- Stable over time
- Occur naturally and frequently in speech
- Robust to environmental distortions
- Variations in voice caused by speaker’s health or ageing should not degrade the performance of the feature extraction method.
- Difficult to imitate or mimic using the speech of imposters
- Difficult to duplicate using the speech of imposters.

In practice, it is unlikely that a single feature would fulfil this full list of requirements. As noted by Kinnunen (Tomi Kinnunen & Li, 2010), there is no globally best feature, and a trade-off must be made between speaker discrimination, robustness and practicality. However, the complexity of the speech signal can be leveraged by extracting multiple complimentary features and combining them to improve the speaker discrimination of the system (D. Reynolds et al., 2003). There have been many features proposed for speaker recognition, broadly categorised based on the duration of speech required for their extraction and on the level of the information they capture. Typically, the magnitude spectrum is wrapped into perceptually motivated scales, such as mel scale, bark scale, and equivalent rectangular bandwidth (ERB) scale. Additional Fourier analysis is then taken to derive Cepstral features. Speaker features such as modulation spectral features (Falk &

Chan, 2010) and features that incorporate phase information (L. Wang & Nakagawa, 2009) have shown robustness against reverberation. “The cepstral analysis (Assaleh, 1995; Assaleh & Mammone, 1994a; Furui, 1981; Sethuraman & Gowdy, 1989; Zilovic, Ramachandran, & Mammone, 1995) and the Mel Frequency Cepstral Coefficients (MFCC) (Davis & Mermelstein, 1980; Murty & Yegnanarayana, 2006; Nakasone, 2003; K. N. Stevens, 1971), are the most common short-time feature extraction approaches, and the linear predictive Cepstral coefficients (LPCCs) (Huang, Acero, Hon, & Foreword By-Reddy, 2001), and perceptual linear prediction (PLP) coefficients (Hermansky, 1990), are examples of frequently-used short-term spectral features”.

On the other hand, temporal features capture dynamics ignored by short-time features. First order and second order delta features fall into this category. Additional instances are modulation spectral features which represent frequency information about sub-band signal envelopes (Atlas & Shamma, 2003; Falk & Chan, 2010), which correlate with speaking rates. Prosodic features capture high-level speech information such as rhythm, stress, and intonation (Shriberg, 2007; Shriberg et al., 2005): although not as discriminating as short time features, they provide additional information. Thus, these features relate to a mix of physical characteristics, e.g., gender and age, and learned or acquired ‘behavioural’ factors, like speaking style or health issues. High-level features, extracted over a duration ranging from seconds to minutes, capture exclusively behavioural attributes, including phonetic-level information, such as accent, and word-level information such as semantics and idiolect, an individual’s lexicon (D. Reynolds et al., 2003; Shriberg, 2007). Short-term features are an advantage from the perspective that they are easy to extract and do not demand a minimum speech duration requirement for the system. However, they are affected by noise and other sources of mismatch between enrolment and recognition conditions in a significant way. Longer-term features, while requiring a lot more speech, and probably being more computationally expensive to extract, are generally robust to noise and mismatch (Tomi Kinnunen & Li, 2010).

“In this thesis, short-term spectral features, Mel frequency cepstral coefficients (MFCC) and the gammatone frequency cepstral coefficients (GFCC) specifically, have been used exclusively”. The various phases of their calculation have similarities with those of MFCC (Qi, Wang, Xu, & Tejedor, 2013). Similarly, the MFCC feature vectors in this method are computed from the spectra of a sequence of windowed speech frames. The sensitivity to additive noise and reverberation is one of the major disadvantages of MFCC. Therefore, recently, a deep study was conducted (Zhao & Wang, 2013) which confirmed the intrinsic noise robustness of GFCC relative to MFCC. In addition, this study includes deep details about high points that make GFCC more robust to additive noise compared with MFCC, by carefully examining all differences between two features using the speaker identification system. This study shows that the cubic root rectification presents more robustness to features than log because the cubic root operation makes features scale variant (energy level independent) which helps to maintain this information, while in the log operation in MFCC features do not encode this information. More details about the calculation process and the properties of the MFCC and GFCC can be found in chapter 4.

2.4 Machine Learning and Speaker Modelling

Machine Learning refers to an artificial process that optimises a feature extraction stage to partition the data into relevant classes. There are two primary methods of classification, namely: unsupervised classification (clustering); and supervised classification (discrimination). These two have been applied to a diverse range of work including physics, mathematics, statistics, engineering, artificial intelligence, computer science, and the social sciences; see (Webb & Copsey, 2011) for more information. An important and growing body of literature has investigated various machine-learning techniques in the field of audio content analysis. Speaker models are built from speaker features. The objective of modelling technique is to generate speaker models using speaker-specific feature vectors. Modern classifiers used in speaker recognition technology include Gaussian Mixture Models (GMM) (D. Reynolds & Rose, 1995), Hidden Markov Models (HMM) (Yuk, 1996), Support Vector Machines (SVM) (W. M. Campbell, Campbell, Reynolds, Singer, & Torres-Carrasquillo, 2006). Vector Quantization (VQ) (Soong, Rosenberg, Rabiner, & Juang, 1985), and Artificial Neural Networks (ANN) (C. Wang, Xu, & Principe, 1997). The GMM is currently recognised as the state of art and dominant modelling and classification technique for speaker recognition (D. A. Reynolds, 1995). The GMM models the Probability Density Function (PDF) of a feature set as a weighted sum of multivariate Gaussian PDFs. On the enrolment stage of speaker recognition, a model is trained from a set of feature vectors. In text-independent speaker recognition, there is no correspondence between the speech content of the enrolment and recognition utterances. Therefore, the model must be general enough to describe the typical feature space of a speaker but discriminating enough to distinguish between the feature areas of different speakers. “In this thesis, GMM-UBM was used in the experiments relating to speaker recognition”.

2.4.1 Gaussian Mixture Model (GMM)

The Gaussian Mixture Model (GMM) method (Finan, Sapeluk, & Damper, 1996) is commonly regarded as the state-of-art modelling and classification technique successfully applied in many pattern recognition problems including speech recognition and speaker identification, image coding and many others. In a diversity of practical applications, a family of finite mixture densities can approximate the distribution of the parameters where the density function is a weighted sum of component densities. The component densities are usually modelled as Gaussians. It can be shown that any continuous probability density function can be approximated arbitrarily closely by a Gaussian mixture density (D. A. Reynolds et al., 2000). In its classical form (D. Reynolds & Rose, 1995), the GMM apply the expectation maximisation algorithm (EM), which iteratively updates the means, covariance's and weights for each class, and converges to a set of parameter vectors, providing the maximum value of the expectation function. Each set consisting of means, variances and weights constitute a class model. The resulting models provide multivariate probability density functions for each class with the highest expected values for giving training data.

The state-of-the-art systems discussed in this thesis use the classic paradigm of Maximum a Posteriori (MAP) adapted Gaussian Mixture Model (GMM) presented by Reynolds (D. A. Reynolds et al., 2000). For speaker identification, features of a test speech signal are matched with GMMs of all the enrolled speakers. The speaker with the highest score is chosen as the output. For speaker verification, each test utterance has a claimed speaker. Two theories represent whether the speech from the claimed speaker or not. A theory test has been shown to derive a likelihood ratio utilising the GMM of the claimed speaker and the GMM representing everyone else, usually a universal background model (UBM) (D. Reynolds, 2002; D. A. Reynolds, 1995). The likelihood ratio is then normalised and compared with a threshold to either accept or reject the original claim. In earlier studies, the GMM of a speaker is trained directly from his/her training data utilising the expectation-maximisation (EM) algorithm. Later this training approach is changed by adapting a pre-trained UBM. The GMM-UBM framework is proven the better option. A Gaussian Mixture Model (GMM) is a weighted sum of M multivariate Gaussian components, allowing it to model an arbitrary distribution of observations. The likelihood of an observation x given a GMM denoted by λ is:

$$p(x|\lambda) = \sum_{m=1}^{M} w_m p_m(x) \qquad 2.1$$

where x is a D-dimensional vector, w_m is the weight of the m^{th} Gaussian component $p_m(x)$

$$p_m(x) = \frac{1}{(2\pi)^{D/2}\left|\sum_m\right|^{1/2}} \exp(-\frac{1}{2}(x \; - \mu_m)^T \Sigma_m^{-1}(x \; - \mu_m) \qquad 2.2$$

w_m and $\sum_m$ are the mean vector and covariance matrix of the mth component respectively. The component weights w_m > 0 must satisfy $\sum_{m=1}^{M}=1 \;\; w_m = 1$. In practice, for reasons of data requirement and computation, covariance matrices are usually diagonal (Tomi Kinnunen & Li, 2010). Training a GMM involves finding the parameters λ = { w_m, w_m , $\sum_m$ } $_{m=1}^{M}$ given a training sample $x = \{x1, x2,...xT\}$. The log-likelihood LL of x concerning λ is given in:

$$LL = \frac{1}{T} \sum_{t=1}^{T} \log \; p(x_t|\lambda) \qquad 2.3$$

The higher the value of *LL*, the stronger the indication that x originates from the GMM λ . Maximum likelihood (ML) estimation is an approach to maximise the likelihood of the model concerning the given data and can be achieved with the iterative Expectation-Maximisation (EM) algorithm (Bilmes, 1998). The GMM has several properties that motivate its use for representing a speaker:

- One of the powerful properties of the GMM is its ability to form smooth approximations to arbitrarily shaped density. The GMM can be viewed as a parametric pdf based on a linear combination of Gaussian basis functions capable of representing a large class of arbitrary densities.

- GMM can be considered as an implicit realisation of probabilistic modelling of speaker dependent acoustic classes with each Gaussian component corresponding to a broad acoustic class such as vowels, nasals, etc. GMM-based speaker identification algorithms are popular due to their good performance (W. M. Campbell et al., 2006). Recently, state-of-the-art speaker verification systems extract sub-vectors using the UBM and further obtain low-dimensional speaker factor or i-vector to deal with channel variation (Dehak et al., 2011; Kenny, 2005a).

2.4.1.1 Universal Background Modelling (UBM)

Universal background modelling is a single speaker-independent model that is used by all speakers. Furthermore, due to the smaller storage space required, it commonly provides better performance. The UBM presented by Reynolds (D. Reynolds & Rose, 1995; D. A. Reynolds et al., 2000) was utilised where no adequate training data were available for GMM training. The speech used in the training of a UBM is not used for the training of the individual speaker models. In other words, the speech involved in the making of UBM does not include the utterances taken from the target speakers. A Gaussian mixture model (GMM) classifier was used with speaker recognition task for the first time by Reynolds (D. A. Reynolds et al., 2000); since then GMM has been extensively used in speaker modelling. GMM requires enough data to model the speaker well (D. A. Reynolds et al., 2000), to avoid this problem, Reynolds presented GMM-Universal Background Model (GMM-UBM) for speaker recognition tasks (D. A. Reynolds et al., 2000). The speaker-dependent model is then generated from the UBM by performing a maximum a posteriori (MAP) adaptation method utilising speaker-specific training speech. “As a result, the GMM-UBM gives better results than the GMM. The benefit of the UBM-based modelling procedure is that it provides good performance even though the speaker-dependent data is small”.

2.4.2 Training Models

This module is designed to read all the features for a given speaker and then train a statistical model for him/her. The features may be in only one file or be spread along multiple files; the module receives, for each speaker, a list of feature file names to be read. The module configuration indicates a sub function with the training strategy suitable for the features in use. Before the actual training, the features may be passed to a module termed Feature Transform, which applies one, or more transforms: Voice Activity Detection (VAD), Cepstral Mean Subtraction (CMS), Cepstral Variance Normalization (CVN), Feature Warping, and Amplitude Normalization.

2.4.3 Training UBM

The best performing Speaker Recognition is built on a likelihood ratio detector, which in turn needs a background model; this is often a Universal Background Model (UBM, Sec. 2.4.1.1). A module has been developed to train the UBMs since the features need to be collected with a specific strategy. Training UBM reads all the feature files from all the speakers and passes them to the optimized Expectation-Maximization routine (a UBM in this study can only be a GMM, so far).

2.4.4 Testing Models

The model's test is obtained supplying a speaker model and a background model with features, and computing their likelihood. This module thus reads a list of speaker models and loads models one at a time; also, the background model is loaded. For every speaker model, the database under test provides an input file list; all the input files have been previously processed by the Front-ends, so their features are available. Testing Models reads the features for each test, applies feature transforms if needed, and computes the likelihoods; the speaker models can be different from GMMs, and the module adopts the correct testing strategy accordingly.

2.4.5 Scoring

The output of the testing module is stored in a results file: for each speaker model and test file pair, a likelihood ratio is reported; this value should obviously be higher when the model and the test file are related to the same speaker (H0 hypothesis true) and lower in other cases (H1 hypothesis true). For each database task, a key file is provided, in which the mismatching between each model and test file is written. Using the key file, the values stored in the results file are thus grouped into two sets: true speakers for scores obtained when (H0) is true and impostor speakers for scores with (H1) true. The Scoring module then, after applying normalization, passes the two score sets to a subroutine, which plots the Detection Error Trade-off (DET) as explained in Sec. 2.5.2. The computation and storage of notable points in the diagram (EER percentage and DET) then complete the scoring process.

2.5 Speaker Recognition Evaluation Methods

“The majority of the work described in this thesis is concentrated on speaker verification tasks”. Speaker verification includes two kinds of possible errors: False acceptance error, also known as the false alarm probability, and false rejection error, also known as the miss probability (J. P. Campbell, 1997; Juang & Rabiner, 1993; Oglesby, 1995). A false acceptance (or false alarm) error happens when the system accepts a claim of identity from an impostor speaker, while the false rejection (or miss probability) error occurs when the system rejects an authentic speaker as an impostor. In the GMM-UBM framework, the decision is based on the score of a test feature vector for a given speaker model and the UBM. Considering the speaker verification scenario, the speaker model under comparison corresponds to the claimed identity of the speaker, and the decision is one of two possibilities: accept or reject. By counting these errors over a large number of trials, the false acceptance rate (FAR) and false rejection rate (FRR) of the system at a given decision threshold can be obtained. In a system evaluation, genuine speaker and imposter attempts are often referred to as a target and non-target trials respectively.

Setting a decision threshold is a task of reaching a desired trade-off between the FAR and FRR. By sweeping the decision threshold over all possible values, the full trade-off between the FAR and the FRR can be visualised as a detection error trade-off (DET) curve.

2.5.1 The Detection Cost Function (DCF)

The analysis and characterising of how well a speaker verification system has performed is possible by a probabilistic relation between the false acceptance and false rejection probability. A cost-based performance measure C_{Det} can be computed based on the false acceptance and the false rejection probabilities and used to assess the system performance. NIST speaker recognition, evaluation plan (NIST, 2001, 2002, 2004) defining the performance measure parameter C_{Det} as a weighted sum of the false acceptance and the false rejection error probabilities given as:

$$\begin{aligned} C_{Det} = C_{FalseRejection}\, P(\text{False Rejection} \mid \text{Target})\, P(\text{Target}) \; + \\ C_{False\,Accept}\, P(\text{False Acceptance} \mid \text{NonTarget})(1 - P(\text{Target})) \end{aligned} \qquad 2.4$$

Where *P(FalseRejection |Target)* is the probability that an actual target speaker was rejected, *P(FalseAcceptance | Non-Target)* is the probability that a non-target speaker was accepted. The parameters, $C_{False\ rejection}$ and $C_{False\ Acceptance,}$ are the costs (or weights) of the false rejection and false acceptance errors respectively, and *P(Target)* is the a priori probability of the specified target speaker. Table 2.1shows the values of $C_{FalseRejection}$, $C_{FalseAcceptance}$ and *P(Target)* suggested by the NIST speaker recognition evaluation comments for all speaker detection tests.

Table 2.1 Speaker Detection Cost Model Parameters

$C_{\text{FalseRejection}}$	$C_{\text{FalseAcceptance}}$	P(Target)
10	1	0.01

Like all performance analysis features the detection cost fuction (DCF) also has some downside factors, a major pitfall that separates the equal error rate (EER) and Detection Cost Function (DCF) are the insensitivity of the DCF to the changes in the performance of a system (Memon, 2010). The presumption while calculating with the equal error rate (EER) assumes both the boundaries of cost to be uniform which can be represented by the expression; CFalse Rejection= CFalseAcceptance=1. Since the decision in a speaker verification task is binary (accept or reject), a threshold of certainty may be comprised in the decision rule. A claim of identity is then accepted only when the decision can be made with a predetermined level of confidence. By varying this threshold one can change the ratio of false acceptance to false rejection errors (Oglesby, 1995).

2.5.2 Equal Error Rate (EER) and Detection Error Trade-off (DET)

The error rates for the speaker recognition system were initially measured using receiver operating characteristic (ROC) curves (J. P. Campbell, 1997). However in the more recent studies of the speaker recognition systems, are replaced the nonlinear ROC curves by detection error trade-off (DET) plots (Martin, Doddington, Kamm, Ordowski, & Przybocki, 1997). The Detection Error Trade-off (DET) is assumed more efficient in presenting the performance of a system, as it has a linear nature concerning the log system of coordinates compared to the ROC, which is nonlinear.

The Detection Error Trade-off (DET) displays the product of the percentages of the false acceptance and false rejection. The points on the Detection Error Trade-off (DET) curve correspond to different values of the acceptance threshold ξ. As illustrated in Figure 2.7, the false rejection probability is in inverse proportion to the false acceptance probability. Therefore, by reducing the false rejection probability the false acceptance probability will be increased and vice versa. Since the final aim of all speaker verification is to minimise both errors (false rejection and false acceptance), the best compromise can be realised when both errors are equal. The value of the percentage false rejection (or false acceptance) at the point when these two errors are equal is called the Equal Error Rate (EER). As demonstrated in Figure 2.7, a computation of the equal error rate is possible through a graph by observing the false acceptance percentage or the percentage of false rejection at the point of intersection of a 45° line and the curve of the DET. The determination of the quality of a speaker identification system is how small its value for an Equal Error Rate (EER) is: the smaller the value, the better its performance. Figure 2.8 shows the second example of a Detection Error Trade-off (DET) curve and Equal Error Rate (EER). “The work presented in this thesis belongs to the speaker verification task, and the EER and the DET have been implemented as the system performance measure in all cases”.

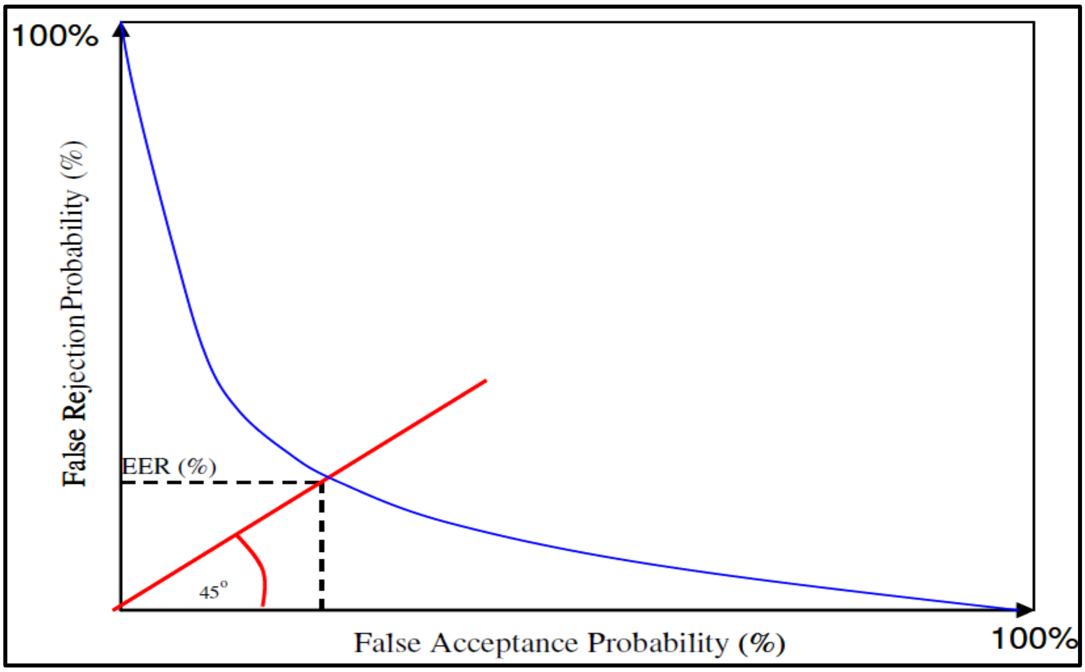


Figure 2.7 Detection Error Trade-off curve (blue) and the process of defining the Equal Error Rates (Memon, 2010)

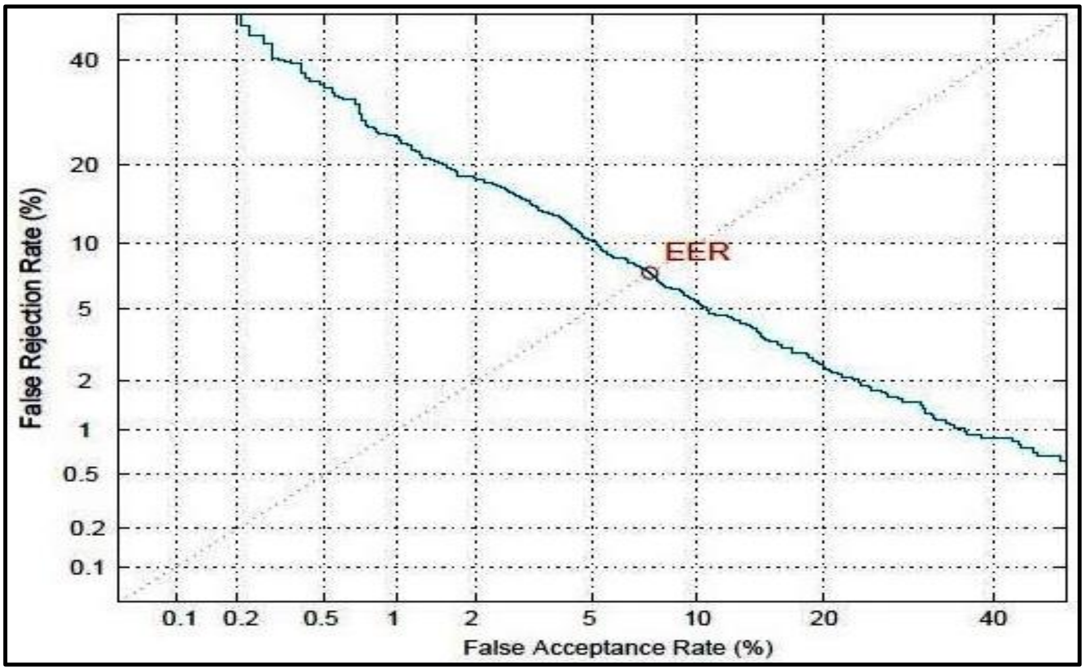


Figure 2.8 DET curves and the EER point

2.6 REVERBERATION

Reverberation is the name commonly given to the effect a room has on an acoustic signal produced by it. When speech or any other acoustic signal is produced in a room, it follows multiple paths from source to the receiver. Some portion of the signal energy that reaches the receiver is transmitted directly through the air, while the remainder is reflected off one or more surfaces in the room prior to reception. Usually, the earliest reflections arrive discretely, while later reflections arrive in rapid succession or concurrently as the number of paths, the sound may take increases. The reverberation process can be modelled as a convolution of the speech signal with a room impulse response. In speech communication systems, such as speech/speaker recognition systems, hands-free mobile telephones, and hearing aids, the received microphone signals are degraded by room reverberation, background noise, and other interference. This signal degradation may lead to the total unintelligibility of the speech and a decrease in the performance of automatic speaker recognition systems. This section presents some acoustic parameters, which are used to define the acoustic characteristics.

2.6.1 Sound in Enclosed Spaces

In real-world acoustic environments, the speech arriving at our ears includes not only direct sound but also its reflections from the surfaces such as walls, floors and ceiling, known as room reverberation. Essentially the reverberant sound is a mixture of delayed and attenuated versions of the original direct sound. Reverberation is usually modelled as a convolution between the direct sound and a room impulse response (RIR). Several factors such as the geometric shape of the room and locations of sound sources and receivers jointly determine an impulse response, which can be divided into three parts: direct sound, early reflections and late reflections. The different components of the sound will be covered in more detail in the following. Figure 2.9 and 2.10 shows these components.

Direct Sound refers to the foremost sound signals that reach the microphone through the distance between it and the source without hitting any surface. If the sound source is not within direct line of sight from the receiver, no sound shall be considered as the direct sound. The time taken by the sound to reach the microphone is dependent on the distance between the source and the receiver and its speed.

Early Reflection is when, a short time after the direct sounds, the process of reflections of the direct sound from objects like walls and furniture, reaching the receiver after the direct sound itself but within a short time span is called early reflections.

Because the propagation varies with distance, these early reverbs are distinguished from the direct sound concerning timing and direction of impact. The variance is dependent on the size of the space and the distance between the source and the microphone. Early reflections can have a positive impact on the intelligibility (Bradley, Sato, & Picard, 2003; H. Kuttruff, 1979; Omologo, Svaizer, & Matassoni, 1998).

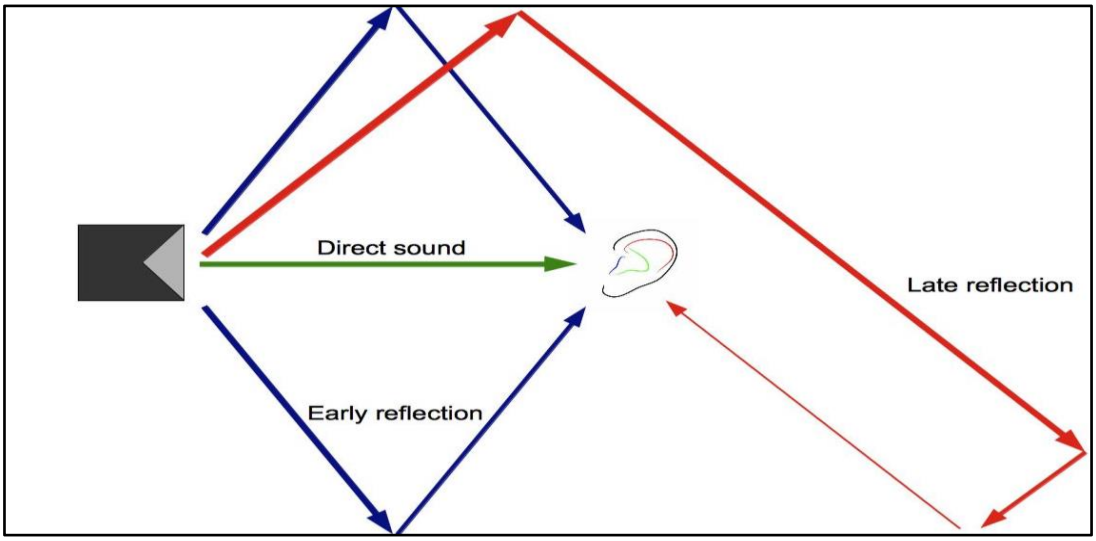


Figure 2.9 different types of reflections

Late Reverberation results from reflections, which arrive with longer delays after the arrival of the direct sound. Together, early reflections and late reverberation corrupt harmonic structure and formants of speech and present a considerable challenge to speaker recognition systems. Late reflections smear the speech spectrum across time (Wu & Wang, 2006). “Also, late reverberation is the principal cause of the ASR degradation (Petrick, Lohde, Wolff, & Hoffmann, 2007)”.

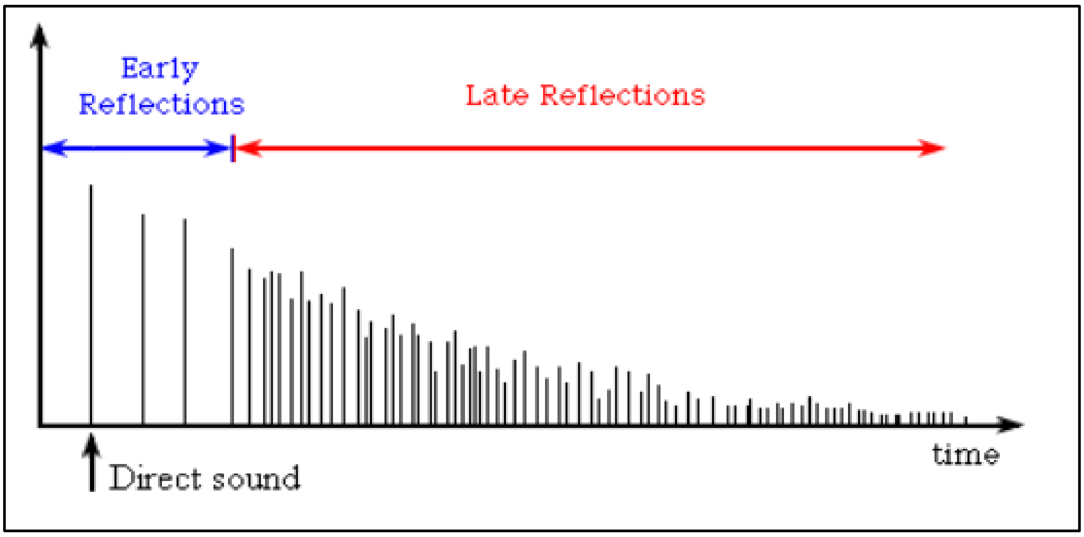


Figure 2.10 the main parts of impulse response

2.6.2 Acoustic Impulse Responses

“In room acoustics, it has been well recognised that the Room Impulse Response (RIR)” describes the reverberation properties of an enclosure for a particular source-receiver position completely and that it involves the direct path signal, an early reflections part affecting mainly the signal’s timbre, being perceived as colouration and late part(Heinrich Kuttruff, 2009).

Figure 2.11 illustrates an instance of room impulse response. Direct-path propagation from the sound source to the receiver provides increase to an early short period of near-zero amplitude, sometimes denoted as the direct path propagation delay, and followed by a peak. The amplitude of this peak due to direct-path propagation may be larger or less than the amplitude of the later reflections, according to the source-receiver distance and the reflectivity of the surfaces in the room. The instance of Figure 2.11demonstrations a relatively strong direct-path element, representing that the source-to-receiver distance is relatively short. The early and the later reflections are shown in the Figure 2.11 as two separate regions of the room impulse response. The beginning of the reflections is often taken as the first 50 ms of the impulse response(Heinrich Kuttruff, 2009) and constitute well-defined impulses of large magnitude relative to the smaller magnitude, and the diffuse nature, of the late reflections. The late reflections are denoted as the tail of the impulse response and constitute closely spaced, decaying impulses, which are apparently randomly distributed. The acoustic parameters RT and ELR have been averaged over a number of octave bands. The spectrum split into about seven bands. These are called the Octave Bands because there is one octave between the bottom and top of each band. The 1/3 Octave Band Filters are very similar to the Octave Band filters described. The difference is that each of the Octave Bands is split into three, giving a more detailed description of the frequency content of the noise. The centre frequencies of these bands are usually: 125Hz, 250Hz, 500Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz

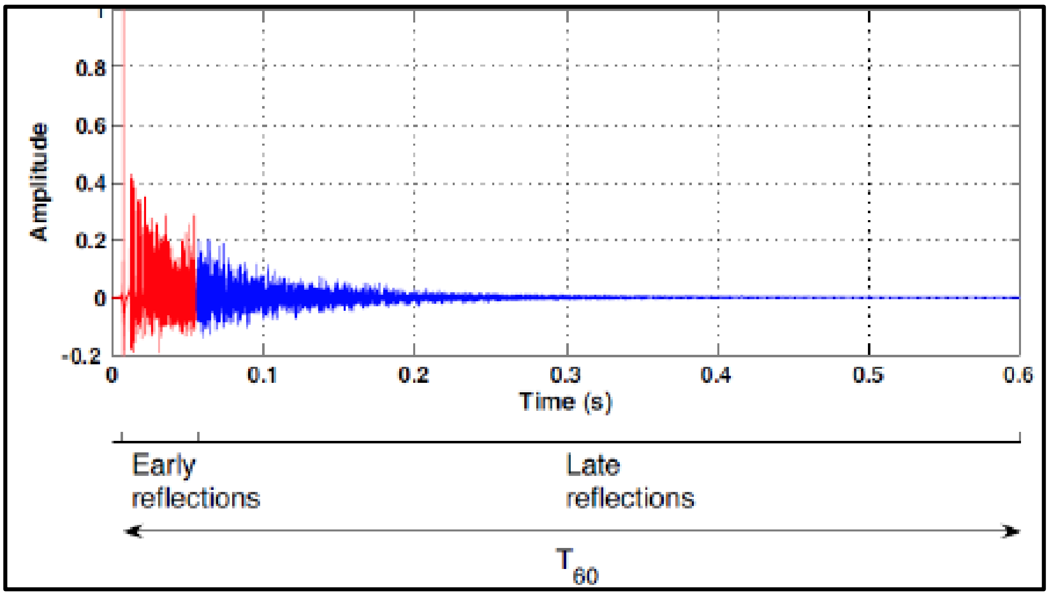
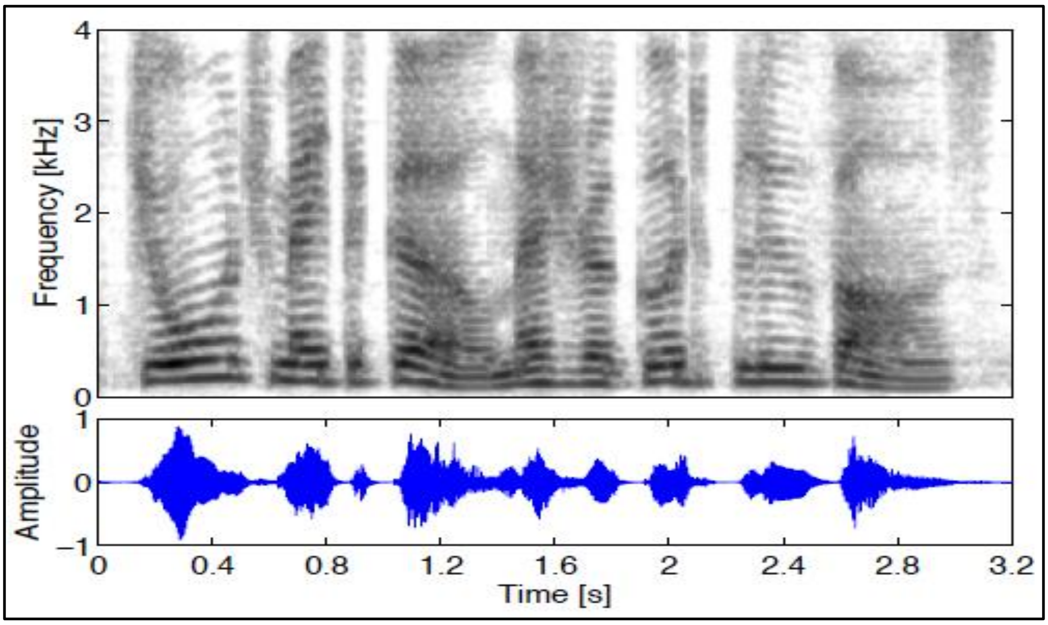


Figure 2.11 representation of the room impulse response

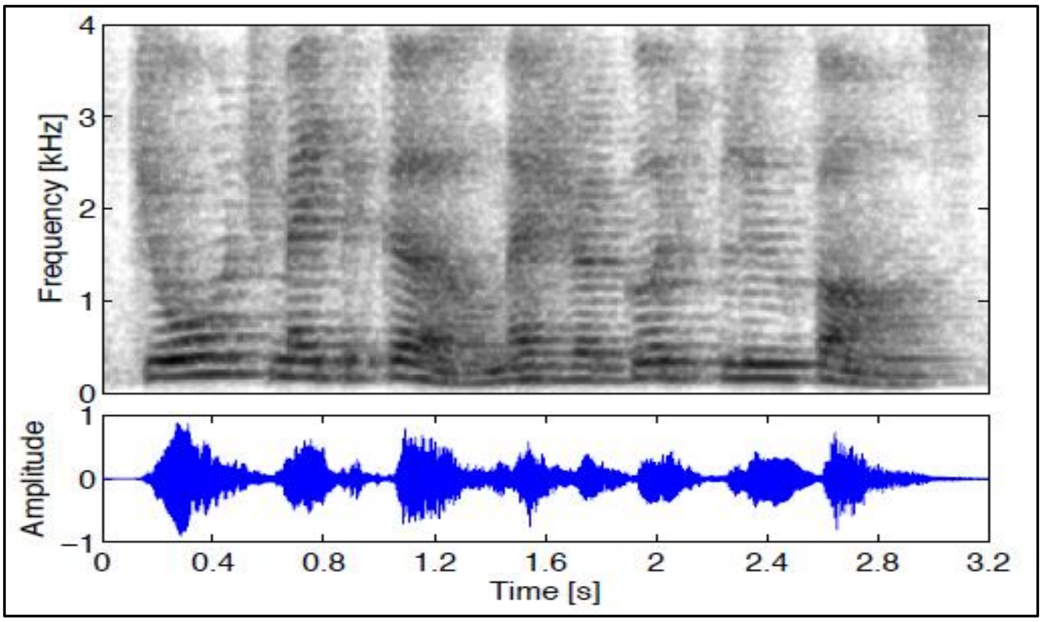
2.6.3 Impacts of Reverberation on the Speech Perception

Reverberation degrades speech recognition accuracy for human listeners. The level of degradation increases with increasing “reverberation time” and decreasing “direct to reverberant energy ratios”. The impacts of reverberation on speech are audible, and visible in the spectrogram and waveform of a speech signal. The perceptual effect of reverberation is typically classified according to the delay time. A reverberation that happens with delays up to a few tens of milliseconds modifies the short-time spectrum of presenting periodically spaced nulls into the speech spectrum. This impact, known as spectral colouration, is particularly noticeable in small rooms with highly reflective walls because the reverberations created have high amplitudes and short delay times (Berkley & Mitchell, 1974). After a delay of perhaps 50m, reverberation may be perceived as distinct copies of the direct path speech (Berkley & Mitchell, 1974), and cause temporal rather than spectral distortion. Several studies have established that reverberation degrades the intelligibility of speech under certain conditions. In a live situation, a normal listener can utilise binaural hearing and possibly other screening approaches to compensate for reverberation (Haas, 1972; Nabelek & Pickett, 1974). In a typical room, intelligibility between speakers and listeners of normal hearing is not affected severely by reverberation, at least in the absence of noise (Nabelek & Pickett, 1974). However, whenever a single microphone is utilised in a room to record speech, the benefit of binaural hearing is lost. In this case, reverberation decreases intelligibility (Haas, 1972; Nabelek & Pickett, 1974).

Most studies have found that long reverberation time affects intelligibility more severely than shorter reverberation time. Studies have shown that the earliest reverberations improve intelligibility because they effectively increase the energy of the speech signal (Kurtovió, 1975; Lochner & Burger, 1961). Figure 2.12 presents an example of spectrograms and time-domain waveforms for one speech portion (a) a clean speech signal, (b) the reverberant version. It is clear in both the spectrogram and time-domain waveform of the reverberant signal smearing of the speech, produced by the late reflections can be observed spectrograms and waveforms of (b) (Naylor & Gaubitch, 2010).



(a)



(b)

Figure 2.12 Spectrograms and waveforms of (a) an anechoic signal, (b) the reverberant version with a measured distance 1 m in an office room with RT= 0.5 s (Naylor & Gaubitch, 2010)

2.6.4 Room Acoustic Parameters

The subjective effect of room acoustics is primarily specified by the reflections that are introduced when a sound source is activated in a room. Special aspects of the subjective effect are regarding the various properties such as the rate of the decay of reflections, the direction of the arriving reflections concerning the listener and the diversity of the response concerning frequency. Reverberation parameters are usually utilised to quantify room acoustics. It is defined and computed from decay curves. Given measurements of the room impulse response, the audio factors can be calculated directly. This section presents a review of the relevant acoustic parameters. The international standard ISO3382 (Standard, 1997) defines all of these parameters.

2.6.4.1 Reverberation Time

Reverberation time is a very significant factor in acoustic space. Reverberation time is the first objective parameter studied in the history of room acoustics by W. C. Sabine about 100 years ago (Sabine, 1922). He used organ pipes as a sound source and revealed that by predicting the time it takes for the sound to become inaudible when switched off, the ratio of sound decay could be associated with the size of the enclosure and the quantity of absorption within it. This study assisted in defining the reverberation time parameter. Sabine determined that the reverberation time depends on the volume of the room, *V* and inversely depends on the amount of absorption in the chamber. Sabine's procedure estimates the reverberation time, ignoring the impact of weakening due to spread via the air, as

$$T_{60} = 0.163\,\frac{V}{S\alpha} \qquad 2.5$$

In the above equation, *V* is taken as the room size in cubic meters, the surface area of the walls the room is taken as *S* and in m^2, while means of the absorption coefficient of the walls is represented by α. The level at which the sound became inaudible was assumed to be approximately 60dB below the initial level, and RT$_{60}$ was more properly defined as the time taken for the sound pressure level to decay by 60dB when a stable sound source is switched off. Later definitions (due to the difficulty in achieving a 60dB signal to noise ratio in recordings) used the variation from - 5dB to -35dB and achieved least the squares fit in this range. The time it takes for this line to decay by 60dB is inferred and is denoted as RT$_{30}$. Frequently, when the signal to noise ratio (SNR) is inadequate, RT is computed from smaller dynamic variations. For instance, from the range -5 to -25dB would be RT$_{20}$. Schroeder suggested an important way to determine decay curves from impulse responses known as Schroeder backward integration method (Schroeder, 1965). This is defined as the decay of the squared sound pressure against time when a broadband sound is switched off after having a steady sound energy distribution. It can be utilised to display particular characteristic of the reverberation, such as the variance between the early and late responses. Figure 2.13 illustrates an instance of a decay curve.

$$EDC(t) = \int_{t}^{\infty} h^{2}(t)dt \qquad 2.6$$

In the given formula, the room's impulse response is represented by *h* (*t*). From the decay curves obtained by Schroeder backward integration, reverberation parameters including RTs, EDTs can be subsequently calculated. Early decay time was formally introduced and defined as an objective parameter by Jordan (1970). It is defined as the 60 dB decay time calculated by a line fit to the 0 to -10 dB portion of the decay curve. When Early decay time (EDT) is measured in a diffused field, where the decay curve in dB is linear, Early decay time (EDT) and Reverberation Time (RT) are the same. Otherwise, Early decay time (EDT) has been found to be better correlated with subjective judgment of reverberation than the classical reverberation time (B. Atal, Schroeder, & SEssLER, 1965). Moreover, it is worth noting that EDT is more position dependent, while reverberation time tends to be more evenly distributed at different locations in a room.

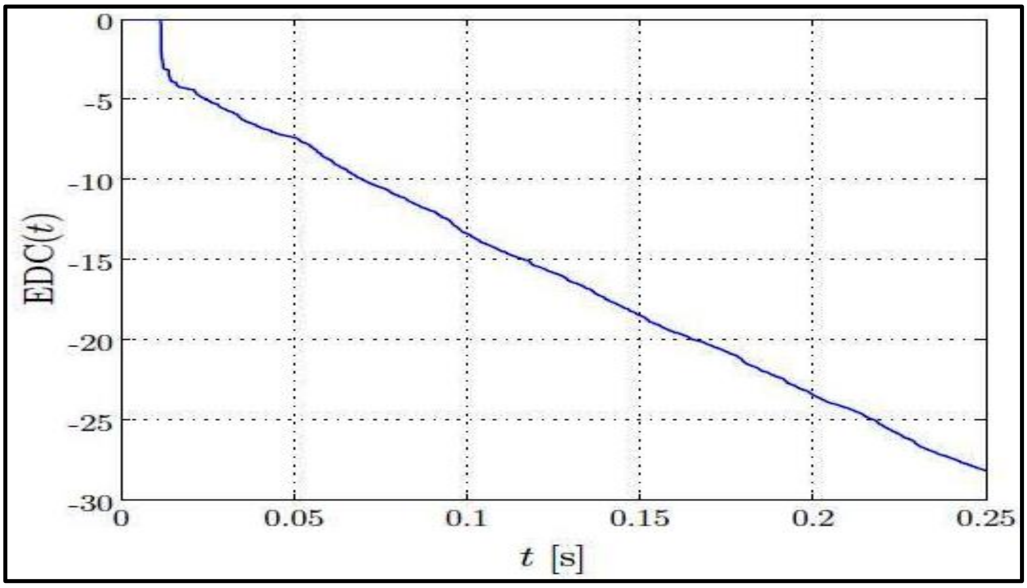


Figure 2.13 Energy decay curve(Kendrick, 2009)

2.6.4.2 Direct-to-Reverberant Ratio

The direct-to-reverberant energy ratio (DRR) is one of the most important parameters when it comes to the analysis of room acoustics. It is absent from most discussions of room acoustics. Only the direct sound (DS) provides information about the localisation and distance of a sound source. The direct-to-reverberant ratio, which describes the energy ratio between the direct and reverberant component of a sound field, is an essential parameter in many audio applications. It dose not only determines the acoustic quality of room but also serves as an integral element in many audio applications. For example, speech enhancement and dereverberation (Bloom, 1982; Jeub, Schafer, Esch, & Vary, 2010; Naylor & Gaubitch, 2010), source localization (Y.-C. Lu & Cooke, 2010), parametric spatial audio coding (Pulkki, 2007), performance evaluation of beam-forming (Jarrett, Habets, Thomas, Gaubitch, & Naylor, 2011) and psychoacoustics, where it is believed that the DRR helps humans to determine the distance to a sound source (Y.-C. Lu & Cooke, 2010; Vesa, 2007, 2009). There is also another important aspect in DRR relating to human hearing. Recent research on human hearing has concluded that DRR may provide absolute distance information, especially in reverberant environments (Zahorik, Brungart, & Bronkhorst, 2005). The knowledge of DRR also helps the derivation of various other acoustic parameters such as reverberation time (T_{60}), and diffuseness (Jo & Koyasu, 1975; Laitinen & Pulkki, 2012). Due to the broad usefulness of the DRR, its estimation accuracy is considered vital. The most primitive method to calculate the DRR is to use the room impulse response (RIR) measured by an omnidirectional microphone. Even though the DRR can be estimated using only the beginning part of RIR (Larsen et al., 2003), reasons such as the need to use intrusive signals to reliably obtain RIR measurements, the requirement to repeat RIR measurements with moving source and receiver positions and the necessity of prior processing to identify the initial part of the RIR, make this estimation process less practical. The DRR measures the ratio between the energy propagating along the direct path (i.e. without reflections) and the reverberant energy(Naylor & Gaubitch, 2010).

$$DDR = \frac{\int_{\tau} h_d(\tau)^2 \, d\tau}{\int_{\tau} h_e(\tau) + h_r(\tau))^2 \, d\tau} \qquad 2.7$$

where $h_d(\tau)$ is the anechoic direct propagation path, $h_e(\tau)$ describes the early arrivals up to some tens of milliseconds and $h_r(\tau)$ denotes the late diffuse reverberation typical of the RIR tail. The direct-path propagation is often assumed to be the most significant magnitude peak at the beginning of the impulse response; however in practice, due to limited rate sampling of the RIR, it can often be difficult to identify it precisely. “The DRR is dependent on the distance between the source and the microphone, the directivity factor of the source and the reverberation time of the room”.

2.7 Related Work of Speaker Recognition in Reverberant Conditions

Reverberation is a typical distortion in daily acoustic environments. The issue of robustness to reverberation has received a lot of attention in the speech community, and many methods have been proposed in the literature.

Castellano et al. (1996), investigated the impact of reverberation in closed set text-independent speaker recognition. Multiple binary classifier models exist where neural networks were utilised as the recognition method. The measure used to quantify the recognition accuracy was the number of correctly classified speech frames. In this work, the feature vectors compared were line spectrum pairs (LSP), reflection coefficients, and Mel-Cepstrum coefficients. The result was strongly dependent on the location of the speaker, room dimensions, and reverberation time.

Lin, Jan, and Flanagan (1994), proposed the use of microphone arrays to record speech in various environments and use this as input to the speaker identification system. The system evaluated the use of speech signals contaminated with reverberation created by a computer model of room acoustics and transduced by various simulated microphone arrays.

Jin et al. (2007), proposed a new approach to increase the robustness of speaker recognition in far field microphone situations such as meeting scenarios. The new technique is presented for reverberation compensation and feature warping of training and testing signals to increase speaker identification performance in contaminated conditions. One disadvantage of this work is the requirement for multiple training signals acquired in reverberant environments.

Gammal (2004), examined the impact of enrolment on speech created from reverberant conditions that are dissimilar from those surrounding the speech used in the testing stage. The results confirmed that using speech contaminated with a similar level or less reverberant than the test speech always improved performance.

Nakatani and Miyoshi (2003), proposed a new blind dereverberation method of recording the speech signal with a single microphone. For applications such as speech recognition, reverberant speech causes serious problems when a distant receiver is used in recording. In this method, harmonic structure is estimated and used to approximate the direct sound in a reverberant sound. The dereverberation operator is calculated as the average ratio of the approximated direct harmonious sound to the reverberant sound and is shown to give the estimate of the inverse transfer function that can be used for the dereverberation.

Akula and de Leon (2008), used a similar approach to that described in (Nakatani & Miyoshi, 2003). They differed in the method of creating the reverberation filters, by utilising the modified image method. They also used higher levels of reverberation than those used by (Nakatani & Miyoshi, 2003) to generate the training room impulse responses. A GMM model with Expectation model (EM) was used to estimate the parameters of GMM such as weights, mean vectors, and covariance matrices. The feature vector is constructed by using MFCC features.

Peer et al. (2008), presented a comprehensive study of the effect of reverberation on speaker verification and investigated approaches to reduce the effect of reverberation. They trained background models in different reverberant conditions.

A reverberation classification was performed to find the best matching background model. The corresponding speaker models were used for SV. The method showed significant improvement in performance.

Garcia-Romero et al. (2012) dealt with noise and reverberation using multi-conditional training. In particular, they created noisy and reverberant training data. Gaussian PLDA subsystems were trained in each of the multiple training environments and combined to yield SV scores. Three training scenarios were discovered. The first one trained subsystems independently across training conditions. The second one presumed that all subsystems shared the same latent variable, and the last one produced only one set of parameters shared by all subsystems, by pooling all the training data together. They described results in noisy and reverberant conditions individually.

De Leon and Trevizo (2007), suggested training speaker models in multiple reverberant conditions (i.e. rooms) to decrease the mismatch created by reverberation. Multiple models were derived for each speaker and utilised for speaker identification as if they were from different speakers. Later, they improved the system by incorporating a GMM-UBM framework (Akula et al., 2009). Two training scenarios were suggested. The first one trained a room-independent UBM and speaker models were adapted from it. The second one also trained a RI-UBM, and then room-dependent UBMs were adapted from its utilising room specific training data. Speaker models have adapted from the UBMs afterwards. Through testing, the first scenario went via all the speaker models and the speaker with the highest score was the output. The second scenario showed room classification utilising the UBMs to choose the closest training room. The speaker models associated with the selected room were employed for speaker identification.

Robust speaker features are studied to combat reverberation. Falk and Chan (2010), suggested spectral modulation features, based on auditory filter banks. Specifically, input speech was passed through a 23-channel Gammatone filter bank. Envelopes of filter outputs were extracted and fed to an 8-channel modulation filter bank to obtain the spectral modulation features, which were shown to be robust to reverberation.

L. Wang and Nakagawa (2009), developed a novel phase extraction technique that converted phase into coordinates on a unit circle. They confirmed that a combination of phase-based features and MFCC was useful in reverberant test conditions.

Borgstrom and McCree (2012), suggested a technique to enhance reverberant speech. Evaluations of reverberation time estimation and SV established good performance. It has been established above that reverberation corresponds to a convolution between an RIR and anechoic speech. With the short time Fourier analysis, the convolution could not be converted to frame level multiplication because typically reverberation time is in the order of hundreds of milliseconds and is much longer than the analysis window (20-30 milliseconds). In their study, an assumption was made that the STFTs of the RIR and anechoic speech were convolved with each other. In this case, reverberation was characterised as the channel wise convolution of STFTs of the RIR and anechoic speech. The convolution was equivalent to multiplication in the spectral modulation domain after Fourier analysis. The original reverberant speech was enhanced in the spectral modulation domain and resynthesized to time domain for subsequent SV.

An alternative speech enhancement technique is blind dereverberation (Sadjadi & Hansen, 2014), focused on restoring the direct sound and early reflections of reverberant speech as the late reflections were believed to be detrimental. Similar to Wiener filtering, this function is applied to the reverberant spectrum and weakened the effects of late reflections. Their evaluations indicated that the proposed method worked well in both speaker identification and speaker verification.

A multi-channel feature enhancement method was applied in the log-spectral domain by (Kang, Kang, Lee, Cho, & Kim, 2014). In the proposed approach, the authors extended the interacting multiple models (IMM) algorithms initially designed for the single channel scenario such that it could be fitted to multi-channel processing. The proposed technique has two significant benefits. First, no a priori knowledge of the room impulse response (RIR) is needed. Second, the parameters concerned with acoustic reverberation and background noise are sequentially updated in a frame-by-frame manner instead of on an utterance-by-utterance or file-by-file basis for tracking the nature of their variation with time. This type of real-time update of the RIR parameters is essential in handling the possible movements of the talker or microphones. From various experiments in noisy reverberant environments, it has been confirmed that the proposed algorithm outperformed the traditional single-channel algorithm.

2.8 Chapter Summary

In this chapter, we have presented a short explanation of the basic concepts necessary to contextualise the work presented in this thesis. First, we defined the speech production system and the process of speech signal generation, followed by an overview of the basic structure of speaker recognition and the most important applications of speaker recognition were discussed. Furthermore, a review of the feature extraction methods that are used in speaker recognition system was presented.

Moreover, the central ideas and methods for modelling speakers with Gaussian Mixture Models were discussed. The classical paradigm of speaker recognition based on

Gaussian mixture models was summarised. The role of the speaker recognition evaluations was highlighted in this discussion. In addition, this chapter has reviewed the fundamental properties of acoustic environments and presented some elementary theoretical acoustic properties, which are essential for understanding why particular models are utilised, and the main parameters of room acoustics were discussed. Finally, some works related to speaker recognition under reverberant conditions have been reviewed in this chapter.

CHAPTER THREE: DATA SETS

One of the most significant challenges in making the automatic speaker recognition system more robust is the inability to collect sufficient amounts of data to train acoustic models and perform meaningful evaluations. The collecting of speech samples requires hundreds of hours of work in recording data as well as a transcription for training and evaluation purposes. Once acoustic models have been trained, experimental evaluation needs an extraordinary amount of test data, which must be different from the training data, to acquire statistical descriptions of the performance of any robust speaker recognition technique. This chapter intends to cover some details regarding the data sets of speech and the impulse responses that were used in this work, followed by the methods employed to generate the impulse responses.

3.1 Anechoic Chamber at Salford University

An anechoic chamber is acoustically akin to rising above the ground outdoors because there are no reflections from the surface of the room such as the walls, floor or ceiling. That means it is ideal for testing the response of loudspeakers or microphones because the chamber does not affect the measurements. Furthermore, it is the best place for generating representations of concert halls using virtual acoustics, city streets and other spaces.

The anechoic chamber is immensely quiet which makes it ideal for testing very quiet products or people hearing very quiet sounds ("Anechoic chamber-Salford University"). An anechoic chamber is designed to be isolated from any sound reflection from walls, floor, and ceiling. The walls, floor and ceiling of the inner chamber are made of heavy Accrington brick and concrete to prevent sound getting into the room. Two heavy acoustic doors with rubber seals are used to minimise airborne sound.

The full inner room is mounted on a set of springs-neoprene rubber mounts to decrease vibration. Every surface is covered with absorbent materials to reduce reflections from the walls of the chamber. The floor you walk on is a wire trampoline stretched between the walls with an acoustically transparent catch net below. The Background noise level of the room is (-12.4dB), and the Cut-off frequency is 100Hz. The working area of the room is 5.4 x 4.1 x 3.3m. Figure 3.1 shows the anechoic chamber at the University of Salford.

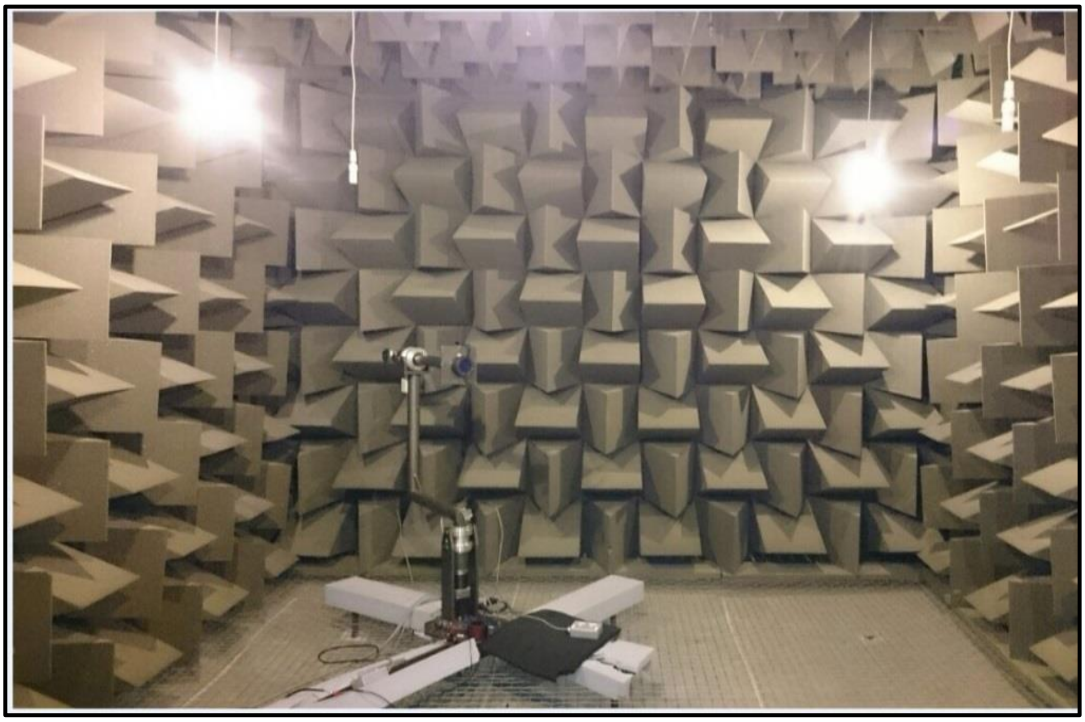


Figure 3.1 Anechoic chamber at University of Salford

3.2 Speech Datasets

This section presents a brief description of the datasets, which are used for the evaluation of speaker recognition algorithms in the experiments described in this thesis.

A benchmark database is essential for the study of speaker recognition. The selection of an appropriate speech data set is of fundamental importance in testing performance while developing speaker recognition methods. Real world speaker recognition is usually utilised in non-ideal environments, including acoustic reverberation. In addition, some applications include recognising an individual later than the date of the provided speech sample, so reliability over an extended period is necessary. In this thesis, Salford University–Anechoic Chamber (SALU-AC), which was recorded by researchers at the University of Salford, has been adopted.

3.2.1 Salford University Anechoic Chamber (SALU-AC) Dataset

The intention behind creating speech databases for speaker recognition is to acquire rich voice messages concerning measuring inter and intra-speaker variability. The SALU-AC corpus design was a joint effort of the PhD students from the department of acoustics, Salford University, UK. The data set was evaluated for a PhD project regarding speaker recognition. The SALU-AC data set was collected in an anechoic chamber at Salford University from 110 volunteer speakers (58 male, 52 female); all volunteers had lived in the United Kingdom. The reason for collecting such as data set is the study of speech needs, the availability of particular conditions such as very high SNR and almost without reflecting sound. The dataset has been designed to provide speech material for the development and evaluation of automatic speaker recognition systems. The text language is English and is read by speakers. There was no formal rehearsal, and perfect pronunciation is not obtained, nor is it necessary, for obtaining the specific and unique identifying characteristics of individuals. The details of the dataset are as follows:

1) ***Volunteer Speakers***: In this database, the audio speech samples were recorded in an anechoic chamber at Salford University from 110 volunteer speakers (58 males, 52 females). The speaker group exhibits relatively small variation in age, profession, and educational background. Each volunteer was instructed to utter three speech samples with different duration times (approximately 60 seconds for the first sample and around 40 seconds for the other samples).

2) ***Language:*** All speakers are recorded in the English language. The speakers rely on using random general text from different readable resources. (Newspapers, books, leaflets, articles, etc. to obtain text/language independent samples.

3) ***Recording Equipment***: the equipment for recording is Zoom H6 portable solid-state recorder. The Zoom H6 can encode audio using a variety of compression algorithms, sample rate, bit rate, and file format. It supports two kinds of recording format: compressed recording, which includes MP2 and MP3, and uncompressed recording, which includes linear pulse code modulation (PCM). The record type can be stereo, mono, and the file can be saved in the (.Wav) format. "In this database, the voice messages were recorded into the most commonly used file types (.wav). The distance between speaker and microphones was approximately 0.25m and the sampling frequency chosen was 44100 kHz".

4) ***Audio Speech Samples:*** The audio speech samples for each speaker are divided into short 5-second sample utterances, to give 30 utterances. The duration of the samples for each speaker is 150s. The database thus contained 3300 speech utterances.

5) ***Silence Suppression:*** Silence removal block is used to eliminate the unvoiced and a silent portion of the speech signal. For this purpose, the input signal is divided into small segments (frames), and root means square (RMS) of each individual segment is calculated and compared with a specific threshold value. The total length of each individual segment is equal to the product of time duration and sampling frequency of segment (Y. Lu, 2010). Accuracy and performance of silence removal block depend on a total number of segments. The total number of segments can be calculated by dividing the total length of an input signal by length of the individual segment.

   Figure 3.2 shows two signals before and after silence removal.

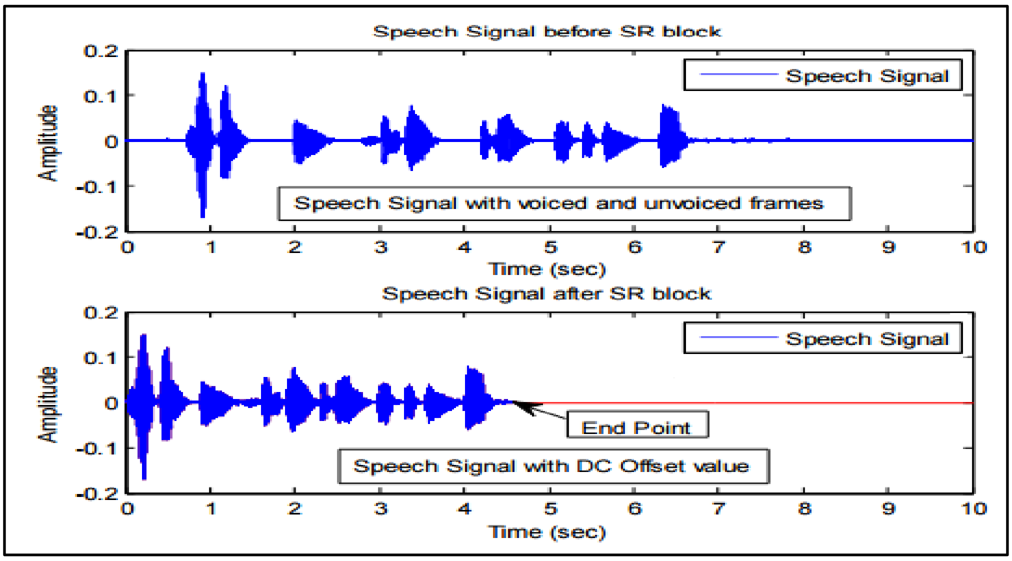


Figure 3.2 Graphical representation of silence removal

6) Normalisation Stage: The issue of normalisation needs to be addressed so that the speech signals can be added in the correct proportion to avoid misinterpretation.

The default method is normalisation of the mixed or compared signals to the same perceived level, and it is a significant factor for reliability that the input signals have the same level (Omologo et al., 1998). The convoluted signals are processed to have the same RMS.

7) Availability: demonstration of the SALU-AC database is available from the website of University of Salford data repository-Figshare (https://salford.figshare.com/). Academic researchers can contact the authors to obtain a free personal passcode for the complete dataset.

3.3 Requirements for Training and Validating Examples

Similar any statistical method, the accuracy and reliability of the machine learning method are built upon training on a large set of realistic examples. In supervised training, the common practice is to use as large as possible a training data set to attain good generalisation, and the validation is achieved utilising examples never used in the training phase (Haykin, 1999). Usually, all available examples are split into equally sized training and validation sets. If satisfactory generalisation is not achieved, an enlarged data set with more examples is warranted. Due to the complexity of room impulse responses, rooms having similar reverberation times may show quite different impulse responses, especially the early reflection patterns. Real room measurement and sampling is certainly the most convincing way to obtain the required data set. However, it is unrealistic on the scale of a study like this to measure such a large number of spaces and obtain all the required impulse responses.

3.4 Computational Room Modelling Techniques

Any linear sound propagation phenomenon is ruled by the well-known wave equation (Helmholtz equation). An impulse response from a source to a receiver position should be theoretically obtainable from solving the equation. Unfortunately, the wave equation takes analytic forms only in occasional cases. Different approximations have to be made in solving practical problems. Three categories of computational methods for room acoustics are found in literature namely, ray-based approaches, wave-based approaches and statistical approaches. The statistical approaches (Lyon, DeJong, & Heckl, 1995) are often used to predict noise levels in coupled systems such as structure-borne noises. However, they do not model temporal behaviors and therefore are not appropriate for modelling impulse responses. The wave-based approaches (Botteldooren, 1995) are suitable to and mainly used for modelling very low-frequency sound propagation. Furthermore, the very heavy computation makes them inappropriate for generating a large number of samples. “The ray-based approaches (geometrical acoustics based computational approaches), comprising ray-tracing (Krokstad, Strom, & Sørsdal, 1968) and image-source methods (Allen & Berkley, 1979), are the mainstream computational techniques for modelling acoustically critical spaces such as concert halls and theatres (H Kuttruff, 1995), therefore, are the apparently possible ones for sample generation”. Typically, the ray-tracing technique can simulate energy impulse responses but not the precise sound pressure of impulse responses. The image source technique exceeds the ray-tracing technique in terms of accuracy but is difficult to model curved and diffusely reflecting surfaces (H Kuttruff, 2000). Besides, the number of image sources grows exponentially as a function of the order of reflections, making it computationally inefficient in finding high order reflections. In practical uses, the ray tracing and the image-source method may be applied together: the image source method is used to simulate the early reflections taking the benefit of its accuracy; the ray tracing is used to handle the late reflections to save the computing time. "A successful and famous example hybrid computational simulator is CATT-Acoustics (CATT-Acoustic, 2010), which is now a commercially available room acoustic simulation package".

3.5 Impulse Response Data set

The room is often assumed as a linear, passive and time-invariant transmission system of sounds (Heinrich Kuttruff, 2009). Impulse responses are used to characterise the properties of the room under such assumptions. Alternatively, transfer functions, which are the Fourier transform of the impulse responses, can be adopted.

This section defines the measured and simulated impulse response database. Moreover, in this section, we define some methods, which are used in work described to simulate the impulse response data set. Two methods were used in this work to generate artificial room impulse responses (RIR); the first technique is called Image Source Method (ISM) and the commercial acoustic prediction software CATT acoustic

3.5.1 Aachen Impulse Response (AIR) Database

The Aachen Impulse Response (AIR) database is a set of impulse responses that were measured in a wide variety of rooms, which, offers a wide range of reverberant conditions (168 RIRs with T60 ranging from 0.1 s to about 4 s). The primary purpose of the AIR database was to allow for realistic studies of signal processing algorithms in reverberant environments (Jeub, Schafer, & Vary, 2009). It is a free database and available in the "Aachen Impulse Response Database" 2009). An impulse response was used to evaluate the performance of the system. Figure 3.3 shows room properties and measurement setup for the meeting and lecture room.

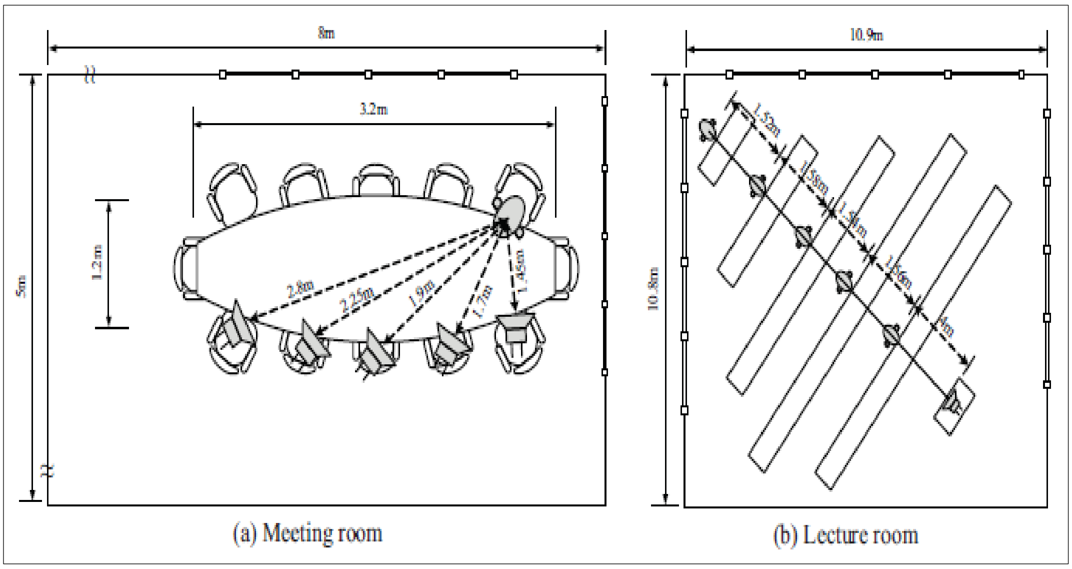


Figure 3.3 Room properties and measurement setup for the meeting and lecture room (Jeub et al., 2009).

3.5.2 Simulated Impulse Response

In this study, two methods were used to generate artificial impulse response. In the following sections, more details about those methods are given.

3.5.2.1 Image Source Models

The image source model (ISM) is a common technique in the acoustics and signal processing industry, which has a broad range of applications in acoustic engineering, including source separation, reverberation prediction, acoustic source localisation, speech intelligibility and enhancement, and much more. Aresearch work was developed a new speech synthesis system, which is based mainly on the fractal dimension to create natural sounding speech in (Fekkai & Shafik, 2013). The imaging technique is particularly suitable for rectangular enclosures (Santon, 1976). In the rectangular box-shaped chamber, it is simple to arrange all image sources for a certain order of reflection (Allen & Berkley, 1979).The image source method allows the calculation of sound behaviour using ray propagation assumptions, travelling directly from source to receiver but also indirectly. It can then be extrapolated as sound reaching the receiver from two sources; the second source is the image source behind the mirror. The image source position is calculated from the source position and reflector position and angle. The straight-line distance between image source and receiver contains the information required to model the actual reflected sound path. Each reflective surface itself produces an image source, and a second order image source is produced by the combined reflection of the two surfaces, giving us four active sources instead of one. The greater the number of image sources the higher level of accuracy in estimated calculation of reverberation time. “A design of the image- source method is illustrated in Figure 3.4. An image source method, which is developed and implemented in Matlab by Eric A. Lehmann, was used in this work”.

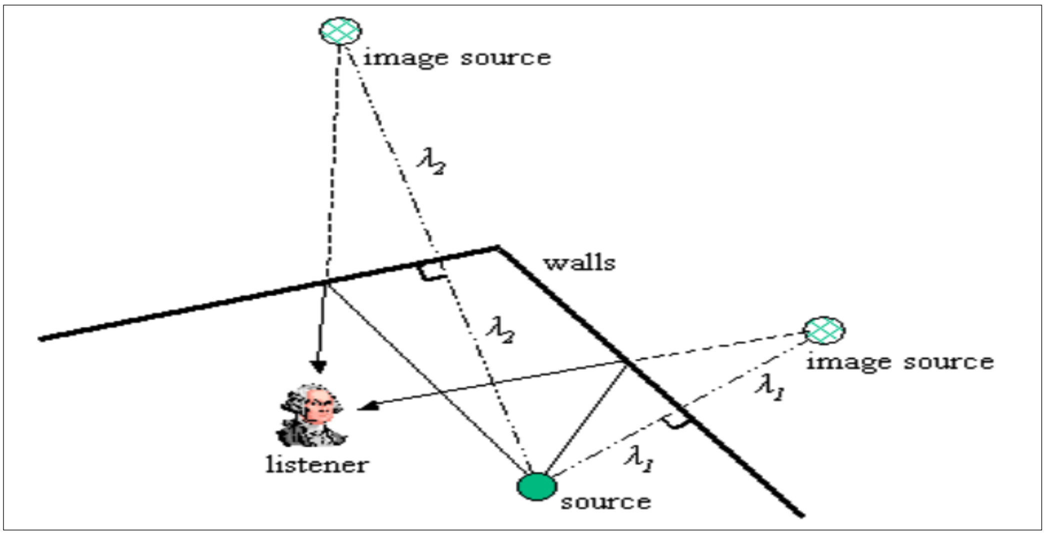


Figure 3.4 Image source method process

3.5.2.2 CAT –Acoustic Software

A geometric room model with variable dimensions is defined, surface properties and source and receiver locations are specified and from this, a large number of realistic room models generated. For creating the rooms, each size is randomly chosen to use an arbitrary number, constrained within appropriate (realistic lengths) bounds. Impulse responses were then predicted using CATT acoustic for each model. A box-shaped room and a fan-shaped room model were used to generate the impulse responses. An enormous number of valid geometries were generated, and from these, room impulse responses were generated using the sequence programmer built into the software; a number of responses can be generated in sequence. A room model is generated and illustrated in Figure 3.5. This model had an Omni-directional sound source positioned on the stage 2m above the floor and within a rectangular area in the center of the stage ensuring the source is at least 1m from any surface (ISO, 3382 : 1997). Receiver placement close to any surface was avoided (at least 1m from any surface) as recommended in ISO 3382. Receivers were not placed exactly in the center of the room to avoid any irregularities where reflections exactly constructively interfere. Suitable absorption properties (appropriate as wall material properties were only applied to walls etc.) were arbitrarily selected from a database of materials provided with the software. Figure 3.6 shows a screenshot of CATT-Acoustic software.

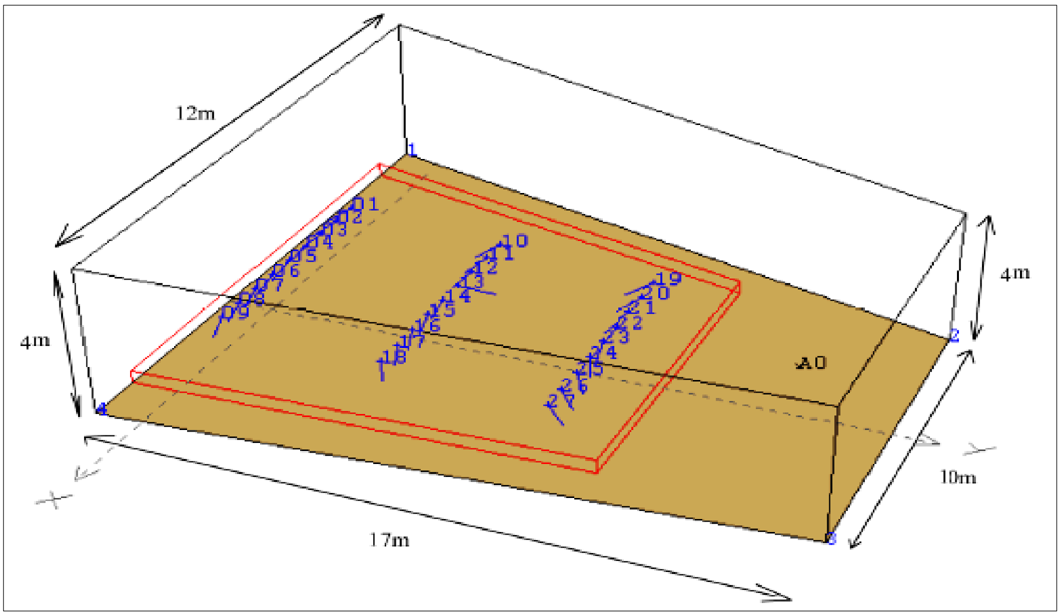
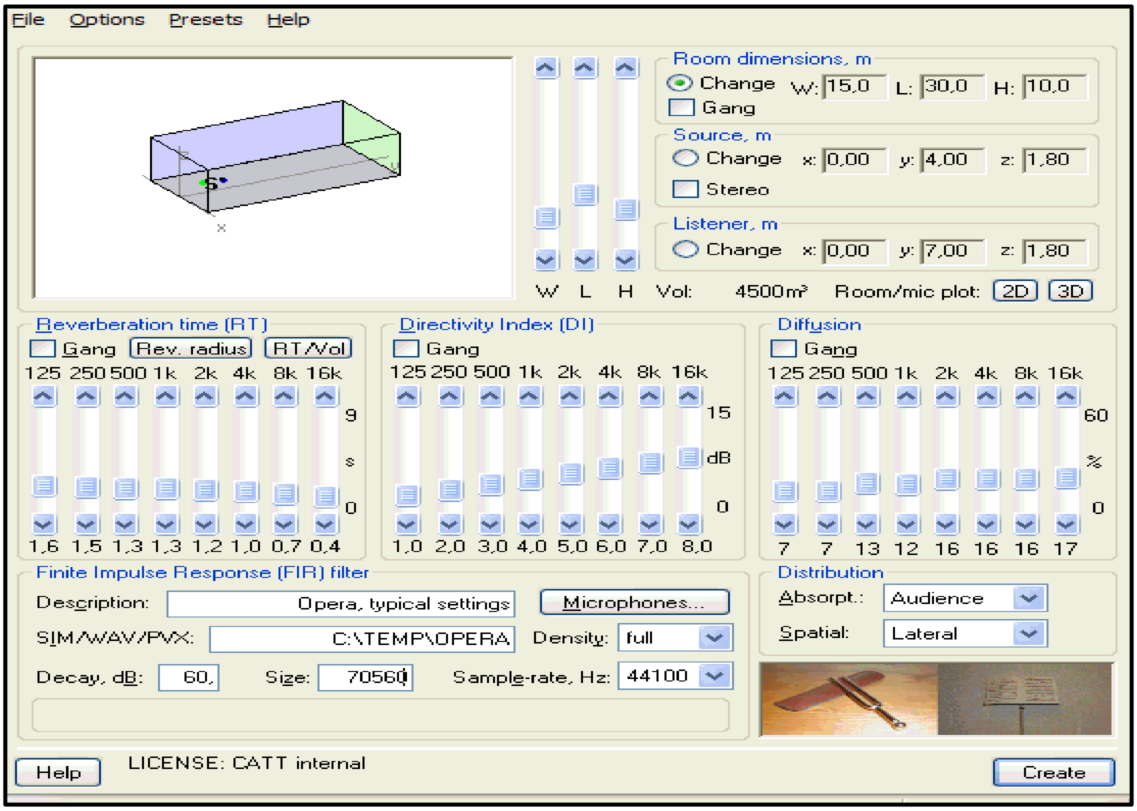


Figure 3.5 View from CATT acoustic showing one of the randomly generated room geometries, the sound source is labelled as A0,



Room dimensions, m
● Change W: 15.0 L: 30.0 H: 10.0
□ Gang

Source, m
○ Change x: 0.00 y: 4.00 z: 1.80
□ Stereo

Listener, m
○ Change x: 0.00 y: 7.00 z: 1.80

Vol: 4500m³

Reverberation time (RT) □ Gang (9 – 0 s)

125	250	500	1k	2k	4k	8k	16k
1.6	1.5	1.3	1.3	1.2	1.0	0.7	0.4

Directivity Index (DI) □ Gang (15 – 0 dB)

125	250	500	1k	2k	4k	8k	16k
1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0

Diffusion □ Gang (60 – 0 %)

125	250	500	1k	2k	4k	8k	16k
7	7	13	12	16	16	16	17

Finite Impulse Response (FIR) filter
Description: Opera, typical settings
SIM/WAV/PVX: C:\TEMP\OPERA Density: full
Decay, dB: 60. Size: 70560 Sample-rate, Hz: 44100

Distribution
Absorpt.: Audience
Spatial: Lateral

LICENSE: CATT internal

Figure 3.6 CATT Pure Verb screenshot(CATT-Acoustic, 2010)

3.6 Chapter Summary

This chapter has described the different speech datasets that are used in this work. Two separate data sets have been reviewed. Furthermore, this chapter has outlined methods for generating a database of the impulse responses. Sets of room geometries and surface material properties are randomly generated within a framework, limiting the models to realistic geometries. The range of parameters that the simulated RIRs show covers those that may be predictable from most real rooms.

CHAPTER FOUR: EVALUATION OF EXISTING METHODS

4.1 Microsoft Speaker Recognition (MSR) Toolbox

MSR is a well-known open source tool used for speaker recognition. The MSR was developed by Microsoft Research as a MATLAB toolbox to help with speaker recognition research (Sadjadi, Slaney, & Heck, 2013). It provides researchers with a group of tools and a test bed to build baseline systems for experiments quickly. The MSR provides two kinds of tools for speaker modelling GMM-UBM and i-vector paradigms. The front end of this toolbox is responsible for transforming the speech signals into acoustic features in the feature extraction process. The cepstral features, especially the Mel frequency cepstral coefficients (MFCC) feature, are the most commonly used features in this toolbox.

The back end, however, includes the training (enrolment) and testing (recognition) phase. The training (enrolment) phase is responsible for estimating a model for each registered speaker to generate a reference model. The test segment is scored against all enrolled speaker models to determine the identity of the speaker (speaker identification) or against the reference model of a claimed speaker to make a decision on whether the speaker is the target speaker or an imposter.

4.1.1 Proposed Algorithm Framework for MSR

1) ***Pre-processing*****:** This step included a different process: reading and segmenting the input speech signal into frames and then multiplying each frame by the Hamming window to maintain the continuity between the first and the last points. In addition, the Speaker Activity Detection (SAD) was used to remove the silence from the speech signal.

2) ***Feature Parameters Extraction:*** This step refers to extracting Cepstral features for each audio, speech in the enrolment and recognition phases. This process is responsible for converting the audio speech signal for each speaker to Cepstral.

   The features of each speaker are then stored in a unique file. The created file is classified into training files (train, i=1, Number of user j=1,… number of utterances for ith speaker), which represent the feature space for each user (i) that has been used to create the reference model, and testing file (Test i=1…. Number of users), which represents the feature space for each user (i) in the recognition phase.

3) ***Training UBM from Background Data:*** This process is responsible for creating background models from a huge number of speakers by fitting GMM to acoustic characteristics utilising binary splitting and Expectation Maximisation (EM).

4) ***The Maximum a Posteriori (MAP)*****:** this process is responsible for adapting speaker-specific GMM from UBM. The output of this process represents the reference model for each speaker.

5) ***Scoring Verification*****:** This computes the verification score between the reference model of a claimed speaker and the recognition features of an input speech signal (Test I); the score is measured as the log-likelihood ratio between two models.

6) ***Decision-making*****:** Based on the score obtained from scoring verifications and a defined threshold, a decision is made whether a recognised feature of the input speech signal (Test I) belongs to the claimed speaker (Target) or not (Imposter).

7) ***System Evaluation:*** Making an evaluation of the performance of the system depends on the output of verification. “In this work, equal error rate (EER), and detection error trade-off (DET) curve are adopted. Figure 4.1 shows the MSR framework”

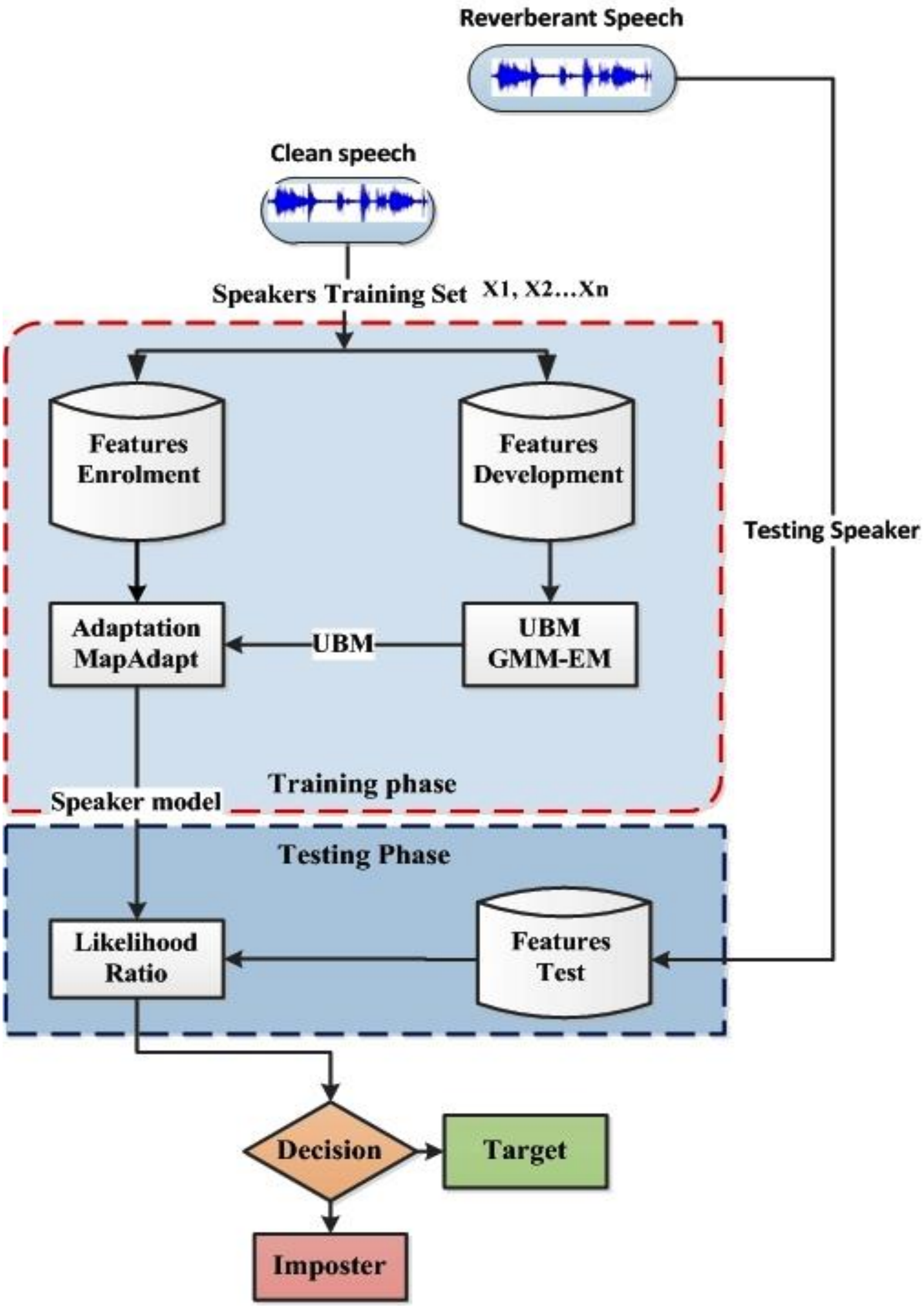


Figure 4.1 the framework of the MSR toolbox

4.2 Evaluation of the Performance of a Speaker Recognition System in Reverberant Conditions

There have been several studies in the literature reporting the effect of reverberation on speaker recognition systems. Therefore, this section stems from a study into the behaviour of typical speaker recognition systems in a reverberant environment. Validation tests were carried out with clean speech and with speech corrupted by reverberation to varying degrees. The first motivation for this study is to use the samples recorded in a reverberation room at Salford University to evaluate and validate the impact of real reverberation time on system performance. The second motivation behind the present study is to use speech obtained from the SALU-AC dataset, after convoluting it with the simulated impulse response, which is acquired from different simulated rooms. The image source method was adapted to take into account real acoustic conditions in the spaces. Statistical relationships between recognition accuracy and reverberation time (RT) have therefore been established. Results show reverberation can to different extents degrade the recognition performance. "Figure 4.2 shows the framework of the evaluation of the baseline system. This part of the study was presented at International Journal of Information and Electronics Engineering, Vol.5, No.6, November 2015 under the title "Automatic Speaker Recognition System in Adverse Conditions-Implication of Noise and Reverberation on System Performance''.

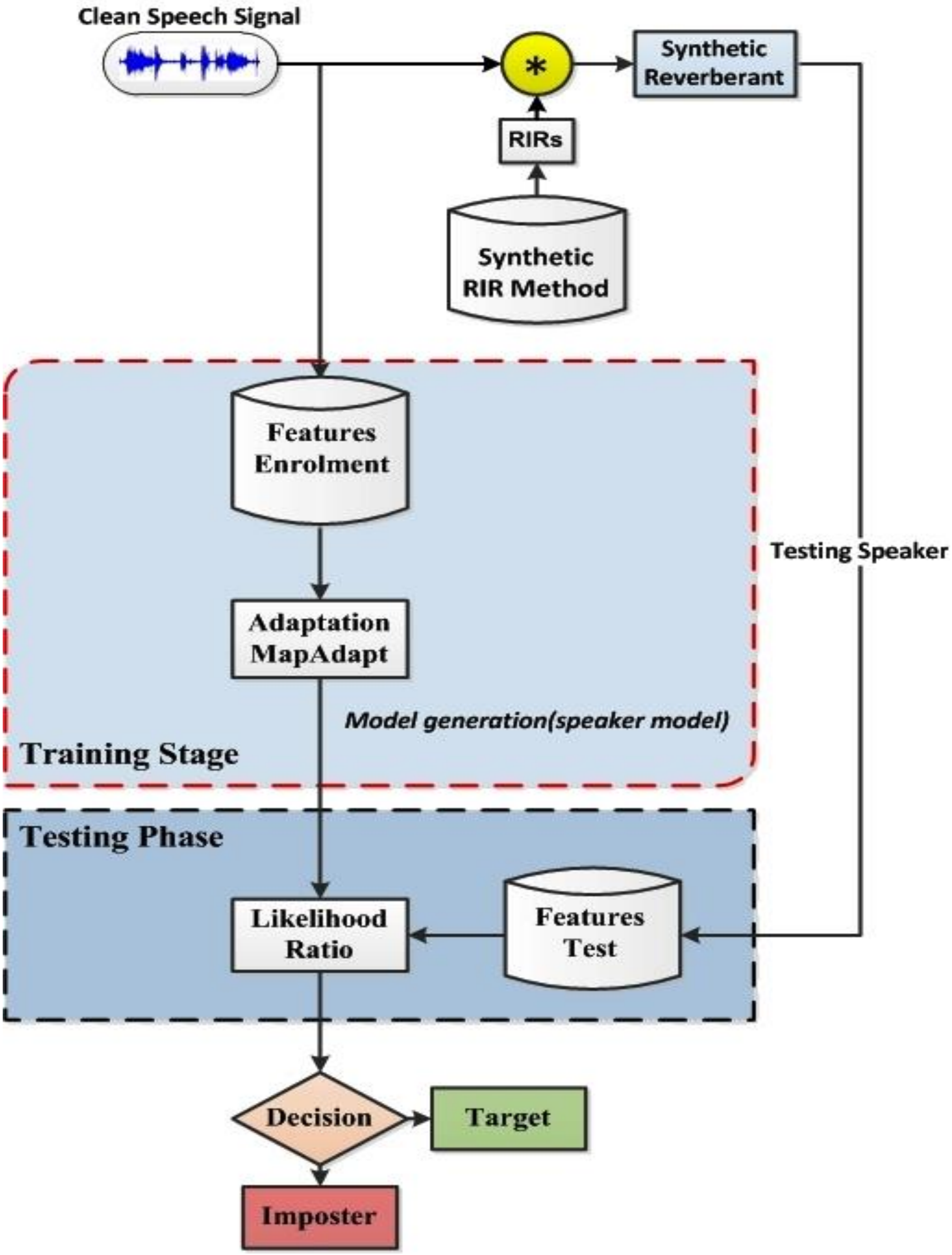


Figure 4.2 the framework of the proposed system

4.2.1 Experimental Setup

This work is presented from a study to implement and evaluate the behaviour of the typical speaker recognition system (MSR) toolbox in a reverberant environment.

A Gaussian Mixture Model was applied for creating reference models and classification with 256 mixtures. In addition, the mel-frequency cepstral coefficients (MFCC) were used as a feature space. Starting from the clean signals, a large variety of reverberant speech was created by means of convolution with synthetic room impulse response (RIRs). Several reverberant acoustic channels were created through the use of the image source method (ISM) (Allen & Berkley, 1979). In total 100 RIRs were randomly generated varying the enclosure properties, the microphone and source positions and the source radiation characteristics. A variation of the original image method was implemented to simulate the source directivity through a parameterised model by Eric A. Lehmann (Lehmann). The microphone is omnidirectional and can be located anywhere in the room. The source is randomly located. More details on ISM can be found in section 4.3.2.1. The system was evaluated using the SALU-AC dataset; the speech samples were obtained from 100 speakers (half male and half female) from these data sets.

Each speaker provided 20 utterances; each utterance has a 5s duration of approximately 50s in total duration; 18 utterances (90s) were selected for the training phase with the remaining 2 (10s) used for the testing phase to produce exclusive training and test data sets. The speech samples, which are used in testing stage, are different from those used in training stage. The purpose of all the experiments is not testing speaker recognition against the talker, but rather the primary goal of this study was to test speaker recognition against reverberant conditions. Therefore, limited speech samples have been used. To validate our work, the popular (MSR) toolbox is adopted. The training material comprises the original signals used for the clean baseline model and reverberant versions produced by convolution with the synthetic room impulse responses for the reverberant models, with reverberation time (0.11, 0.23, 0.53, 0.7, 1, 1.5, 2, and 2.5s).

Figure 4.3 demonstrates the reverberant speech signals produced by the ISM for T60=0.1, 0.5, and 1s. Table 4.1 shows the specifications of these rooms that are used to generate impulse responses.

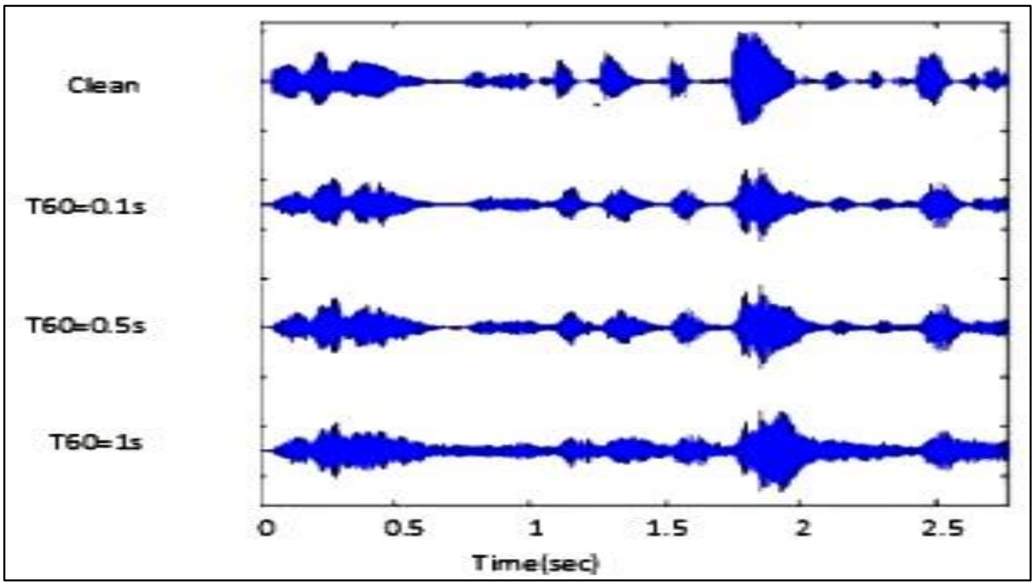


Figure 4.3 Waveforms, top to bottom: clean and reverberant speech to T60 =0.1, 0.5,1s

Table 4.1 Reverberation Specifications for experiment

Specification	Reverberation Model						
	Room1	Room2	Room3	Room4	Room5	Room6	Room 7
Room dimensions	3*4*2.5	5*4*3	6* 6*3	8*6*3	8*8*4	9*10*4	11* 12*4
Room volume	30 m^3	60 m^3	108 m^3	144 m^3	256 m^3	360m^3	528m^3
RT$_{60}$	0.11 , 0.3, 0.5 , 0.7 , 1, 1.5 , 2 s, 2.5s						

4.2.2 Results and Discussion

The present study was designed to determine the effect of reverberation time on a speaker verification system via the implemented MSR toolbox. “The result of this study has shown that MSR does not work well in reverberant conditions”. To quantify the relationship between the recognition results and the effects of reverberation, two types of speech sample were used. Baseline results using clean signals, i.e. Training, and testing with independent clean speech samples were established. A simple percentage error rate is not adequate to indicate the performance of the system since false acceptance, and false rejection has different impacts on the system. In speaker recognition and other biometric security systems, the EER is often used as a combined single measure for error. The introduction of the EER measure gives a more suitable tool for the evaluation of the performance of detection systems in general and speaker verification systems in particular. “The EER is the value where the false negative rate (FNR) and false positive rate (FPR) are equal”. “The lower the EER, the higher the reliability of a biometric system, therefore, Error equal rate (EER) is used as evaluation methods in this work”. The accuracy of the system using clean signals was 100%. However, the results of the present study also suggest that the system using reverberating samples recorded in the reverberation room with high reverberation time was only 15%. This result clearly shows that reverberation time has significant effects on speaker recognition. On the other hand, Figure 4.4 shows the system accuracy, using clean speech in the training stage and speech samples corrupted with different reverberation time from various rooms. As shown in this figure, there is a clear trend of decreasing system performance with increasing value of reverberation time and dimensions of the rooms. For instance, the system accuracy was (97.7%, 95.6%) using speech samples corrupted by RT=0.33s, 0.53s from the first and second room. However, the system efficiency degraded to 84.2% when the reverberation time increased to 1.5 seconds. “To conclude, the percentages of the system performance show significant degradation with an increase in the reverberation time and the room’s dimension”.

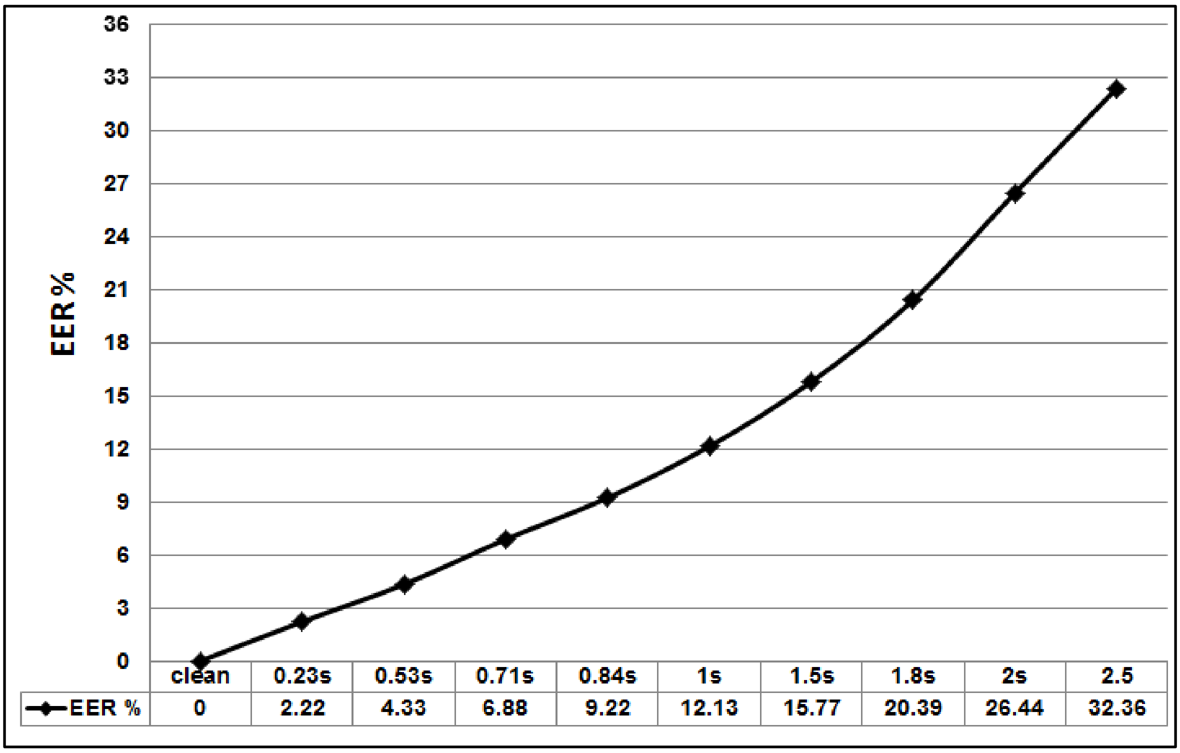


Figure 4.4 The speaker recognition system performance with different reverberation time from various rooms

4.3 Evaluation of the robustness of MFCC and GFCC Features in Reverberant Environments

Speech transmits information about the message to be conveyed, the speaker, and the language. Therefore, speech features must provide a sufficient representation of the speech signal. A number of sources can lead to redundant or inaccurate information being added to the speech signal, which affects speaker-specific information and design of a speaker model. Such sources comprise interference from the environment and distortions added by the transmission channel. In speech/speaker recognition tasks, it is required that the speech features represent the specifics of particular voice with sufficient accuracy.

Previous studies have primarily concentrated on using MFCC, GFCC features in noisy conditions, and there have been several studies in the literature reporting that GFCC is more robust than MFCC in noisy conditions. The objective of this study, therefore, aims to investigate the experimental robustness of both features in reverberant conditions. Performance in terms of equal error rate and detection trade-off plot under various reverberation times is quantified via simulation. Results from the study are presented and discussed.

4.3.1 Mel Frequency Cepstral Coefficients (MFCC)

The most popular feature extraction technique is Mel Frequency Cepstral Coefficients (MFCC), as it is less complex in implementation and more efficient and robust under various conditions and it is based on knowledge of the human auditory system (Poonkuzhali, Karthiprakash, Valarmathy, & Kalamani, 2013). Figure 4.5 shows the main steps of MFCC calculation.

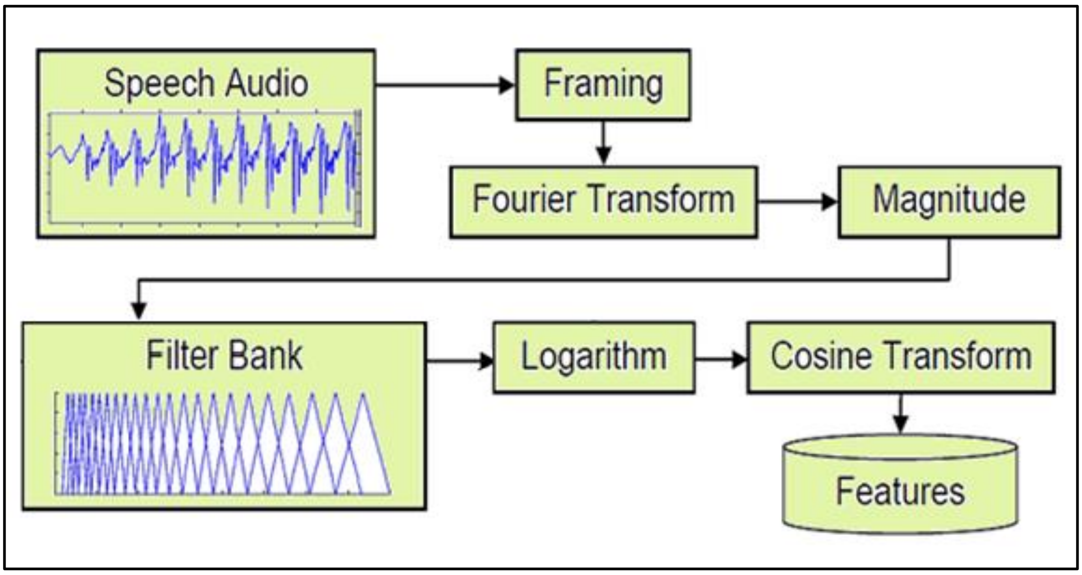


Figure 4.5 Main computation stages of MFCC

4.3.1.1 Pre-emphasis

Pre-emphasis refers to filtering that emphasises the higher frequencies. Its purpose is to balance the spectrum of voiced sounds that have a steep roll-off in the high-frequency region. For voiced sounds, the glottal source has an approximately -12 dB/octave slope (Picone, 1993; Rabiner & Schafer, 1978). However, when the acoustic energy radiates from the lips, this causes a roughly +6 dB/octave boost to the spectrum. As a result, a speech signal when recorded with a microphone from a distance has approximately a -6 dB/octave slope downward compared to the true spectrum of the vocal tract. Therefore, pre-emphasis removes some of the glottal effects from the vocal tract parameters. The most commonly used pre-emphasis filter is given by the following transfer function (Brümmer & du Preez, 2006).

$$H_p(z) = 1 - \alpha z^{-1} \qquad 4.1$$

where the value of α controls the slope of the filter and is usually between 0.4 and 1.0 (Picone, 1993). Figure 4.6 shows the power spectral density of a speech waveform before and after applying pre-emphasis. We can notice in the original signal the drop at higher frequencies compared to the pre-emphasised signal using α = 0.95, in which the power is better distributed across the relative frequencies. This comparison can also be seen using the spectrograms in Figure 4.7. Notice that the high frequencies are more prevalent in the pre-emphasised signal. Furthermore, Voice activity detection (VAD) is an essential step in processing a speech signal for speaker recognition. An energy-based method was suggested by Reynolds (D. A. Reynolds, 1995). It is used for the detection of speech. In simple cases, the audio, segmented into short frames, is used to compute the frame energy level. An energy threshold is defined which is sufficient to decide whether the frame contains speech or not.

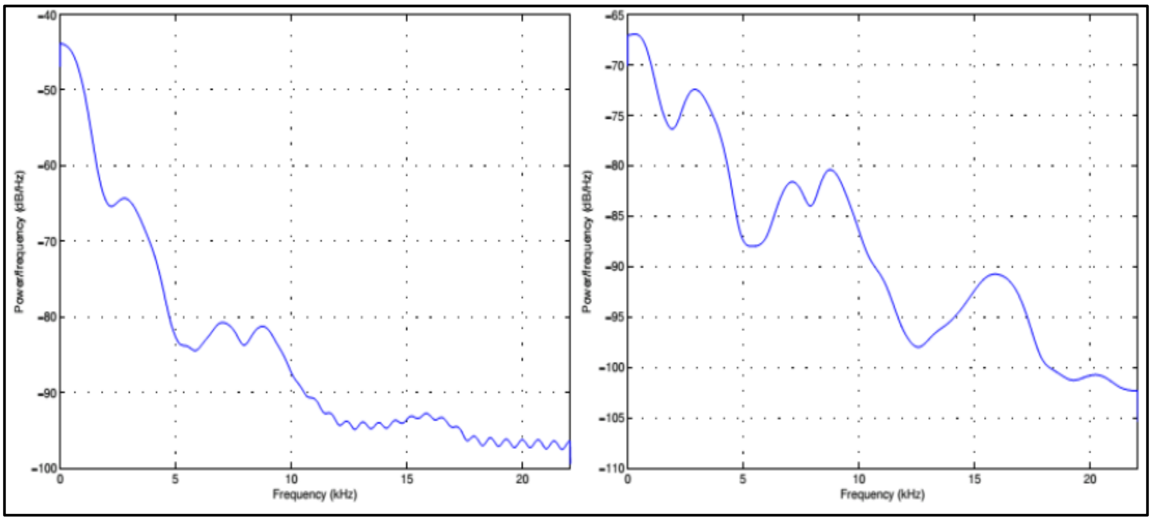


Figure 4.6 Power spectral density speech signal sampled at 44100 Hz before/after pre-emphasis (Beigi, 2011)

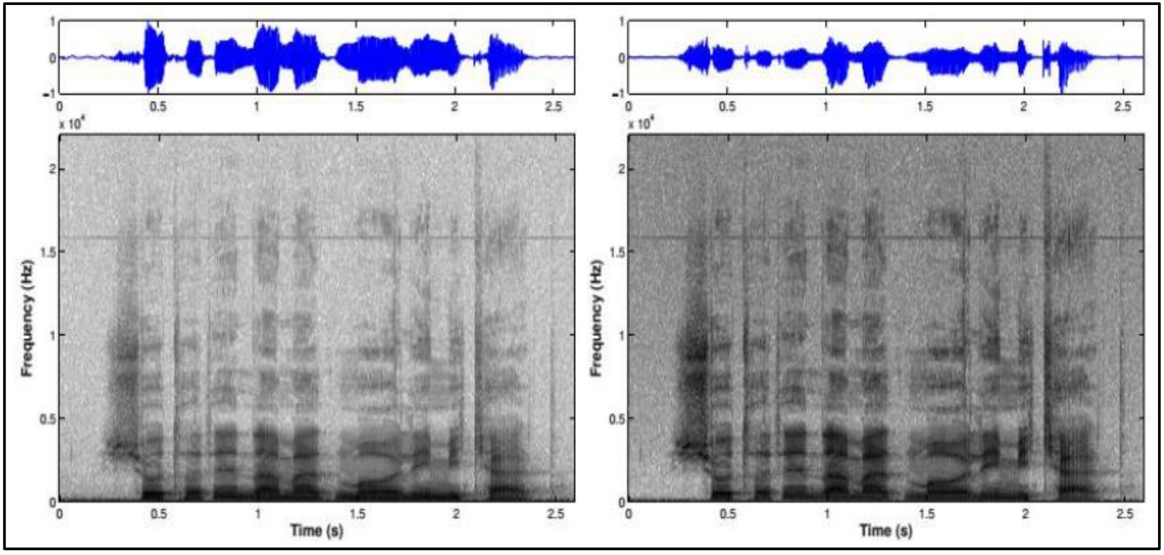


Figure 4.7 Spectrogram of a speech signal sampled at 44100 Hz before and after pre-emphasis (Beigi, 2011)

4.3.1.2 Frame Blocking and Windowing

A speech signal is a slowly time-varying or quasi-stationary signal. For stable acoustic characteristics, speech needs to be examined over a sufficiently short period. Therefore, speech analysis must always be carried out on short segments across which the speech signal is assumed to be stationary. Short-term spectral measurements are typically carried out over 20 ms windows, and advanced every 10 ms (Benesty, Sondhi, & Huang, 2008; Deller Jr, Proakis, & Hansen, 1993; Tomi Kinnunen & Li, 2010). Advancing the time window every 10 ms enables the temporal characteristics of individual speech sound to be tracked, and the 20 ms analysis window is usually sufficient to provide reasonable spectral resolution of these sounds and at the same time short enough to resolve significant temporal characteristics. Figure 4.8 shows an example of framing. The purpose of the overlapping analysis is that each speech sound of the input sequence would be approximately centred in some frame. On each frame, a window is applied to taper the signal towards the frame boundaries. Hanning or Hamming windows are used because this prevents any of the sharp edges seen from similar rectangular windows (Picone, 1993), and helps to enhance the harmonics, smooth the edges and to reduce the edge effect while taking the DFT on the signal. Equation 4.2 defines the Hamming window

$$w[n] = \begin{cases} 0.54 - 0.46\ \cos\dfrac{2\pi\, n}{L} & 0\ \le n \le \text{L-1} \\ 0 & \end{cases} \qquad 4.2$$

Figure 4.9 demonstrates the window at the time, and frequency domains and Figure 4.10 shows the original and windowed speech signal.

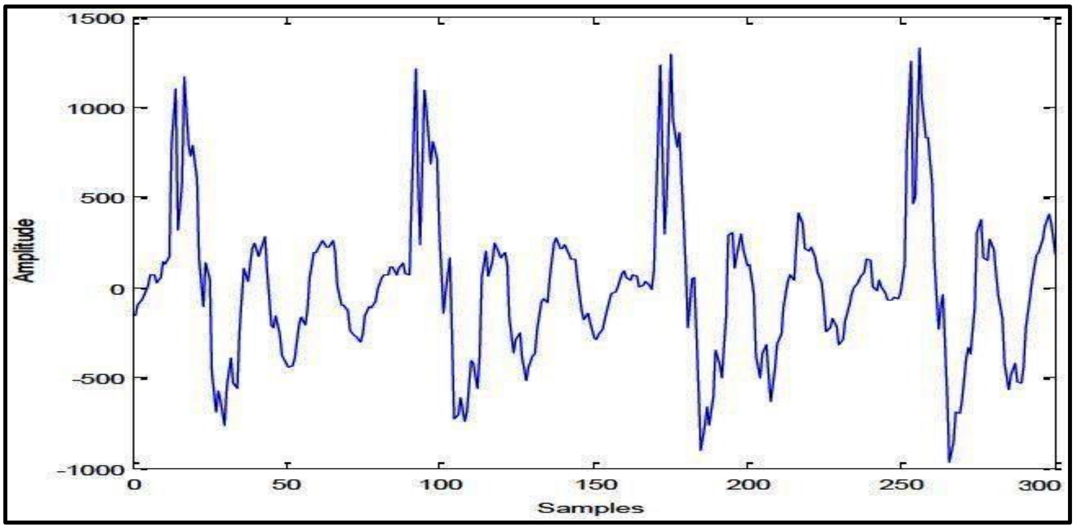


Figure 4.8 Example of framing

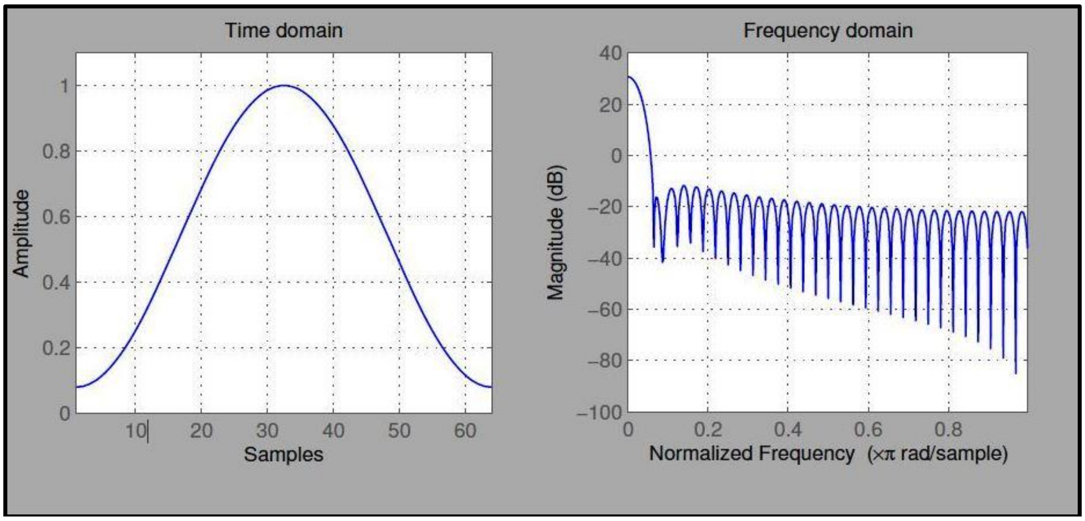


Figure 4.9 Hamming window (Beigi, 2011)

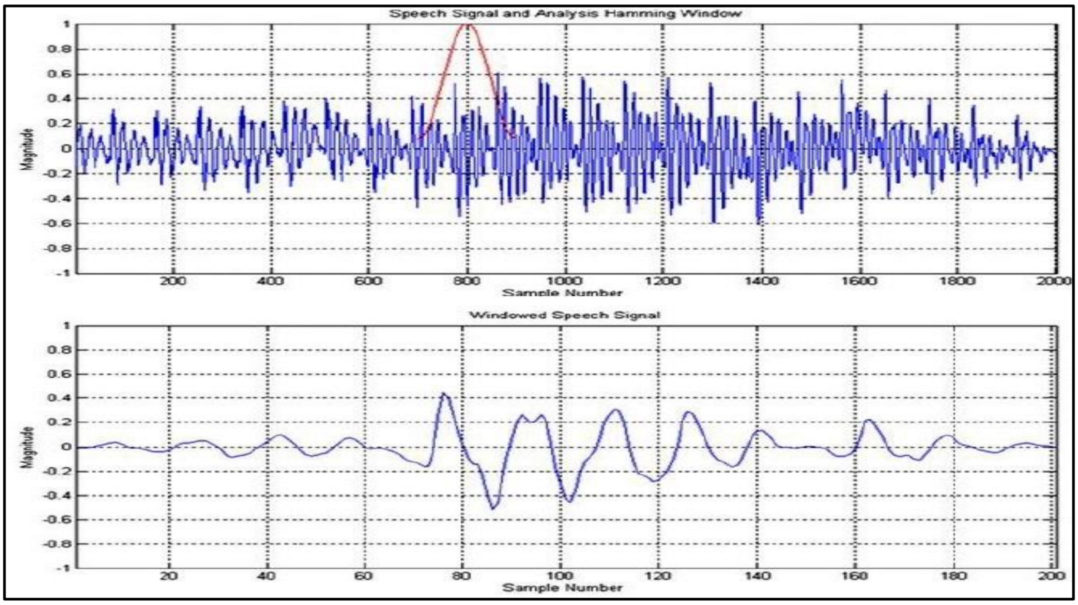


Figure 4.10 Original and Windowed Speech Signal

4.3.1.3 Discrete Fourier Transform(DFT) Spectrum

Spectral analysis shows that different timbres in speech signals correspond to the different energy distribution over frequencies. Therefore, a Discrete Fourier Transform (DFT) is performed to obtain the frequency magnitude response of each frame. When we perform a DFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapped (Kurzekar, Deshmukh, Waghmare, & Shrishrimal, 2014). Figure 4.11 shows an example of applying a Discrete Fourier Transform (DFT).

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}} \qquad 0 \le k \le N\text{-}1 \qquad 4.3$$

where N is the number of points used to compute the DFT.

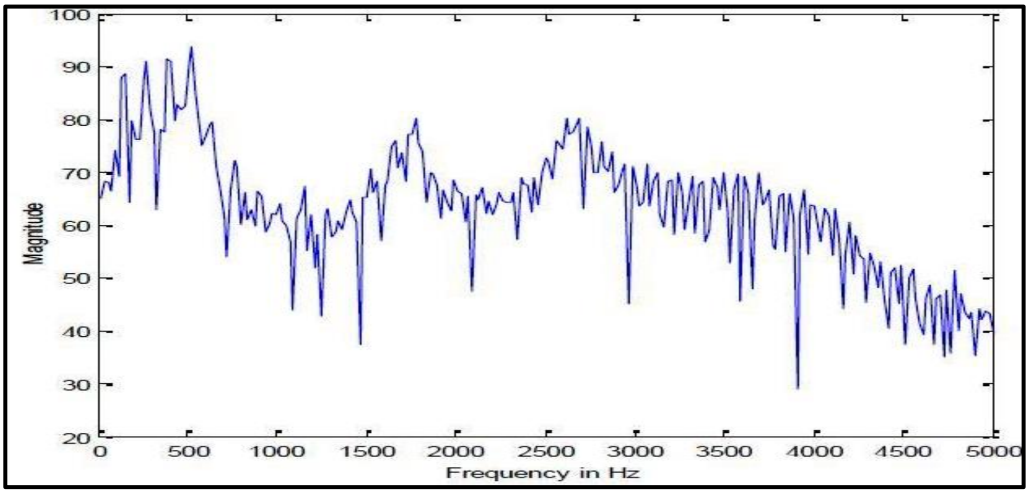


Figure 4.11 Example of DFT

4.3.1.4 Mel-Spectrum

Mel-Spectrum is calculated by passing the Fourier transformed signal through a set of band-pass filters known as a mel-filter bank. The Mel scale relates the perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely, what humans hear. The mel scale is approximately a linear frequency spacing below 1kHz, and a logarithmic spacing above 1kHz (S. S. Stevens, Volkmann, & Newman, 1937). The approximation of mel from physical frequency can be expressed as

$$f_{\mathrm{mel}} = 2595 \log_{10}(1 + \frac{f}{700}) \qquad 4.4$$

where f denotes the physical frequency in Hz and f_{mel} denotes the perceived frequency (Deller Jr et al., 1993). Filter banks can be implemented in both time domain and frequency domain. For MFCC computation, filter banks are generally performed in the frequency domain. The centre frequencies of the filters usually are equally spaced on the frequency axis. The most commonly used window functions are triangular. However, in some cases, the Hanning windows are used (Picone, 1993). The triangular windowed filter banks with mel-frequency warping are given in Figure 4.12. The mel spectrum of the magnitude spectrum $X(k)$ is computed by multiplying the magnitude spectrum by each of the triangular mel weighting

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 H_m(k)] \; ; \qquad 0 \le \mathrm{m} \le M\text{-}1 \qquad 4.5$$

where M is the total number of triangular mel weighting filters(Ganchev, Fakotakis, & Kokkinakis, 2005; Zheng, Zhang, & Song, 2001). $H_m(k)$ denotes the weight given to the *kth* energy spectrum bin contributing to the mth output band.

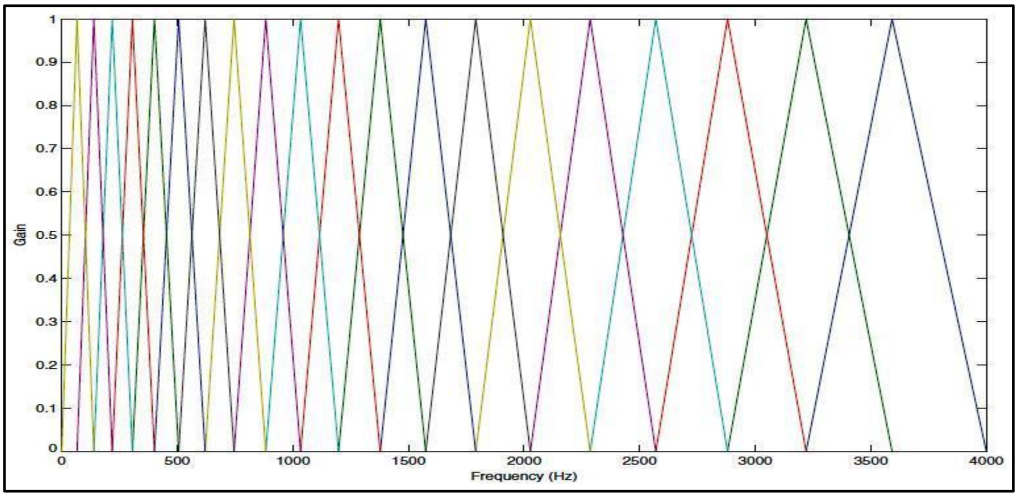


Figure 4.12 Shape of the Mel filter bank

4.3.1.5 Discrete Cosine Transform (DCT)

Since the vocal tract is smooth, the energy levels in adjacent bands tend to be correlated. The Discrete Cosine Transform (DCT) is applied to the transformed mel frequency coefficients to produce a set of cepstral coefficients. Prior to computing Discrete Cosine Transform (DCT), the mel spectrum is usually represented on a log scale. This results in a signal in the cepstral domain with a frequency peak corresponding to the pitch of the signal and a number of formants representing low-frequency peaks. Since most of the signal information is represented by the first few MFCC coefficients, the system can be made robust by extracting only those coefficients by ignoring or truncating higher order Discrete Cosine Transform (DCT) components (Picone, 1993). Finally, MFCC is calculated as (Picone, 1993).

$$c_n = \sum_{k-1}^{k} (\log D_k) \cos\left[m(k - \frac{1}{2}) \frac{\pi}{k} \right]; \quad \text{m=0, 1, 2… k-1} \qquad 4.6$$

where c_n represents the MFCC and *m* is the number of the coefficients here m=13 so, total number of coefficients extracted from each frame is 13. Traditional MFCC systems use only 8–13 cepstral coefficients. The zeroth coefficient is often excluded since it represents the average log-energy of the input signal, which only carries little speaker-specific information(Gupta, Jaafar, Ahmad, & Bansal, 2013).

4.3.2 Gammatone Frequency Cepstral Coefficients

The ability of humans to achieve speaker recognition in noisy and reverberant environments has motivated research into the robust speaker recognition from the viewpoint of computational auditory scene analysis (D. A. Reynolds, 1994). Furthermore, the sensitivity to additive noise condition is one of the main drawbacks of MFCC, which prompted researchers to search for a more robust feature to alleviate these drawbacks. Auditory features were proposed firstly by (Shao et al., 2007; Shao & Wang, 2008) to improve the robustness of speaker recognition for noisy speech. This type of feature simulated the auditory process of the human ear since the filters used to extract these features are based on the psychophysical observation of the total auditory system known as gammatone filter bank (Beigi, 2012; Shao & Wang, 2008). This filterbank consists of 128 filters (or sometimes 64 filters) centred on the frequencies that are quasi-logarithmically spaced from 50Hz to 8 kHz which model the human cochlear (Shao et al., 2007). The Gammatone filter bank is responsible for decomposing the input signal into the time-frequency (T-F) domains, which represents the difference of Cochleagram (D. Wang & Brown, 2006). Cochleagram keeps the higher frequency resolution at the low-frequency range for the same number of frequency components, which makes it different from the linear frequency resolution of the spectrogram. The time frame of Cochleagram is known as the gammatone feature (GF). Finally, discrete cosine transform (DCT) is applied to GF to reduce dimensionality and de-correlate the components, and the results of this reduction are known as Gammatone frequency cepstral coefficients. Figure 4.13 shows the framework of the main steps of the Gammatone Frequency Cepstral Coefficients (GFCC) feature calculation.

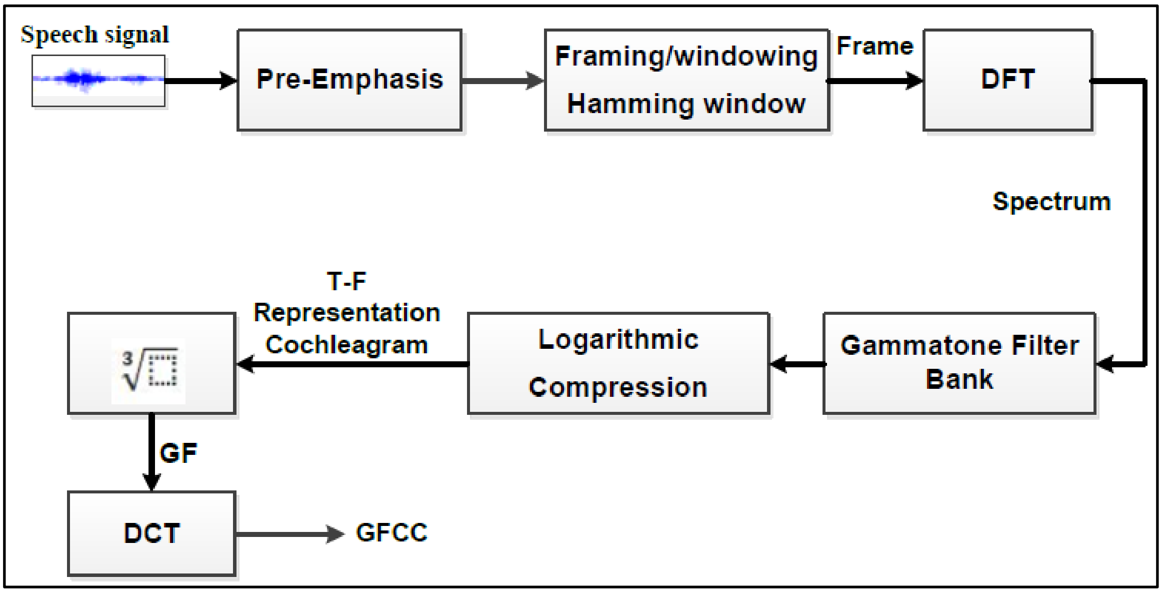


Figure 4.13 Calculation of the GFCC parameters

4.3.2.1 Gammatone Filter-Bank

The Gammatone filters are used to simulate the operation of the human auditory system. A gammatone filter with a central frequency f_c can be defined as:

$$g(t) = at^{n-1}e^{2\pi bt}\cos(2\pi f_C + \varphi) \qquad 4.7$$

Where $\boldsymbol{\varphi}$ refers to the phase (but is generally set to zero), the constant *a* controls the order of the filter is defined by the value *n* which is normally set to a value less than 4. *b* is the decay factor and defined as

$$b = 25.17\left(\frac{4.37\,fc}{1000} + 1\right) \qquad 4.8$$

4.3.2.2 Pre-emphasis

This step is similar to the pre-emphasis phase of the MFCC counterpart. It is used to help decrease the dynamic range and to accentuate the frequency components that hold most of the key information required for the speech signal. The pre-emphasis is defined as a second order filter as follows:

$$H(z) = 1 + 4e^{-2\pi b/\,f_S}z^{-1} + e^{-2\pi b/\,f_S}z^{-2} \qquad 4.9$$

where *b* is defined in (4.8), and *fs* is the sampling frequency. Figure 4.14 shows the gammatone filter output after applying pre-emphasis filter.

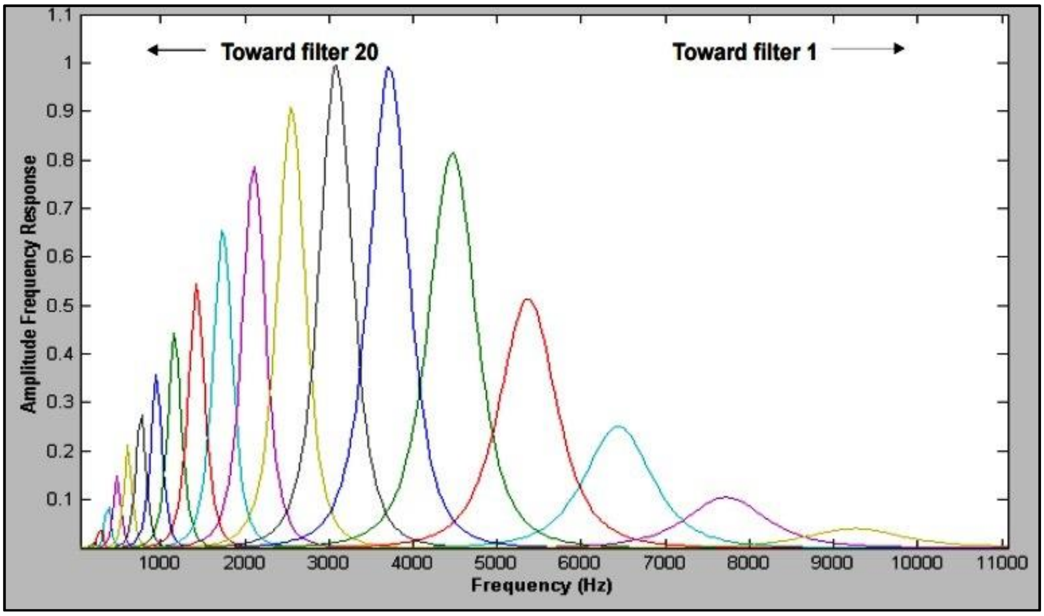


Figure 4.14 gammatone filter output after applying pre-emphasis filter (Abdulla, 2002)

4.3.3 Differences between GFCC and MFCC

There are several differences between MFCC and GFCC features. Firstly, there are differences in frequency scaling. GFCC, based on Equivalent Rectangular Bandwidth (ERB) scale, has a finer resolution at low frequencies than MFCC (mel scale). Secondly, there are differences in the nonlinear rectification step prior to the DCT. MFCC utilises a log while GFCC utilises a cubic root. Furthermore, the log operation transforms convolution between the excitation source and vocal tract (filter) in addition to the spectral domain. Besides these two major differences, some other notable differences are summarised in Table 4.2 (Zhao & Wang, 2013).

Table 4.2 Differences between MFCC and GFCC

Category	**MFCC**	**GFCC**
Pre-emphasis	Yes	Yes\ No
Frequency bands	26-39	64
Frequency scaling	Mel-Scale	EBR
Nonlinear Rectification	Logarithmic	Cubic root \Logarithmic
Intermediate T-F representation	Mel-Spectrum	Variant of Cochleagram

4.3.4 Experimental Setup

This experiment aims to investigate the robustness of two common features used with speaker verification in a reverberant environment. Both features have been extracted from the same speech signal and written to the disc in HTK format. The output of the MFCC and GFCC are used as a new vector. The system is divided into two subsystems; both systems will be employed in parallel via training and testing. To make a fair comparison, MFCCs and GFCCs used in these experiments were both 22-dimensional. A Gaussian mixture model was applied for creating reference models and classification with 256 mixtures. In this experiment, the system was evaluated using the SALU-AC data set; the speech samples were obtained from 100 speakers (50 male and 50 female) from this data set. Each speaker provided 20 utterances; each utterance has a 5s duration of approximately 50s in total duration; 18 utterances (90s) were selected for the training phase with the remaining 2 (10s) used for the testing phase to produce exclusive training and test data sets. The speech samples, which are used in testing stage, are different from those used in the training stage. The responses, however, were simulated using the commercial software CATT-Acoustic, which is employed to generate synthetic room impulse responses (RIRs) with T_{60} from differently reverberating rooms. A large variety of reverberant speech was created by means of convolution with synthetic room impulse responses. Note that the test utterances are different from the training ones. In addition, simple energy-based voice activity detection was applied to remove the large chunks of silence in the excerpt, and the speech was sampled at 16 kHz.” Error equal rate (EER) and detection trade-off curve (DET) are used as evaluation methods in this work; EER is the most widely used performance measure for speaker verification systems. The system accuracy, using both features in the clean environment was 0% EER. The purpose of all the experiments is not testing speaker recognition against the talker, but rather the primary goal of this study was to test speaker recognition against reverberant conditions. Therefore, limited speech samples have been used”.

4.3.5 Result and Discussion

The findings of the present study suggest that the GFCC feature outperformed traditional MFCC features under different reverberation times. Moreover, to show the robustness of a speaker verification system, both features were examined using clean speech in the training stage, while the speech was contaminated with different reverberation time in the testing stage. Figure 4.15 depicts the box plot obtained with the MFCC and GFCC setups, respectively. This simplest possible box plot displays the full range of variation (from min to max), and it corresponds to the standard deviation according to the percentage EER for each feature performance with different reverberation time. In this figure, the MFCC and GFCC appear to have closer centres. “The box plot indicated that the GFCC feature demonstrates better performance compared with the MFCC as it has a lower median, of around 3.2”. Table 4.3 shows the details of the EER % value of each of the features investigated using both data sets. Furthermore, Figure 4.16 shows the system performance using the MFCC and GFCC features. In this figure, the x-axis represents the reverberation time level, and the y-axis represents the EER %.” As shown in these figures, it is noticeable that the GFCC was more robust intrinsically as features in reverberation condition”. The robustness of a speaker verification system is increased when reverberation time tends to be longer. For example, with RT=0.23s, 0.53,0.71 and 0.8s, the accuracy of the system using the GFCC feature as determined by the EER is 1.74%, 3.11, 4.66 and 9.22% compared with 2.22%, 4.33%, 6.88%, and 9.22% for MFCC. However, the performance of the GFCC shows steady degradation when the reverberation time is increased to 2s, and the performance for both features becomes close. The highest improvement against MFCC was found in the reverberation time range 0.53s to 1.8s. Furthermore, Figure 4.17, and Figure 4.18 show the DET curves for both features with different reverberation time values. These figures indicated that the false positive rates (FPR) for the GFCC feature are less than MFCC; even the False Negative Rate for both features is close. More explanation of results depending on the DET curves with different reverberation times is shown in Appendix A.

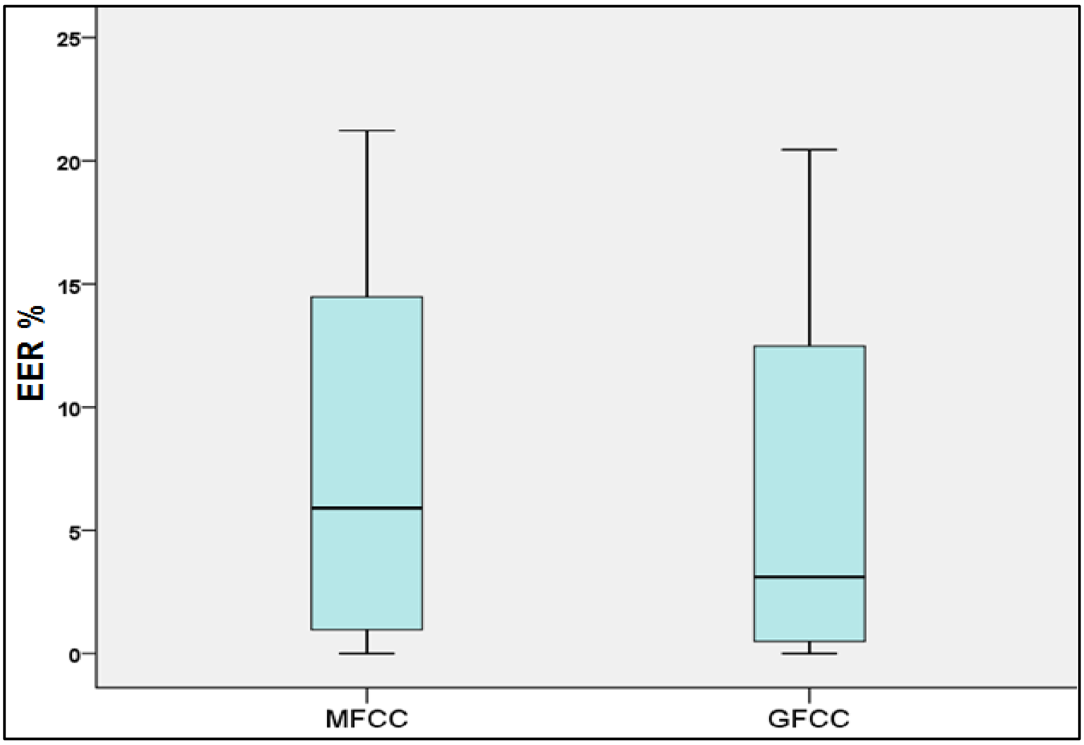
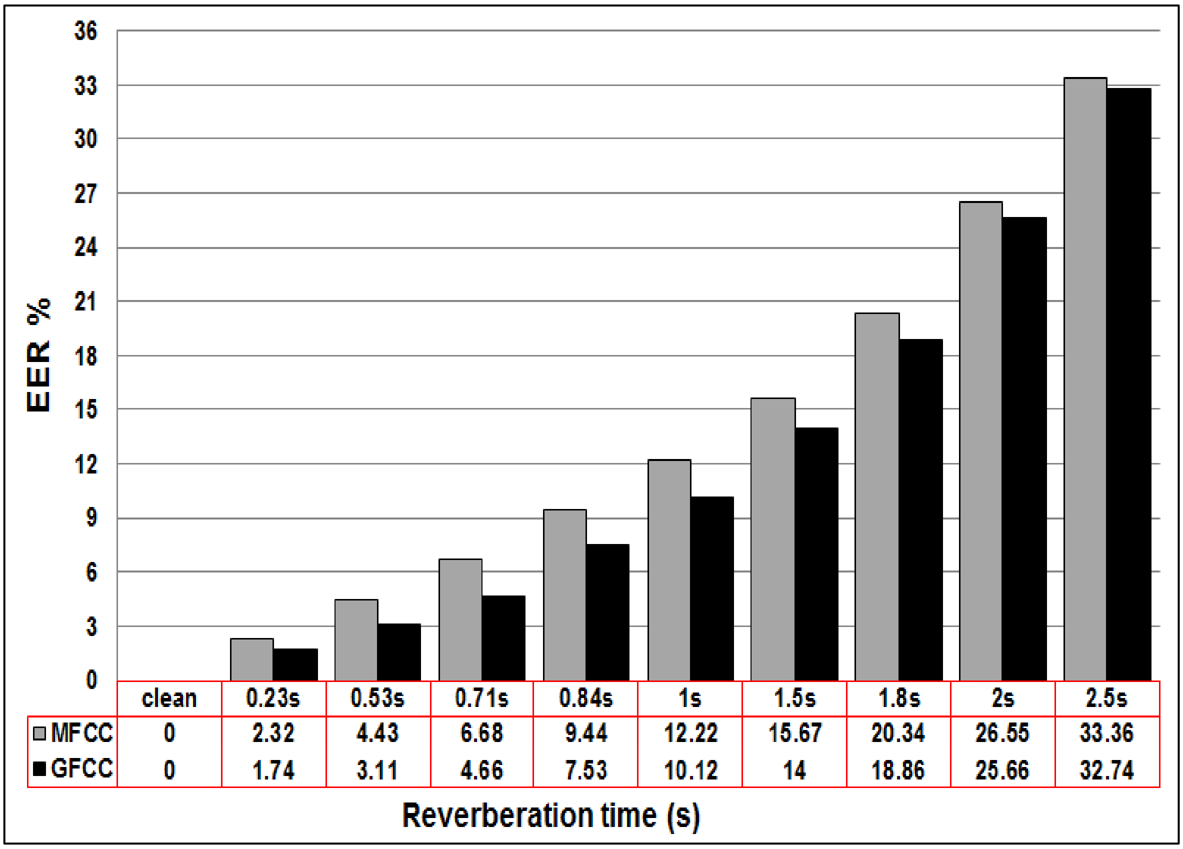


Figure 4.15 Box plots of EER (%) over the two features' results using different reverberation time

Table 4.3 Summary of the EER for both features with different RT

RT(s)	EER % Using SALU-AC database	
	MFCC	GFCC
Clean	0	0
0.23	2.32	1.74
0.53	4.43	3.11
0.71	6.68	4.66
0.84	9.44	7.53
1	12.22	10.12
1.5	15.67	14
1.8	20.34	18.86
2	26.55	25.66
2.5	33.36	32.74



	clean	0.23s	0.53s	0.71s	0.84s	1s	1.5s	1.8s	2s	2.5s
□ MFCC	0	2.32	4.43	6.68	9.44	12.22	15.67	20.34	26.55	33.36
■ GFCC	0	1.74	3.11	4.66	7.53	10.12	14	18.86	25.66	32.74

Figure 4.16 EER (%) for both features with different reverberation times using the SALU-AC dataset

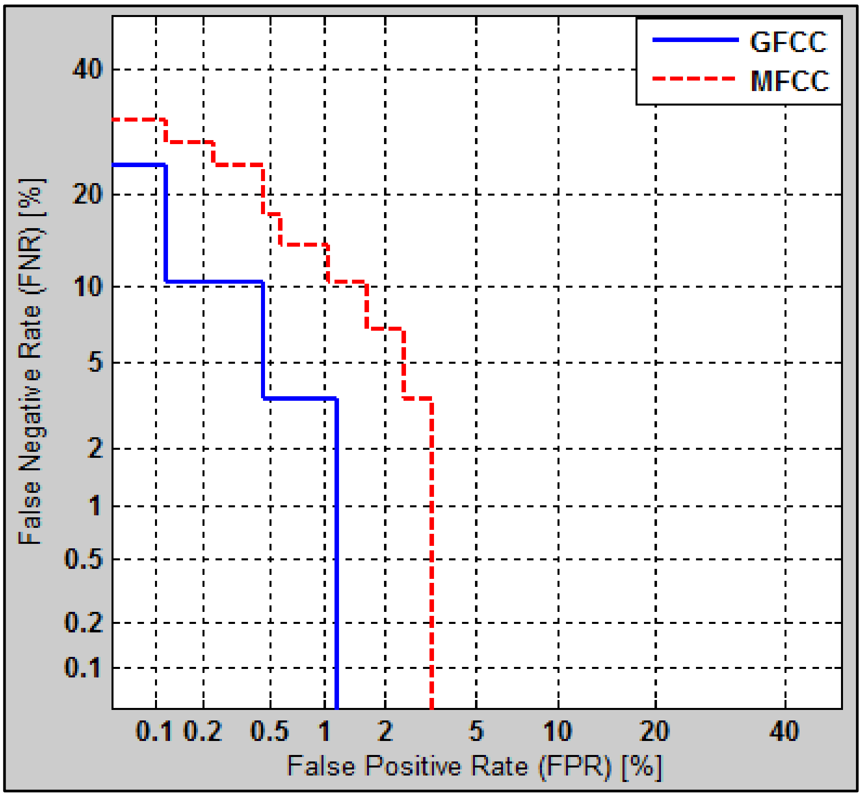


Figure 4.17 DET plot with RT=0.53s using the SALU-AC database

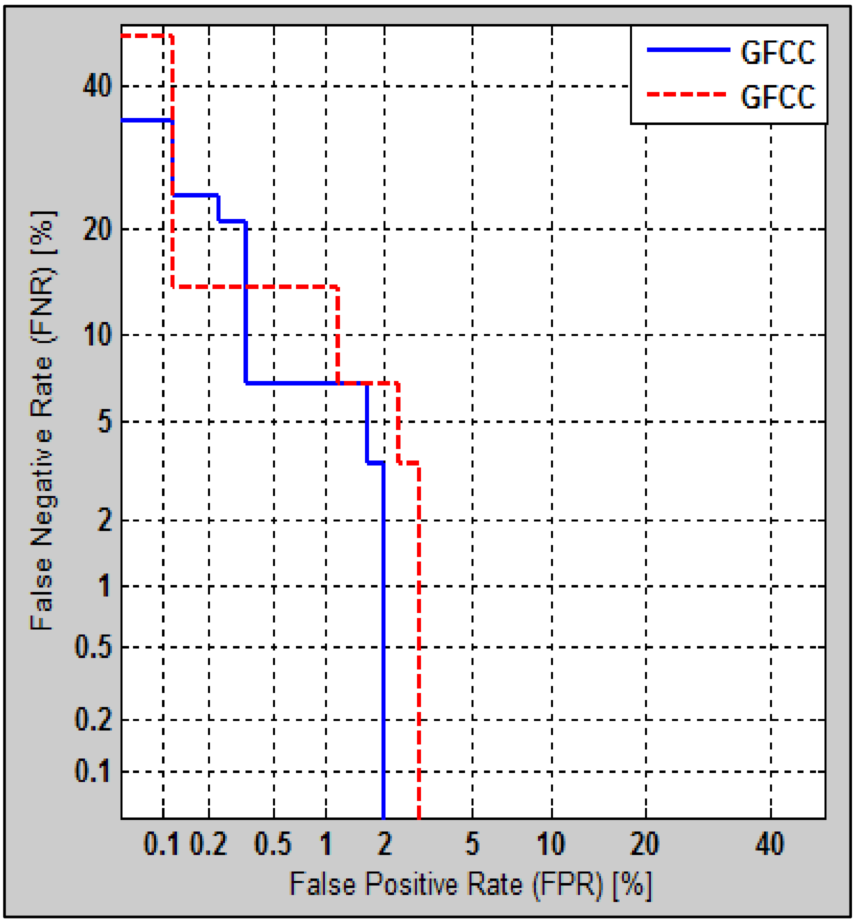


Figure 4.18 DET plot with RT=0.71s using the SALU-AC database

4.4 Evaluation the Effect of Reverberation Ttime and Source-Receiver Distance on the Performance of Speaker Recognition

Speech recorded by a distant microphone in a room may be subject to reverberation. Although there has been much research about the effect of reverberation time on speaker recognition performance, none of the previous studies investigated the suitability of the reverberation time and the distance as room acoustic parameter, which is directly related to the degradation in the speaker recognition performance in reverberant environments.” In this experiment, the performance of speaker recognition in reverberant environments has been mainly associated to the reverberation time (T_{60})and to the distance between the source and the receivers, keeping all the other elements affecting the room impulse response fixed (source directivity and orientation, room sizes and wall absorption coefficients)”. The correlations between these parameters play a crucial role in the overall performance of speaker verification systems. “The results of this experiment confirmed that both the reverberation time and the source to receiver distance could affect the system performance”.

4.4.1 Test Methodology

Firstly, it is necessary to investigate the correlation of room acoustic parameters to ASR performance apart from the RT and the source-receiver distance. For this, a dataset of simulated RIRs with RT values ranging from 0.33 to 2s with different source-receiver distances was created. The commercial software CATT-Acoustic was employed to generate and synthetic impulse responses from different rooms with different dimensions and acoustic properties. These RT values of the selected rooms are representative of a wide range of typical acoustic scenarios, and taking into account that the ASR deteriorates for distant speech applications, eight different distances were used in each room. Source-receiver distances of 0.5, 1, 2, 3, 4, 5, 6 and 7 m were used. This can be considered a representative range of speaker–microphone distances in typical far-microphone recordings. In this experiment, the performance of speaker recognition in reverberant environments has been mainly associated to the reverberation time T_{60} and to the distance between the source and the microphones, keeping all the other factors affecting the RIR fixed (source directivity and orientation, room dimensions and wall absorption coefficients). For all the above cases, speech signals derived from the SALU-AC data set were convolved with the artificial RIRs to represent the corresponding reverberant signals. Note that the test utterances are different from utterances used for training. In addition, simple energy-based voice activity detection was applied to remove the large chunks of silence in the excerpt. The speech was sampled at 16 kHz. "The performance of the speaker recognition system was evaluated using EER and detection trade-off curve (DET) as evaluation methods in this work. Figure 4.19 shows the block diagram of the experimental setup".

4.4.2 Result and Discussion

"The results provide confirmation that both reverberations time and source to receiver distance can affect the system performance". In order to measure the strength and direction of the relations between the percentage EER and both reverberation time and the distance, we used the Pearson's correlation in the SPSS software. The Pearson's correlation presents a sample correlation coefficient; it has a value between (+1 and −1). The Pearson's correlation between the EER (%) and both RT and the distance are shown in Table 4.4, and this indicates that the EER (%) has a strong correlation with RT ($R = 0.966$, $p = 0.000$). The other factor (source-receiver distance) shows a strong correlation ($R = 0.899$, $p = 0.006$) with an EER (%). That means the EER (%) is affected by reverberation time and distance.

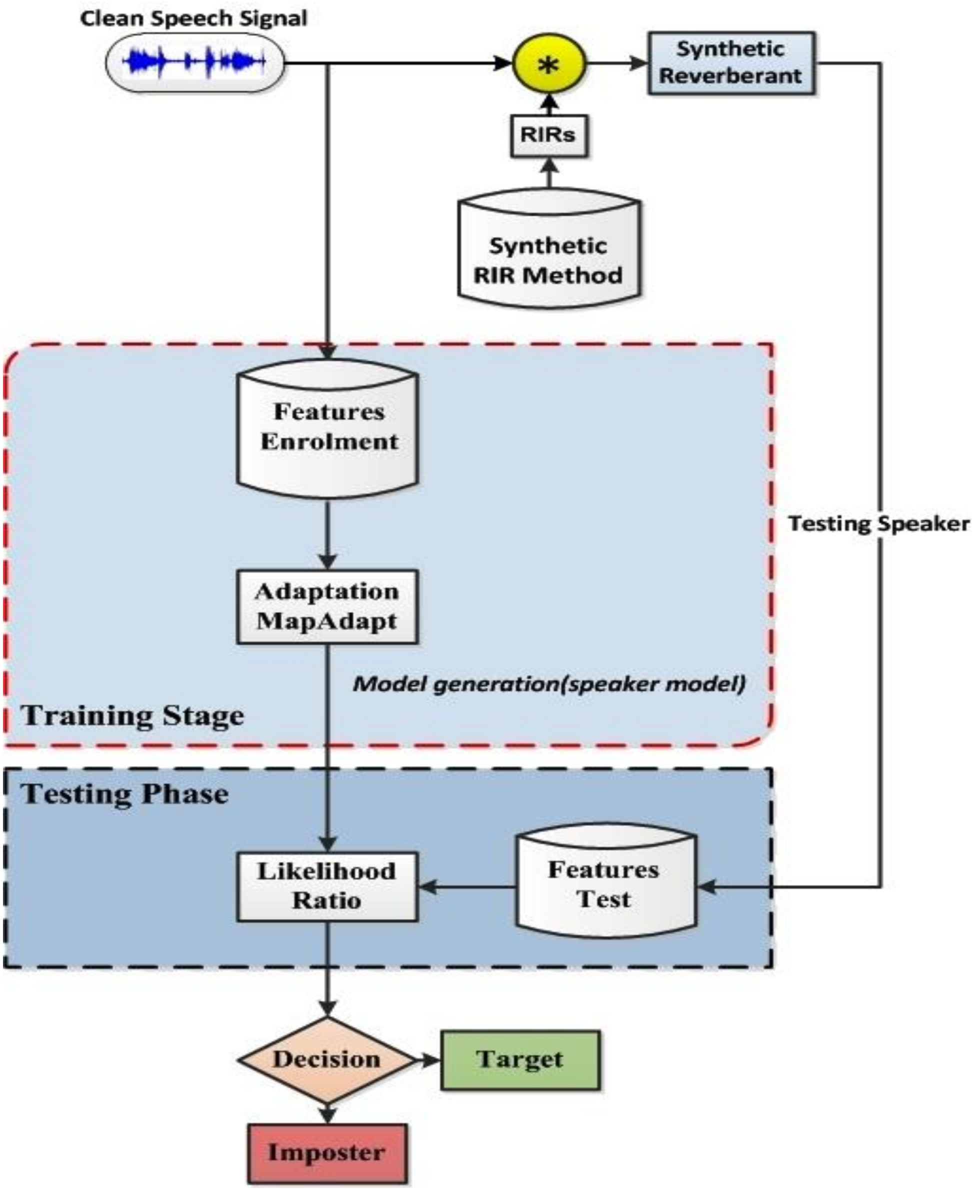


Figure 4.19 Block diagram of the experimental setup

In this case the better system performance will be associated with lower RIGHT and lower distance. The effect of both factors on the percentage EER is conducted by comparing the Standardised Beta Coefficient of the model, as shown in Table 4.5.

Table 4.4 The correlation between RT, DRR and EER (%)

	EER (%)	
	EER (%)	Correlation
RT Pearson Correlation Sig. (2-tailed)	.926 .000	Strong
Distance Pearson Correlation Sig. (2-tailed)	.899 .006	Strong

Table 4.5 Linear of prediction EER according to the RT and the distance

Model	Unstandardized Coefficients		Standardised Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant) RT Distance	-4.827 10.620 .952	.370 .242 .059	 .926 .899	-13.042 43.952 16.074	.000 .000 .000
a. Dependent Variable: EER %					

The effect of both factors on the percentage EER is conducted by comparing the Standardised Beta Coefficient of the model, as shown in Table 4.5. A positive value indicates that the factor causes degradation of the EER (%). Figure 4.20 shows the fitted line of the EER (%) and reverberation time (s). In addition, Figure 4.21 shows the fitted line of the EER (%) and source-receiver distance. From these figures, it is clear that both the reverberation time and the distance are strongly correlated with an EER (%). Moreover, Figure 4.22 shows the system performance acquired with different reverberation times and different distances. In this figure, the x-axis represents the reverberation time, and the y-axis represents the source to receiver distance, while the z-axis represents the percentage equal error rate. “It appears from these figures that both reverberation time and distance can affect the recognition performance”. For example, with RT=0.53s and distance 0.5cm, the EER is 2.33%. This percentage increased to reach 6% when the distance was increased to 7m with a reverberation time from the same room. Furthermore, the percentage EER increases to 14.66% when the reverberation time becomes 1.5s and the distance 5m. This result suggests that the system performs with high accuracy when the source is close to the receiver, as the direct sound is dominant compared to any reflections. In this case, the reverberation time level is low. In the result also confirmed that the system performs low accuracy when the source is far from the receiver, as the direct sound is not dominant compared to any reflections. In this case, the reverberation time level is high. The result is also confirmed that the effect of the source to receiver distance on the system performance is clear and should be taken into account. Moreover, the finding also confirms that degradation in system performance is more likely in larger rooms than small and medium rooms. “This finding verifies the indication that distance in large rooms more strongly affects system performance than in small and medium rooms”.

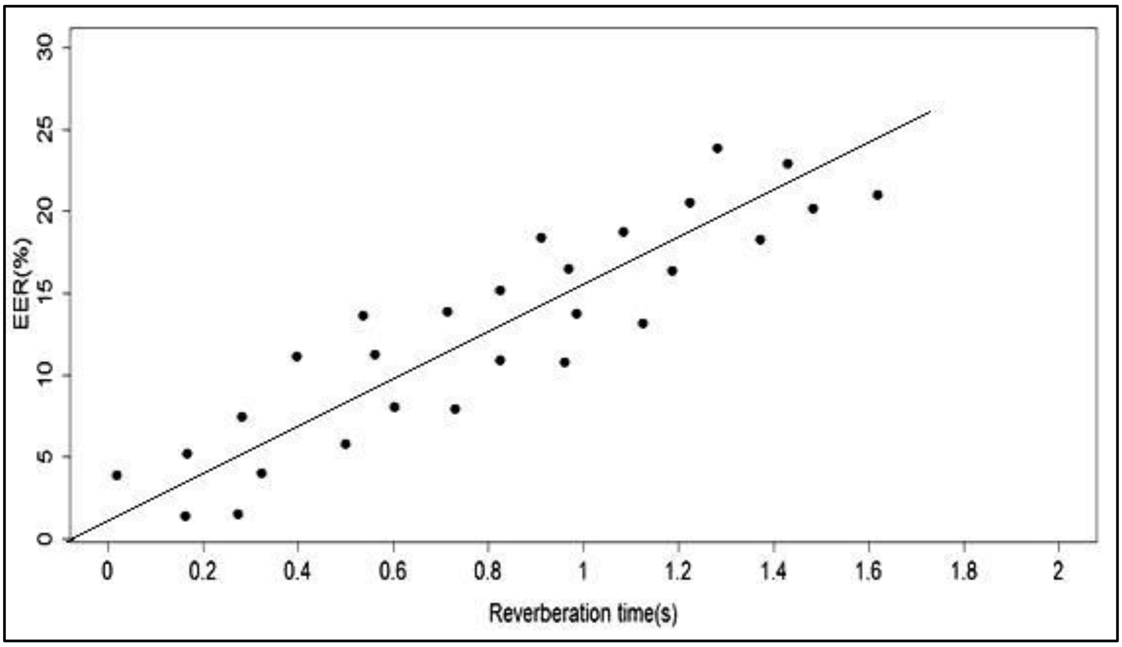


Figure 4.20 Fitted line of the EER (%) and reverberation time (s)

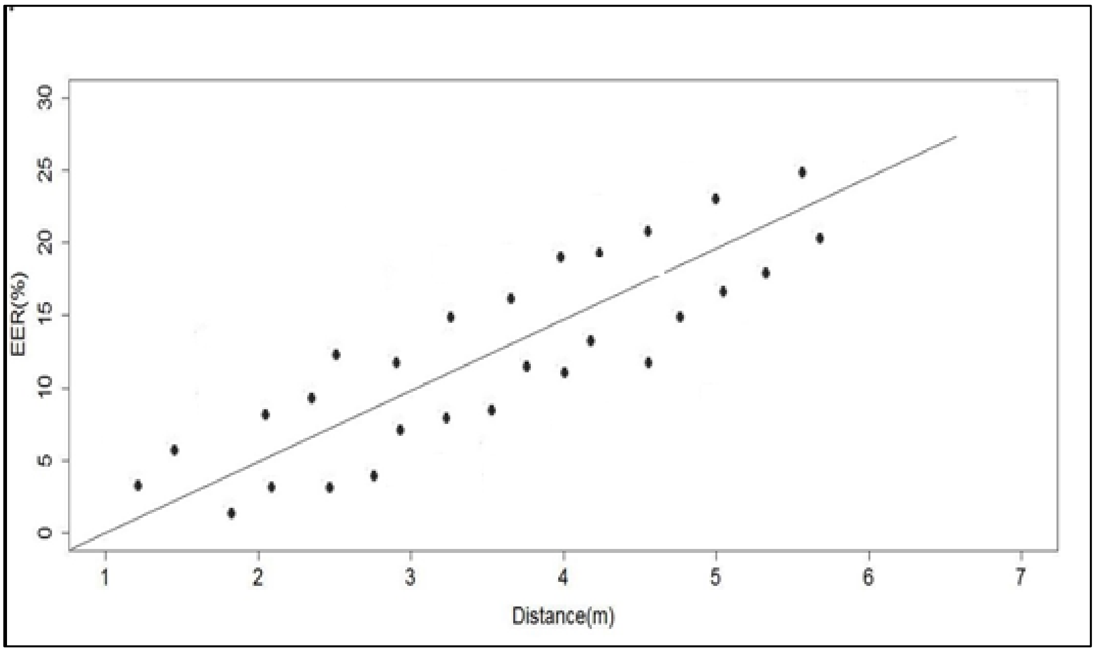


Figure 4.21 Fitted line of the EER (%) and source-receiver distance (m)

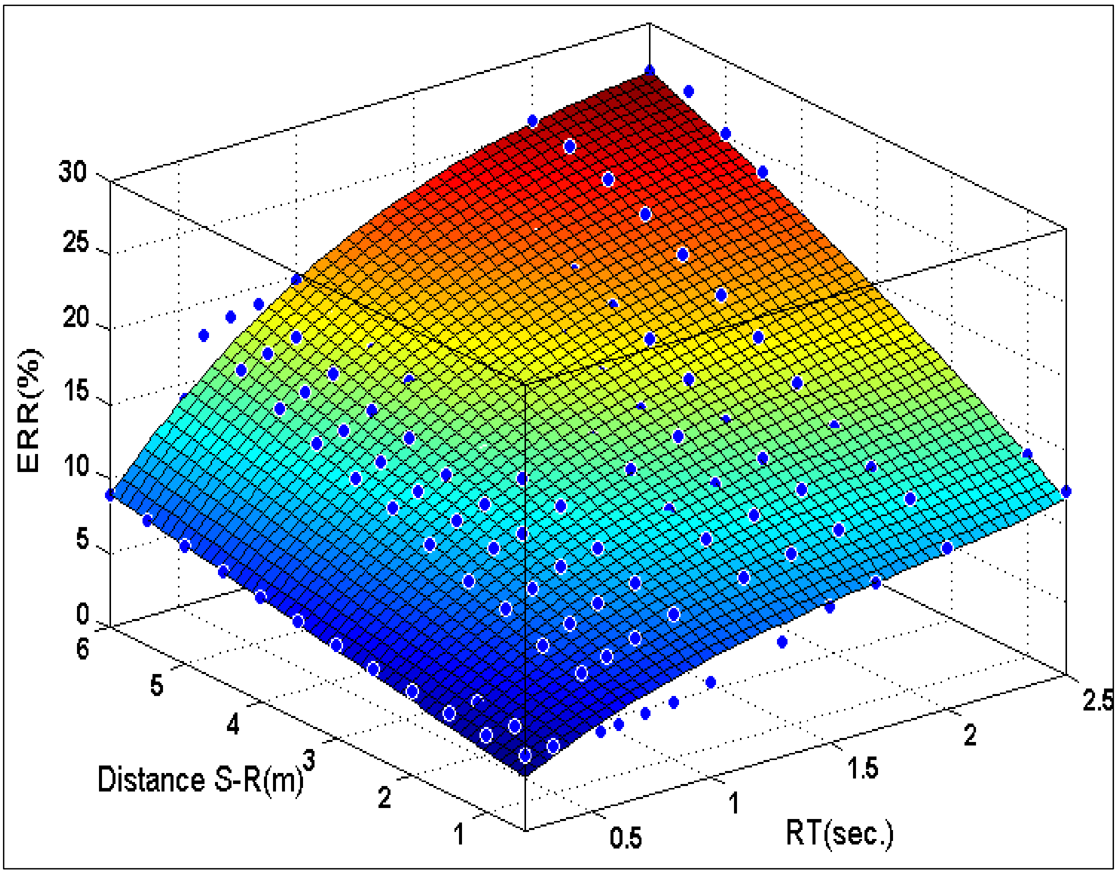


Figure 4.22 Final total error equal rates using

4.5 Chapter Summary

This chapter has given a description of the MSR Toolbox, which was developed by Microsoft Research a MATLAB toolbox to help with speaker recognition research. Moreover, in the first experiment, a speaker recognition system based on the state of the art of the Gaussian mixture Model–Universal Background Model (GMM-UBM) was evaluated, and the impact of real and simulated reverberant speech signals on the performance of the system was examined. In addition, the performance of two common features, the Mel-Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC) was evaluated and compared as a feature of the speaker verification system in reverberant conditions. The evaluation was based on SALU-AC corpora. The results were consistent, indicating that the GFCC feature provides the best overall performance. Finally, the effect of reverberation time and the source to receiver distance on system performance was evaluated. The results confirm that both parameters can affect system accuracy.

CHAPTER FIVE: THE ROBUSTNESS OF SPEAKER RECOGNITION USING REVERBERANT SPEECH TRAINING

5.1 Introduction

Mismatched acoustic transmission channels are responsible for the degradation in reliability of automatic speaker recognition, as many authors have identified (Castellano et al., 1996; González-Rodríguez et al., 1996; Ning et al., 2011). The issue becomes significant in the presence of acoustic reverberation.

The work in this chapter continues the efforts of the author to understand the effect of using speech signals corrupted by changes in reverberation time in the training stage and attempts to improve system robustness by convolving room impulse responses in the training phase of typical Gaussian mixture model based speaker recognition systems.

Three scenarios have been considered in this work. The first scenario is the use of clean speech samples of the enrolment phase, the second includes reverberant samples in the enrolment phase, and the third scenario uses two and four condition training to mitigate the effect of reverberation on system performance. In all scenarios, the best results occur when the reverberation characteristics of training and test speech are as close as possible. Thus, the potentials and limitations of including reverberant samples in the training phase to improve system robustness are identified.

5.2 Inclusion of Reverberant Cases in Training

The received reverberant speech signals for identification or recognition go through a pre-processing stage to decrease the reverberation so that it matches the acoustic conditions in the enrolment phase. Therefore, the removal or reduction of channel effects, to some extent, mitigate the mismatching issue at the cost of added distortions to the speech signals themselves, and therefore its effectiveness is limited. In this work, different scenarios were investigated to improve the performance of the system, such as training/testing matching and training with low and medium reverberant speech from same or different rooms. Reverberant, clean and artificially reverberated speech signals, with reverberation times of up to 2s, were used in training. Various reverberant speech signals were employed in the validation phases to identify how the inclusion of reverberant cases can affect the performance of the system. Training with a reverberant speech similar to the test speech can lead to significant improvement in performance of the system compared with clean speech training. This setup is depicted in Figure 5.1. Moreover, Figure 5.2 and Figure 5.3 shows the training with the low and medium reverberation time scenario; the experiment involves choosing a speaker model of reverberant speech is used encompassing training data corrupted by varying RT values. In this experiment, we use RT values in the range of 0.23, 0.33, 0.53, 0.62, 0.84, 1, 1.2, 1.5, and 2 seconds to generate the reverberant speaker models and for enrolment.

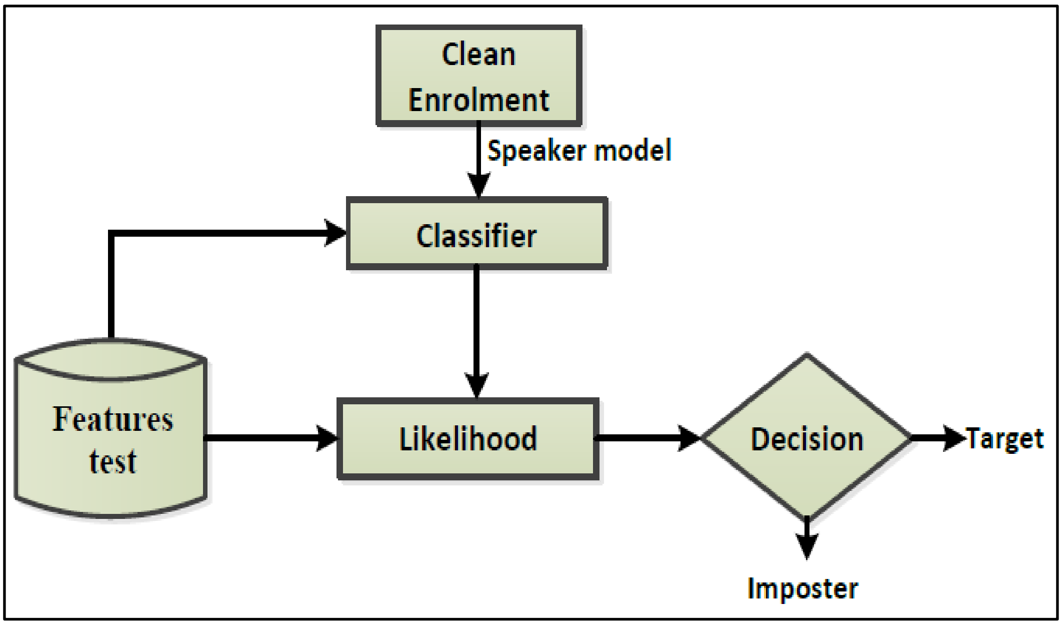


Figure 5.1 The baseline system using clean training and reverberant testing

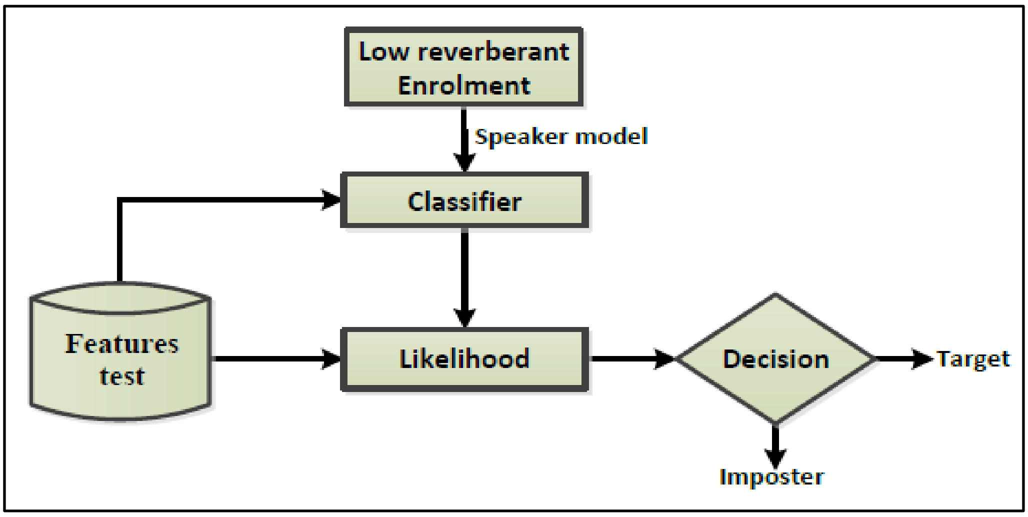


Figure 5.2 The training with low reverberation time

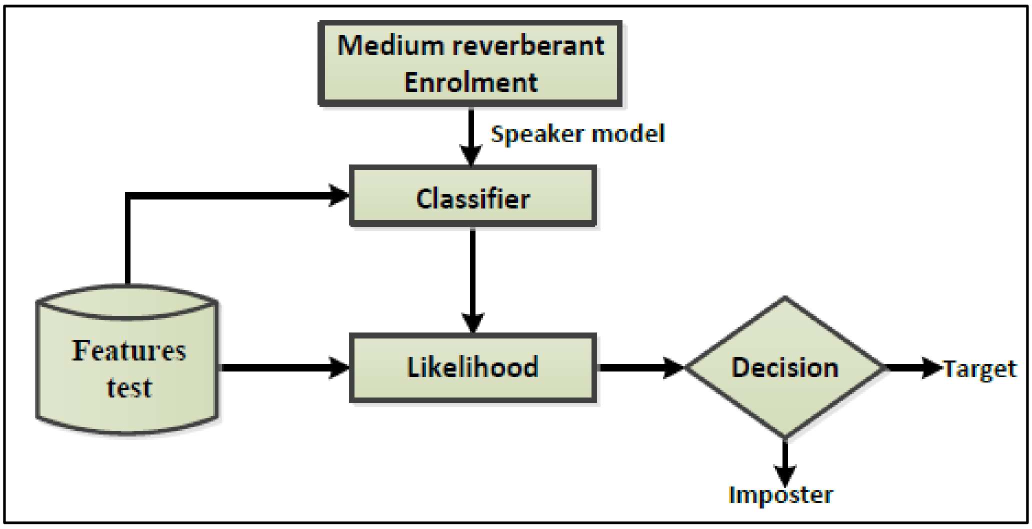


Figure 5.3 The training with medium reverberation time

5.2.1 Train/Test RT Matching

Acoustic matching of speaker models includes training and testing under the same room acoustic conditions, e.g. the same reverberation time. Figure 5.4 depicts this setup. In the training stage, several models are generated for each speaker under different reverberation conditions. First, the reverberant background model for each reverberation time is produced: Clean speech segments of various speakers are convolved with simulated room impulse responses, and a reverberant background model is trained utilising these segments. These reverberant background models have also been used for reverberation classification. Then, employing the GMM training, the reverberant speaker model is adapted from the reverberant background model and the speaker reverberant speech signal. During the verification stage, the models that more closely match the RT of the test data are used.

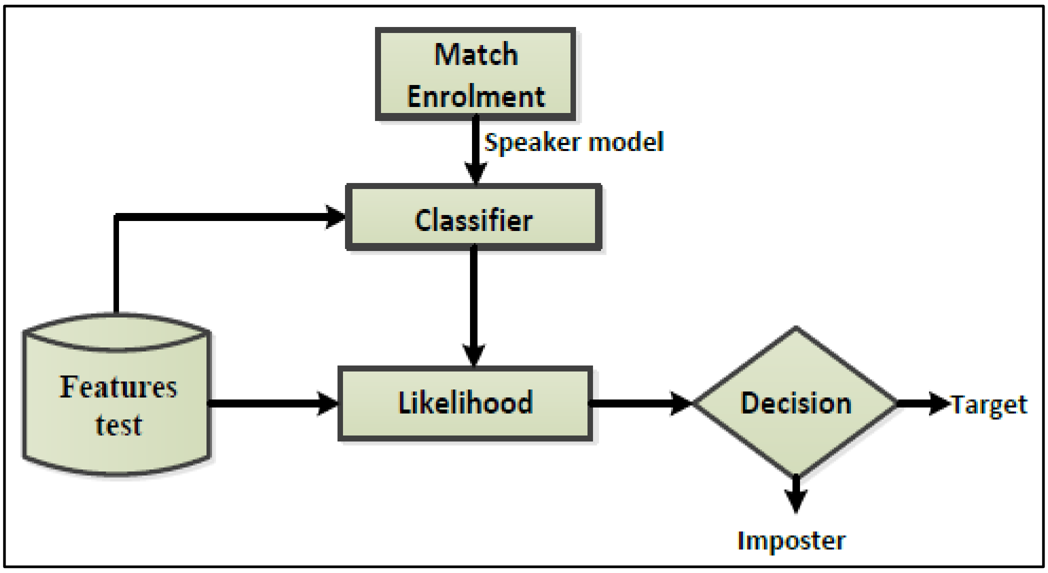


Figure 5.4 The training and testing matching

5.3 Multi-Condition Training

Adding reverberation to the speech samples for training is referred to as multi training when a speaker is registered or enroled in the system. The speech samples of the speaker in different reverberant conditions include the training set and those relating to that specific speaker. Multi-condition training is a promising solution for the two well-established mainstream speaker recognition systems, namely the Gaussian mixture model-universal background model (GMM-UBM) framework and i-vector based framework to tackle reverberant speech samples (Ming et al., 2007; Rajan, Kinnunen, & Hautamäki, 2013). In essence, the multi-condition training regime includes a speech sample together with a large number of possible reverberant conditions in the training phase of the speaker recognition system, so that, hypothetically, the trained system can generalise the reverberation cases in the training phase to any real-world situations in the retrieval phases. In these experiments, multiple utterances from different environments for each speaker were used to increase the match between enrolment and testing phase. Figure 5.5 illustrates the training stage using the multi-speaker utterances approach. GMM is trained for each speaker for both clean and reverberant utterances. In the testing stage for multiple utterances as in Figure 5.6, the extracted feature vectors are scored against all the S-N speaker models. Two types of multi- training conditions were used in these experiments; the first type uses two-condition training.

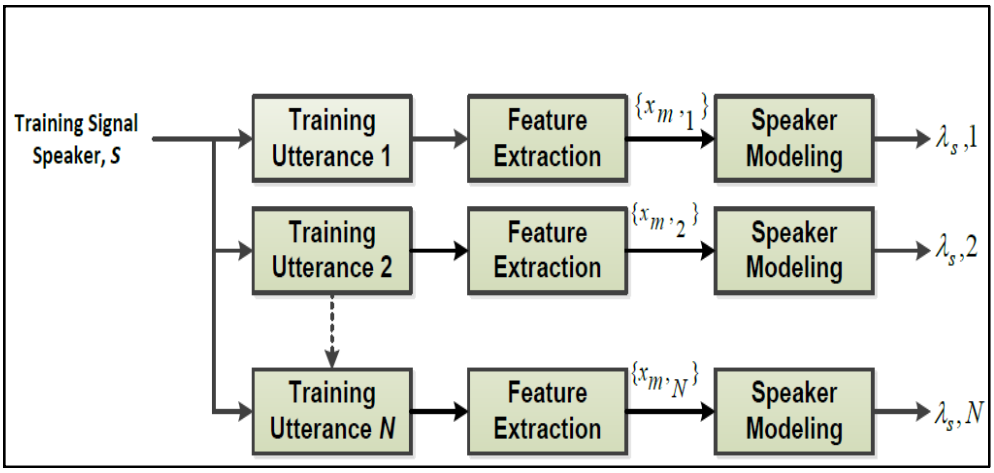


Figure 5.5 The enrolment phase where using signal from various training rooms

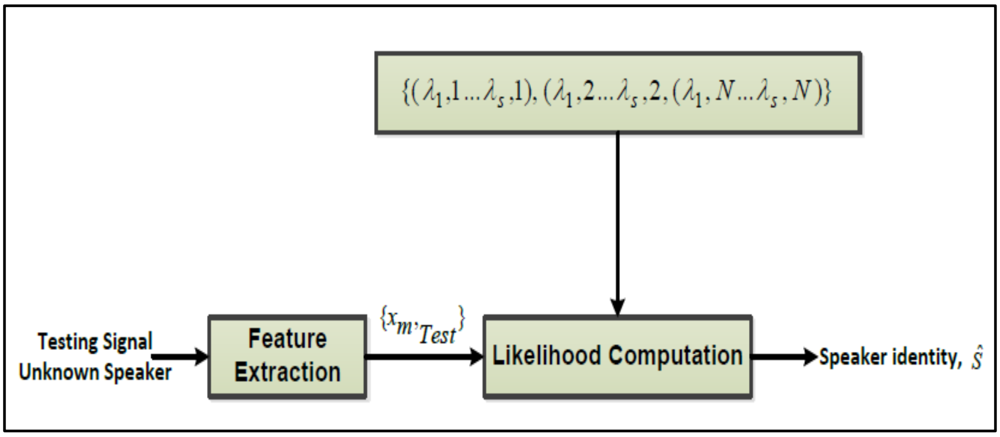

$\{(\lambda_1,1...\lambda_s,1),(\lambda_1,2...\lambda_s,2,(\lambda_1,N...\lambda_s,N)\}$
Testing Signal
Unknown Speaker
Feature Extraction
$\{x_{m,Test}\}$
Likelihood Computation
Speaker identity, $\hat{S}$

Figure 5.6 Tthe testing stage using multiple speaker models for each speaker

During training, two models were defined for each speaker by using clean speech as well as speech convolved with a room impulse response with a moderate reverberation time of 0.53s. While, the second type is a four-condition training, in this type the utterances are convolved with three impulse response (with RTs of 0.53s, 1.0s and 1.5s) creating four conditions including the clean cases. In each test phase, the new room impulse response was generated. As an example, speaker verification (SV) for the two and four-condition training utterances and reverberant test utterances is shown in Figure 5.7 and Figure 5.8.

In the verification stage, the model that yields the maximum log-likelihood is used, which is usually the model with the closest reverberation time. In the final step, the performance of the system where the acoustic conditions are assumed to be known is evaluated. This involves choosing a speaker model trained with utterances convolved with an impulse response with a similar reverberation time to the test case.

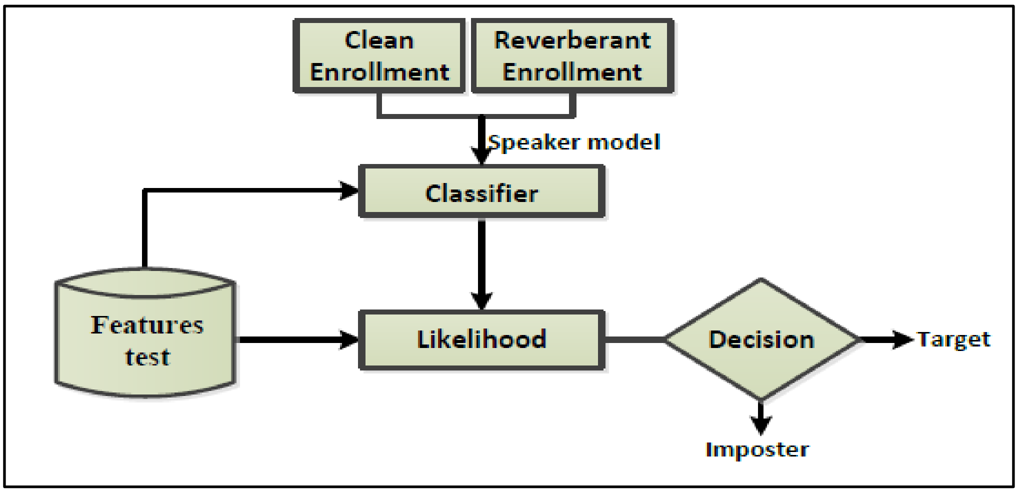


Figure 5.7 The two-condition training

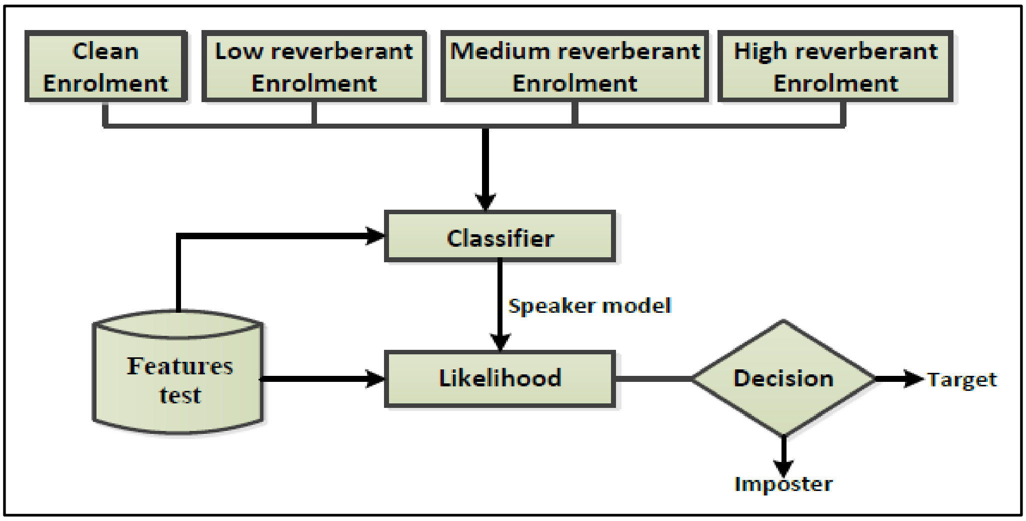


Figure 5.8 The four-condition training

5.4 Experiment Setup

In this experiment, the system was evaluated using speech samples obtained from the SALU-AC data set; the speech samples were obtained from 100 speakers (50 male and 50 female) from this data sets, truncated into 5-second excerpts for training and testing purposes. Each speaker provided 10 utterances; each utterance has a 5s duration of approximately 50s in total duration; 8 utterances (40s) were selected for the training phase with the remaining 2 (10s) used for the testing phase to produce exclusive training and test data sets. The speech data were sampled at 16 kHz. The speech samples, which are used in testing stage, are different from those used in the training stage. Commercial software CATT-Acoustic was employed to generate synthetic impulse responses for training and testing stages with reverberation times for 0.23, 0.33, 0.53, 0.62, 0.84, 1, 1.2, 1.5, and 2 seconds from several rooms; each test utterance was convolved with impulse responses that are obtained from simulation software. Noticed, that the different RIRs and different utterances from the SALU-AC data set were used for testing stage. For the system error evaluation, a set of standard performance metrics to score ASR has been generated by the National Institute of Standards and Technology (NIST) (Doddington, Przybocki, Martin, & Reynolds, 2000). For the system error evaluation, the National Institute of Standards and Technology (NIST) (Doddington et al., 2000) have generated a set of standard performance metrics to score ASR. For statistical testing, there are two kinds of errors; the false positive rate, and false negative rate, sometimes called false alarms. A false positive error occurs when the system falsely confirms an impostor as the target through the impostor verification stage. However, a false negative occurs when the system defines the target as an impostor through the verification target trials. The critical area of the curve where the error rates (False rejection rate (FRR) and false acceptance (FAR)) are equal is called the equal error rate (EER). In general, the lower the EER, the higher the system accuracy. The detection error trade-off (DET) curve is a very useful way of showing the accuracy of the system in a linear plot of bit error rates on a standard scale, referred to by the NIST (Chen & Lin, 2006).

5.5 Result and Discussion

Figure 5.9 depicts the box plot obtained depending on the percentage EER with the baseline, matching, medium RT training, two-condition training and four-condition training setups, respectively. The box plots are used to show overall patterns of response to a group. In addition, they provide a useful way to visualise the range and other characteristics of responses in a large group. This simplest possible box plot displays the full range of variation (from min to max), and it corresponds to the standard deviation according to the percentage EER for each scenario performance with different reverberation time. The diagram below shows a variety of different box plot shapes and positions. In Figure 5.9, the matching training and four-condition training appear to have approximately close centres, which exceed those of baseline and the two-condition training. The baseline seems to have larger variability than the other three scenarios. Depending on the lower max, the box plot indicated that the training\ testing matching scenario and four-condition training scenario produce the best performance compared with the other scenarios. Furthermore, for more investigation, the discussion of the results begins with the summary of the EER for each scenario (Table 5.1 Summary of the Equal Error Rates with different RT). “The evidence from this study suggests that the equal error rate of the training/testing matching scenario for all reverberation time levels is better than the other scenarios”. “The percentage EER obtained from all scenarios is depicted in Figure 5.10”. The x-axis represents the degree of reverberation time in seconds, while the y-axis represents the recognition accuracy of the system based on the equal error rate. The finding highlights that increasing the reverberation time value caused significant performance degradation in the case of the baseline system, consisting of clean training and reverberant testing. “For example, the percentage EER for a baseline is 0.42% with RT= 0.33s. This percentage increased to just 2.44% when the reverberation time increased to 0.53s. Moreover, the EER rose to 19.32% at RT= 2s. However, in the second scenario train/test matching, an inverse relationship was found. The system accuracy shows more reverberation robustness when reverberant speech samples are used in the enrolment stage compared with using clean speech samples for enrolling; especially with (RT< 1.2s) the EER remained below 0.98% and increased to 3.22% at RT= 2s. These amounts are considered a significant relative reduction in EER percentage. However, the performance of the other types of training condition was less than training/testing matching. Regarding two training conditions, there is a clear degradation in system performance when the reverberation time increased, especially over 0.62s. Despite the poor performance relative to other types of training, the RT-matched setup clearly improved performance, and a relative reduction could be seen in the percentage EER at different levels of reverberation time”. “Furthermore, the detection error trade-off curves plotted in Figures 5.11and 5.12 clearly show the false negative (rejection) rate and false positive (acceptance) rate for the training/testing matching are better with different reverberation time values”. It can be seen that the accuracy of false positive rate (FPR) for the training/testing matching and four-condition training (the red and blue line) shows significant improvement compared to the baseline result. More DET curves with different reverberation times are shown in Appendix B. “The conclusion that can be drawn from the present study is that using reverberant training can improve the performance of the system and can to some extent mitigate the performance degradation. Therefore, if acoustic conditions can be somehow estimated and suitably included in the pre-training of the models, the robustness of the system can be improved”.

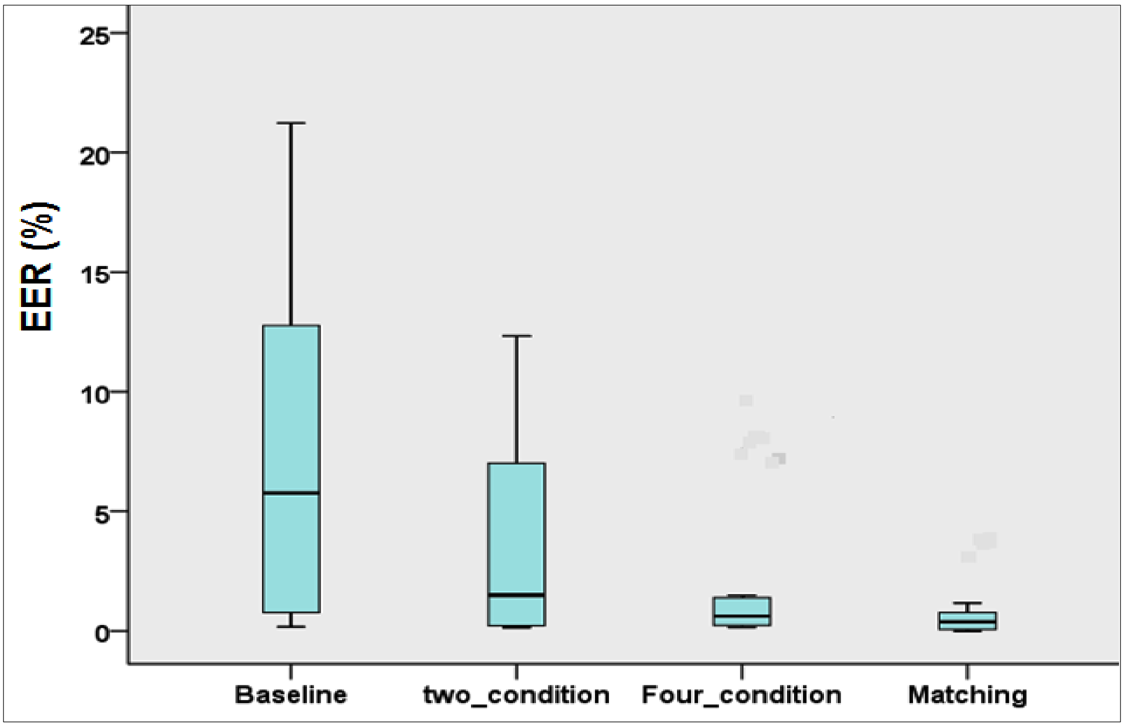


Figure 5.9 Box plots of Error Equal Rate are using different reverberation time for each scenario

Table 5.1 Summary of the Equal Error Rates with different RT tests

RT(s)	EER %			
	Baseline	Matching	Two condition	Four condition
Clean	0	0	0.16	0.15
0.23	0.24	0.22	0.23	0
0.33	0.42	0.51	0.52	0.07
0.53	2.44	0.14	0.13	0.12
0.62	4.55	0.74	0.71	0.31
0.84	6.98	2.26	0.81	0.44
1	9.88	4.34	0.52	0.49
1.2	12.93	7	1.48	0.76
1.5	15.44	9.44	1.39	1.17
2	19.32	12.33	7.44	3.11

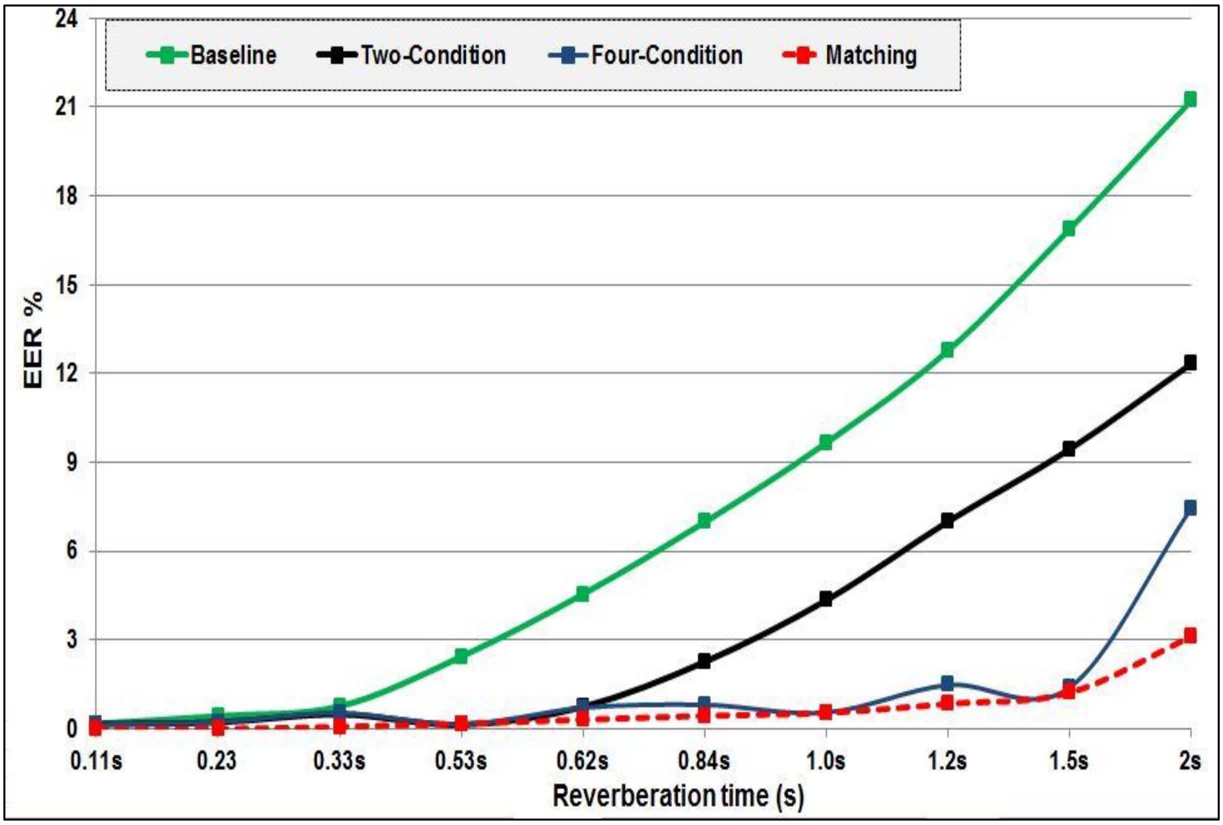


Figure 5.10 System performance using multi-training conditions using SALU-AC data set

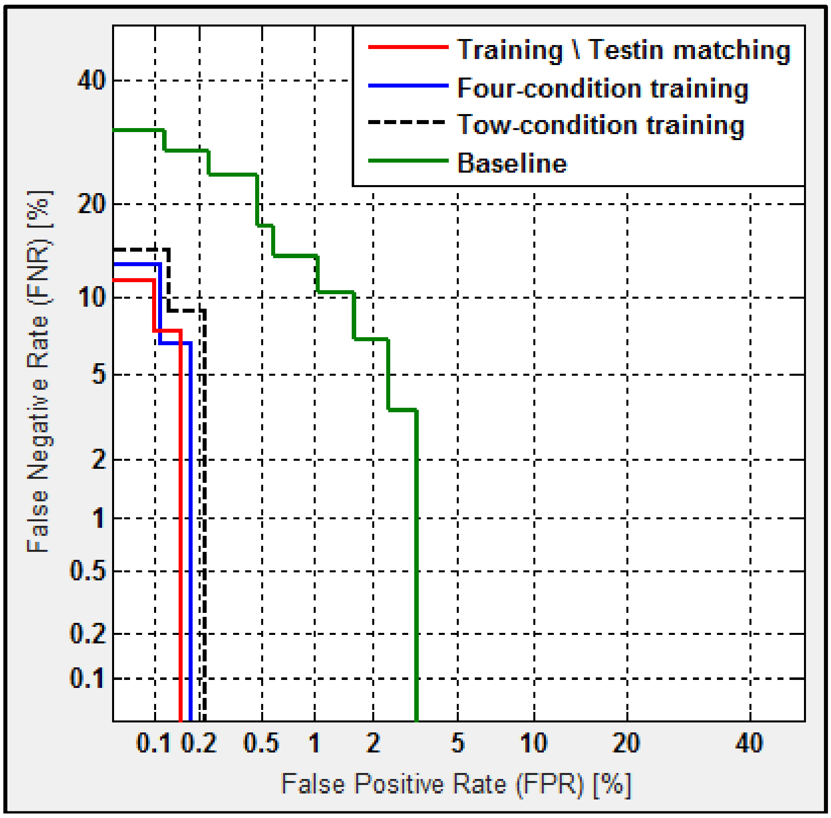


Figure 5.11 DET Curve for multi training scenarios with RT= 0.53s in testing phase

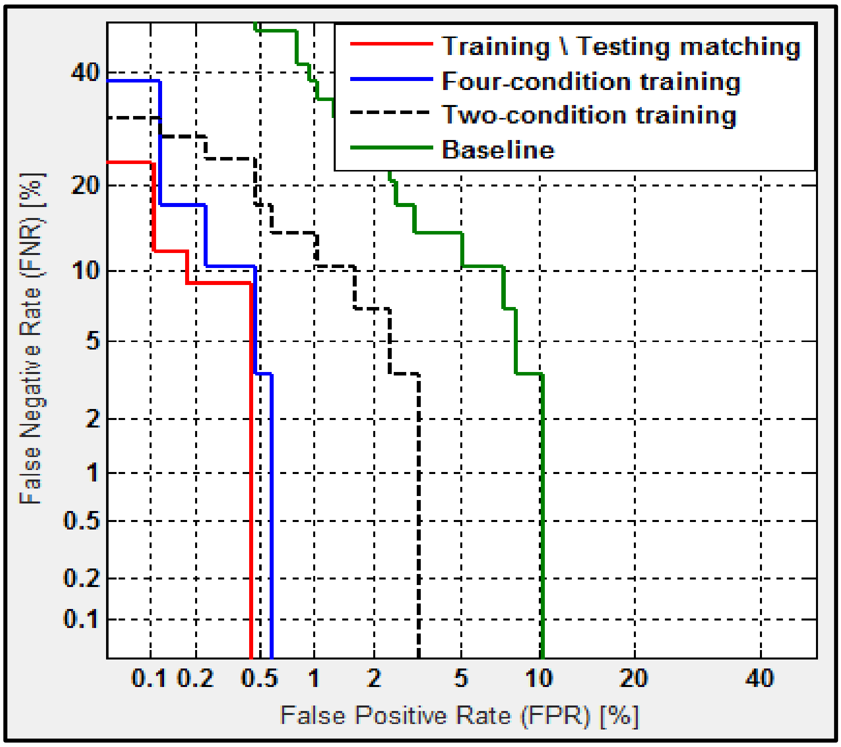


Figure 5.12 DET Curve for multi training scenarios with RT= 1s in testing phase

5.6 Chapter Summary

The work in this chapter improves the robustness of speaker recognition using speech signals corrupted by the reverberation time in the training stage. Three scenarios have been considered in this work: The first scenario used clean speech samples in the enrolment phase and the second included reverberant samples in the enrolment phase. Thus, the potentials and limitations of including reverberant samples in the training phase to improve system robustness are identified. The third scenario is using two and four condition training to mitigate the effects of reverberation on the system performance. The best results occur when the reverberation characteristics of training and test speech are as close as possible, but an exact match is not necessarily needed. “These findings suggest that in general, the inclusion of environmental conditions in the training stage can to some extent mitigate the performance degradation; therefore, if acoustic conditions can be somehow estimated and suitably included in the pre-training of the models, the robustness of the system can be improved”.

THE TRAINING ON THE FLY AND THE APPLIED ESTIMATION METHODS

6.1 Estimation Methods

Reverberation is a well-known phenomenon in the area of room acoustics and speech-processing representing the gradual sound energy decay in an enclosure after the sound source has been switched off. Due to the detrimental effects of temporal and frequency smearing of the speech signals received by microphones that it causes, numerous methods for signal enhancement have been proposed, where a number of them require the reverberation time (RT) parameter value to be known in advance.

This chapter introduces a review of the estimation methods in general as well as to the maximum likelihood estimation (MLE) method.

6.2 Preliminaries to the Addressing Mismatch Problem

Mismatched acoustic transmission channels are known to cause degradation in the reliability of speaker recognition. The problem becomes significant in the presence of acoustic reverberation (Bimbot et al., 2004; Jin et al., 2007; Mammone et al., 1996). Several attempts to overcome the reverberation problems have been reported (Borgstrom & McCree, 2012; Falk & Chan, 2010; McCowan, Pelecanos, & Sridharan, 2001; Ming et al., 2007; Nakatani & Miyoshi, 2003; Ning et al., 2011; Sadjadi & Hansen, 2012; L. Wang & Nakagawa, 2009; Zhao, Shao, & Wang, 2012). Some of these methods adopt multi-microphones, microphone arrays, or even multi-modal schemes to address the reverberation problems; but these methods are often redistricted in real-world applications due to the specific requirements in speech acquisition procedures. Other methods tried to perform channel equalisation, i.e. the removal or reduction of channel effects, or the dereverberation methods. These methods to some extent mitigate the mismatching problem at the cost of added distortions to the vulnerable speech signals themselves, and therefore, their effectiveness is limited. However, successful single channel dereverberation for speaker recognition does not seem to exist. Currently, available single channel blind dereverberation techniques might cosmetically improve perceived cleanness of speech to some extent, but they do not appear to improve the performance of speaker recognition, due to the distortions imposed on the so-processed speech signals.

For speaker recognition, it is evident that the best performance is achieved when the reverberation features of training and test speech phases are identical or reasonably close (Al-Noori, Al-Karawi, & Li, 2015; Ming et al., 2007; Rajan et al., 2013). Therefore, if acoustic conditions can somehow be estimated and suitably pre-trained models called upon, the robustness of the system can be improved. As an alternative to pre-training, multiple models under different acoustic conditions for every single speaker, “training on the fly” as conceptualised previously (Francis F Li, 2016) can be used.

6.3 Estimation Methods

Room reverberation time (RT) is a crucial factor that qualifies the room acoustics (Heinrich Kuttruff, 2009). A priori knowledge of reverberation time can potentially enable advanced training strategies, improving the robustness of speaker recognition in reverberant conditions. Different techniques have been employed in the literature to estimate or measure reverberation time. Early in the 20th century, Sabine (Sabine, 1922) formulated the RT, based entirely on the geometry of the environment (i.e., Volume and surface area) and the absorption attribute of its surfaces. However, such a method requires that the room geometry and absorption characteristics of the surfaces of the room be determined first. Consequently, methods that are based on sound decay curves were developed: A broadband noise, e.g. White noise is radiated in a room, then at the instant when the sound field attains a steady state, the noise source is switched off, and the decay curve is obtained. The time that it takes to the sound pressure level to reduce by 60 dB is term RT_{60}, the early definition of reverberation time. Schroeder developed an integrated impulse response method (Schroeder, 1965), in which the decay above curves can be calculated from the measured impulse response from a source to a receiver.

Existing theoretical estimation or measurement methods do not solve the problem of speaker recognition in reverberant conditions since in such an application speech signals were often acquired from unknown or various spaces or rooms. Several approaches have been developed that can estimate RT directly from the reverberant signals (H. Löllmann, Yilmaz, Jeub, & Vary, 2010; H. W. Löllmann & Vary, 2008; Ratnam et al., 2003; Wen, Habets, & Naylor, 2008).

Essentially, these algorithms establish a parametric statistical model for the sound decay, followed by maximum likelihood estimation (MLE) to estimate the decay rate presented in the signals. The resulting decay curves are used to calculate the reverberation time.

6.3.1 Maximum Likelihood Estimation and Speech Decay Model

An MLE of a set of parameters that best model the decay phases of the reverberated signals are undertaken. The MLE was firstly suggested in the 1910s by Fisher (Aldrich, 1997) and is one of the most important models utilised for parametric estimation in statistics. More specifically, the reverberation time of a room can be estimated from the received speech signals (Kendrick, Li, Cox, Zhang, & Chambers, 2007; H. Löllmann et al., 2010) and this was inspired by the method used by Ratnam et al. (2003). A reverberant speech signal is considered which is given by a speech signal $s(k)$ convolved with the room impulse response $h(\eta, k)$ of length Lh (H. Löllmann et al., 2010):

$$Z(k) = \sum_{n=0}^{L_h - 1} s(k-n).h(n,k) \qquad 6.1$$

A discrete random process model the sound decay *d*(*k*)

$$d_m(k) = A_r v(k) e^{-pkT_S} \qquad 6.2$$

where A_r > 0 represents the real amplitude, decay rate p and $\in(k)$ marking the unit step sequence. The variable T_S = 1/*fs* denotes the sampling period and $v(k)$ is a sequence of independent and identically distributed (i.i.d). Random variables with zero mean the variance of one and normal distribution *N* (0, 1). From the relationship between the decay rate *ρ* and the reverberation time, the following equation can establish T_{60}

$$T_{60} = \frac{3}{p \log 10(e)} \approx \frac{6.908}{p} \qquad 6.3$$

Due to this relationship, the term decay rate and RT will be used interchangeably below. According to our model, *d*(*k*) is a random variable with the Gaussian probability density function (PDF)

$$p_{d(k)}(x) = \frac{1}{\sqrt{2\pi\xi(k)}} \exp\{-\frac{x^2}{2\xi^2(k)}\} \qquad 6.4$$

where

$$\xi(k) = A_r a^k \in (k) \quad \text{and} \quad a = e^{-T_S p} \qquad 6.5$$

The sequence *d*(*k*) for *k* ∈ {0, . . . ,*N* − *1*} is modelled by *N* independent random variables with zero mean and non-identical PDFs having normal distributions. These allow the derivation of a maximum likelihood (ML) estimator for the unknown decay rate or RT, respectively (H. W. Löllmann & Vary, 2008; Ratnam et al., 2003). The decay rate *ρ* is estimated from a given sound decay *d*(*k*) by finding the maximum of the log-likelihood function

$$\hat{p}^{(ML)} = \max_{p}\{L(p)\} \qquad 6.6$$

$$L(p) = -\frac{N}{2}((N-1)\ln(a) + \ln(\frac{2\pi}{N}\sum_{i=0}^{N-1} a^{-2i} d^{2}(i)) + 1) \qquad 6.7$$

The ML estimate for the RT $\hat{T}_{60}^{(ML)}$ is obtained by Equation 6.3.

6.3.2 Optimisation Method

Typically, a room response model contains parameters, which change the behaviour of the model. The optimisation is used to determine the 'best' model parameters that reproduce available experimental room response information. A technique for estimating reverberation time from received speech signals utilising a model of sound decay is defined; this is called the maximum likelihood method. The input to the likelihood function contains the received reverberant signal and all of the model parameters.

As the parameters are unknown, the likelihood function must be 'optimised' so that a possible set of model parameters (e.g. Decay time) responsible for generating the decay is found. In the case of a likelihood function, this parameter set yields the maximum function value of all possible parameters. The optimisation is the procedure of maximising of a wanted quantity or the minimising of an unwanted one. No single optimisation technique is available for solving all optimisation problems in a uniquely efficient manner. Numerous optimisation approaches have been developed to date for solving various kinds of optimisation problems. Locating the optimum of a complicated, multi-dimensional function is a challenging problem. There is no final optimisation procedure, and each technique has its advantages and disadvantages. In this work, the multivariable optimisation methods were used to optimise the obtained results.

6.4 The Training on the fly and the applied Estimation Methods

This work proposed to train system on the fly during its recognition operation. In this section, a maximum likelihood estimation algorithm is proposed for blind-estimation of reverberation time from speech signals submitted for verification. A cluster of impulse responses with the estimated reverberation time is selected. The estimates are used to choose matched acoustic impulse responses or transfer functions for inclusion in the training of the pattern recognition model on the fly, this work was published by the IEEE conference based on taking the notable advantage confirmed by the success in the work that of improving the robustness of speaker recognition via training which is published in IEEE conference (Al-Noori et al., 2015) and other work in (Ming et al., 2007; Rajan et al., 2013). Therefore, a combined method is proposed in this study to mitigate the effect of reverberation. “Instead of including a large number of possible channel conditions, the channel features are predicted, and only appropriate channel models are used to training the system. Hypothetically, the method proposed in this work can be a universal method applicable to any mainstream speaker recognition framework algorithm”. “Experimental results have shown significant improvement in system performance regarding reduced equal error rate and detection error trade-off”.

6.4.1 The Proposed System

The enrolment phase includes estimating a model that represents (summaries) the acoustic (and often phonetic) space of each speaker. During the evaluation phase, either each test segment is scored against all enrolled speaker models to determine who is speaking (speaker identification) or against the background model and a given speaker model to accept/reject an identity claim (speaker verification). The proposed system is illustrated in Figure 6.1 comprises three stages. The first stage included estimation the reverberation time using maximum likelihood method in 1/3 octave band as well as using training on the fly schema. The training on the fly approach is used to select the best match between the training data sets and the estimated reverberation time depending on the simple Euclidean distance schema. The Euclidean distance or Euclidean metric is the "ordinary" straight-line distance between two points in Euclidean space. While, the second stage is created a reference model that represents (summaries) the acoustic (and often phonetic) space of each speaker, during the evaluation phase, each test segment is scored against all enrolled speaker models to determine who is speaking (speaker identification). Finally, in the last stage, each test segment is scored against the background model and a given speaker model to accept/reject an identity claim (speaker verification).

6.4.2 Training on the Fly Schema

This experiment proposes to train the system on the fly during its recognition operation. The proposed system is illustrated in Figure 6.1. The dotted line indicates the training on the fly part, while the blue parts indicated the training and testing parts. On receiving a submitted speech signal for recognition, a maximum likelihood algorithm is used first to estimate the reverberation time RT in octave or 1/3 octave bands in the speech frequency range; these are used to synthesise a model for the virtual channel, or to choose a closet matched one from a channel model bank. Therefore, each impulse response was filtered into in seven-octave bands: 125, 250, 500, 1K, 2K, 4K and 8 kHz to further analyses how the response alters with frequency yielding decay curve estimates for each octave band. Speech stimuli have limited bandwidth and therefore can only efficiently determine objective parameters in the frequency range where speech signals have sufficient energy (Francis Feng Li, 2002). “The speech signals do not have significant energy above 6300 Hz while, at the 8000 Hz band, the speech signals have very little energy. To solve this problem 6300 Hz band is used instead”.

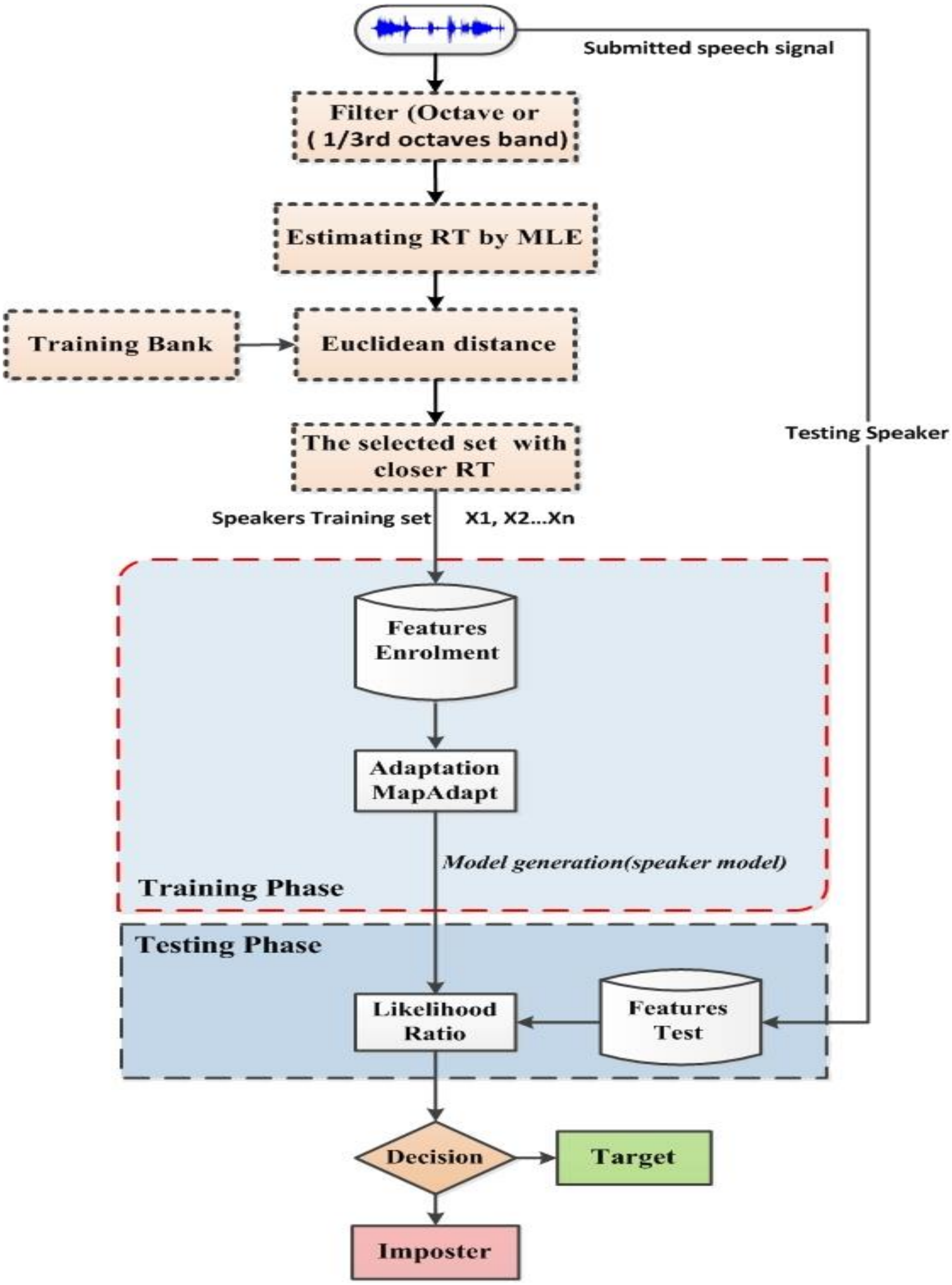


Figure 6.1 The proposed system using estimate RT and ELR

6.4.3 Training Data Bank

Clean speech samples were obtained from SALU-AC dataset, 100 speakers (50 male and 50 female were used). Each speaker provided speech samples excerpt of 3 minutes in total duration. Two minutes were used in training stage and the remaining 1 minutes used in a testing stage. The speech samples divided into 5s duration for each utterance. On average there is about 10 training utterance per speaker were used in the training stage; each utterance has a 5s duration of approximately 50s in total duration; were selected for the training phase with the remaining 2 utterances (10s) used for the testing phase to produce exclusive training and test data sets. Furthermore, two types of impulse response were used to simulate different reverberant conditions. The first set of impulse responses were generated from 30 rooms of various dimensions, room volumes ranged from 270 m^3 to 5472 m^3, simulated using the well-established and tested room acoustics simulator CATT–Acoustic (CATT-Acoustic, 2010). To create the rooms, each dimension is randomly chosen using a random number constrained within appropriate (realistic lengths) bounds. More details on the generating impulse response in 3.5.2.2. The sound source and receiver are moved around, and a number of impulse response captured. For each room, we simulate seven room impulse responses with the source-to-receiver distance from (1 m to 7m). This gave a useful database of impulse responses (210-impulse response) and represented a range of the different position of the source to receiver in each room with broadband reverberation times from 0.2 to 3.0 seconds. Each impulse is labelled by its reverberation times in seven-octave bands: 125, 250, 500, 1K, 2K, 4K and 8 kHz. The reverberation times involved in the submitted speech samples in the same octave bands are estimated. Speech samples acquired in the enrolment phase were convolved with these impulse responses. A set of speaker models in each of the conditions were generated. Each set of speaker models characterises a unique reverberant condition and is used independently for speaker recognition. For cross-validation to the proposed method, the second impulses responses were obtained from the Aachen Impulse Response (AIR) database (and down-sampled to 16 kHz) which, offers a wide range of reverberant conditions (168 RIRs with T_{60} ranging from 0.1 s to about 4 s) (Jeub et al., 2009). From this dataset, four different rooms comprising studio booth, meeting room, office, and lecture room were used, each room offered a 10-impulse response with the different source to receiver distance were used to generate training bank and, 5-impulse response used for the testing phase. Furthermore, the speaker models from the best matching conditions should be used. Reverberation classification using simple Euclidean distance criteria has been proposed to classify the estimated reverberant condition as one of the training bank and select speaker models from the chosen condition for recognition (Akula et al., 2009; Peer et al., 2008). Thus, a matched virtual channel is created. The closest match from the training bank data set of reverberation times ranging from 0.11s to 3.0s was selected. Table 6.1 demonstrated different reverberation time in 1/3 octave band. Moreover, Figure 6.2 shows the 1/3 octave band for reverberation time 0.23s.

Table 6.1 Reverberation time in 1/3 octave band

Averaged RT(s)	Octave bands(Hz)						
	125 Hz	250Hz	500Hz	1kHz	2kHz	4kHz	8kHz
0.33	0.54	0.41	0.34	0.39	0.2	0.27	0.21
0.53	0.78	0.57	0.68	0.51	0.47	0.39	0.35
0.71	0.98	0.89	0.83	0.79	0.69	0.44	0.35
0.84	1.22	0.99	0.88	0.82	0.77	0.68	0.57
1.0	1.65	1.33	1.38	0.82	0.77	0.68	0.57
1.5	1.88	1.77	1.73	1.69	1.44	1.22	0.98
1.8	2.44	2.23	1.89	1.75	1.64	1.44	1.32
2	2.66	2.54	2.11	1.85	1.79	1.64	1.44
2.5	3.22	2.84	2.72	2.61	2.51	1.92	1.69
3.0	4.12	3.62	3.27	3.27	2.55	2.45	1.72

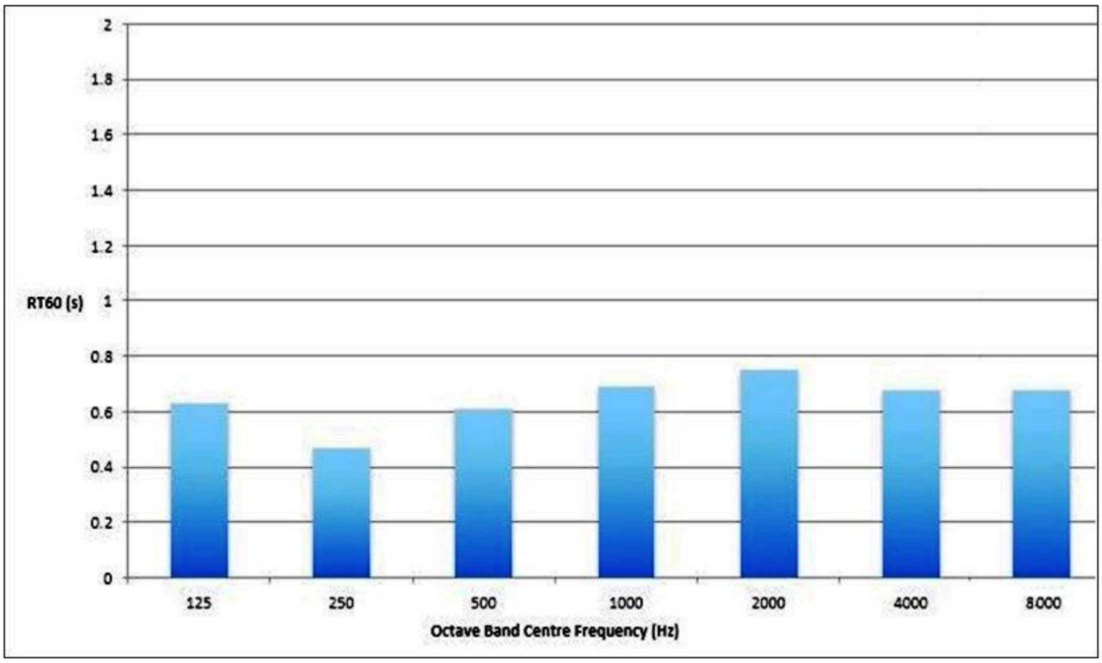


Figure 6.2 The RT60 calculator produces an average of 0.6s across the frequencies with a slight dip in the 250Hz range.

6.4.4 Experiment Setup

The current study involves an investigation of the effects of using the estimated reverberation time in the training stage to mitigate the impact of reverberation on the speaker verification system. Experiments were carried out to validate the proposed method. Clean speech samples were obtained from SALU-AC dataset, 100 speakers (50 male and 50 female were used). Each speaker provided speech samples excerpt of 3 minutes in total duration. Two minutes were selected for the training phase and the remaining 1 minutes used in a testing stage. The speech samples divided into 5s duration for each utterance to produce exclusive training and test data sets. Each speaker model has been tested 10 times, in which one represents the true speaker and the remaining nine are impostors. Simple energy-based voice activity detection was applied to remove the large chunks of silence in the excerpt. The speech samples, which are used in testing stage, are different from those used in training stage. Speech samples acquired in the enrolment phase were convolved with these impulse responses. The MLE method used to estimate the reverberation time that involved in the submitted speech samples in the same octave bands. These emulated reverberant speech samples were used to train the speaker recognition system. In the training stage, GMM training was performed to adapt the target models. Furthermore, feature space GFCC was experimented with. The GFCC has been found essentially more robust as the features for the GMM based speaker recognition in reverberation conditions. In the testing stage, after feature extraction, a log-likelihood ratio test was employed to compute target and impostor scores.

A simple percentage error rate is not adequate to indicate the performance of the system since false acceptance, and false rejection has different impacts on the system. In speaker recognition and other biometric security systems, the EER (equal error rate) is often used as a combined single measure for error. The introduction of the EER measure gives a more suitable tool for the evaluation of the performance of detection systems in general and speaker verification systems in particular. “The EER is the value where the false negative rate (FNR) and false positive rate (FPR) are equal. The lower the EER, the higher the reliability of a biometric system. The DET curve compares the false positive rate against the false negative rate by varying the decision threshold. A very high threshold will result in a very safe system with a very high rejection rate. On the other hand, a low threshold means a very high acceptance rate but also a great impostor acceptance rate. Therefore, the performance of the proposed method was evaluated using equal error rate (EER) values and the detection trade-off (DET) curve”.

6.4.5 Experiment and Results Discussion

The system performance has been evaluated using two types of impulse responses. The first evaluation was conducted using impulse response generated by CATT-Acoustic software from several rooms. However for cross-validation to the proposed method, the second impulses responses were obtained from the Aachen Impulse Response (AIR) database (and down-sampled to 16 kHz) which, offers a wide range of reverberant conditions (168 RIRs with T60 ranging from 0.1 s to about 4 s) (Jeub et al., 2009). From this dataset, four different rooms comprising studio booth, meeting room, office, and lecture room used. In the following sections more explanation on these evaluations.

6.4.6 Evaluation of the Method with Simulated Impulse Responses

To evaluate a universal method for reference matching is not an easy task. Feature selection and machine learning algorithms and training scenarios can all have effects on system performance. The lack of a standardised benchmark regime makes comparisons to other work difficult. “In this experiment, the objective was to identify whether the use of estimated reverberation time can usefully determine a matched virtual channel for re-training and improve the system reliability”. It would be ideal to avoid the use of similar channel models for the training and testing. “In this work, the proposed method was validated utilising text-independent speaker recognition testbed based on the Microsoft Research (MSR) identity toolbox (Sadjadi et al., 2013)”.

The objective parameters were estimated from the reverberated signal, and results were compared to the values that were calculated directly from the impulse response. Estimation error using different lengths such as (1s, 2s, 3s, 4s,5s, 6s, 7s, 8s, 9s,10s) of speech signals, then calculated, which refers to the difference between the estimated RT_{60}, was calculated using MLE and the reverberation time obtained by Schroeder’s backwards integration from impulse responses to determining the proper speech sample length that is used with MLE. The results show significant accuracy in long signals (8s-10s). This seems to suggest that giving longer signals can improve estimation accuracy, which is not surprising. In all cases, the largest error was 0.4 seconds (Al-Karawi, Al-Noori, Li, &

Ritchings, 2015). In this case, the microphone was far away from the source. Under other distance conditions, MLE gives an accurate estimation. Figure 6.3 shows the statistical results of the relation between sample length and estimation error. The MLE gives higher accuracy when the reverberant speech samples are longer. Furthermore, Table 6.2 shows the estimated reverberation time using MLE method and the measured reverberation from impulse response by Schroeder's backwards integration. It can be observed that the proposed MLE method provides a closer estimation value to the RT measured by Schroeder's approach. Figure 6.4 shows the measured reverberation from impulse response by Schroeder's backwards integration. The baseline system was created using MSR toolbox and GFCC features.

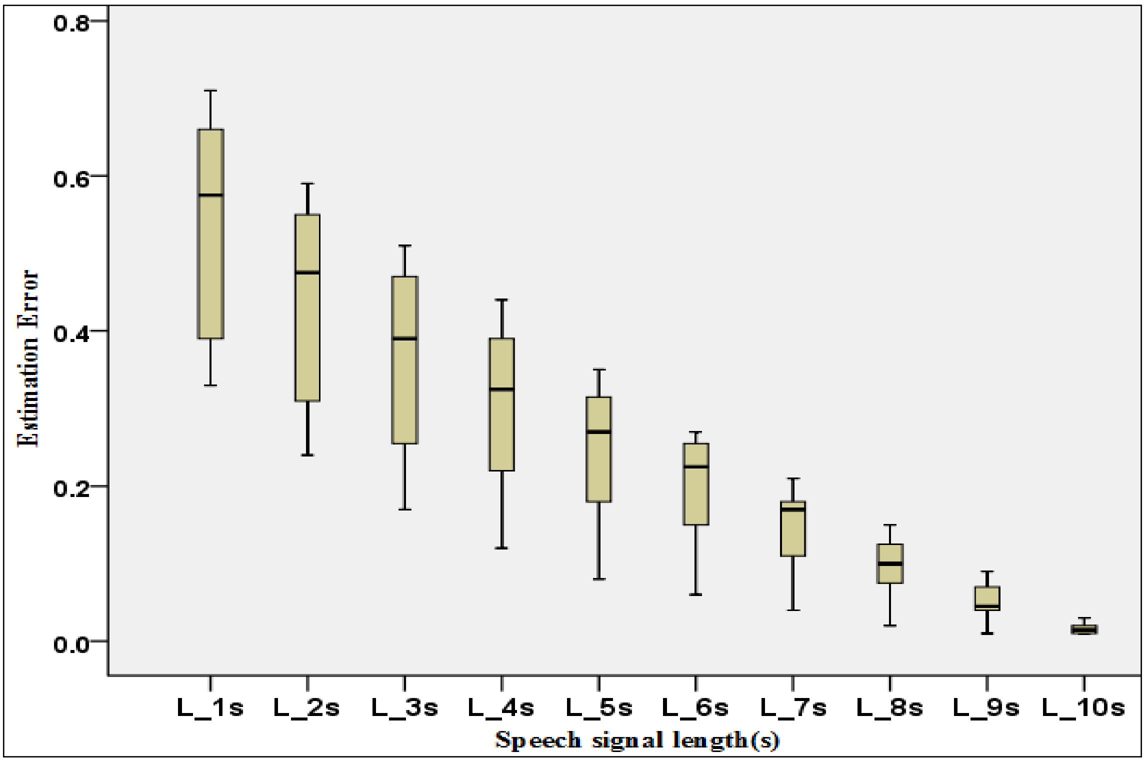


Figure 6.3 The distribution of estimation errors versus the used length of speech samples

Table 6.2 The estimation result, according to different methods

Room types	Distance (Source-Receiver.)	Schroeder backwards integration	Estimated RT by MLE
Studio booth	0.50	0.11	0.15
	1.0	0.18	0.23
	1.50	0.21	0.25
	2.0	0.24	0.23
	2.5	0.27	0.25
Meeting Room	1.70	0.23	0.27
	1.90	0.25	0.29
	2.25	0.27	0.30
	2.80	0.28	0.27
	3.20	0.29	0.28
Office Room	1.00	0.33	0.37
	2.00	0.42	0.48
	3.00	0.51	0.56
	4.00	0.38	0.36
	5.00	0.36	0.38
Lecture Room	4.00	0.71	0.75
	5.56	0.82	0.88
	7.10	1.00	1.2
	8.68	1.1	1.5
	10.2	1.2	1. 7

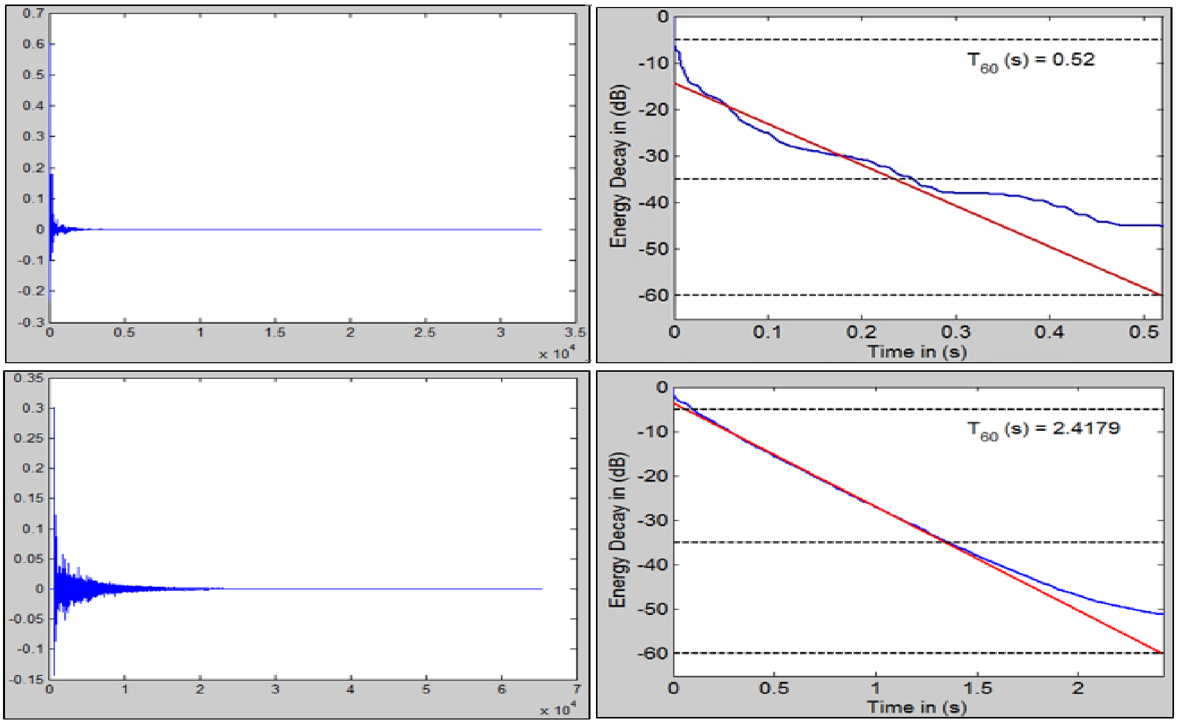


Figure 6.4 Calculation of reverberation time using Schroeder's backwards integration

Furthermore, Figure 6.5 depicts the box plot obtained with the baseline result and training on the fly schema. Each speaker model has been tested 10 times, in which one represents the true speaker and the remaining nine are impostors. This simplest possible box plot displays the full range of variation (from min to max), and it corresponds to the standard deviation according to the percentage EER for each speaker model during 10 times testing using both methods with different reverberation time. The benchmark appears to have a larger variety than the proposed method. Moreover, it is indicated that the proposed method produces the best performance compared with the baseline as it has a lower maximum, approximately 2.2, and median around 1.2. This box plot corresponds to the standard deviation of EER %.

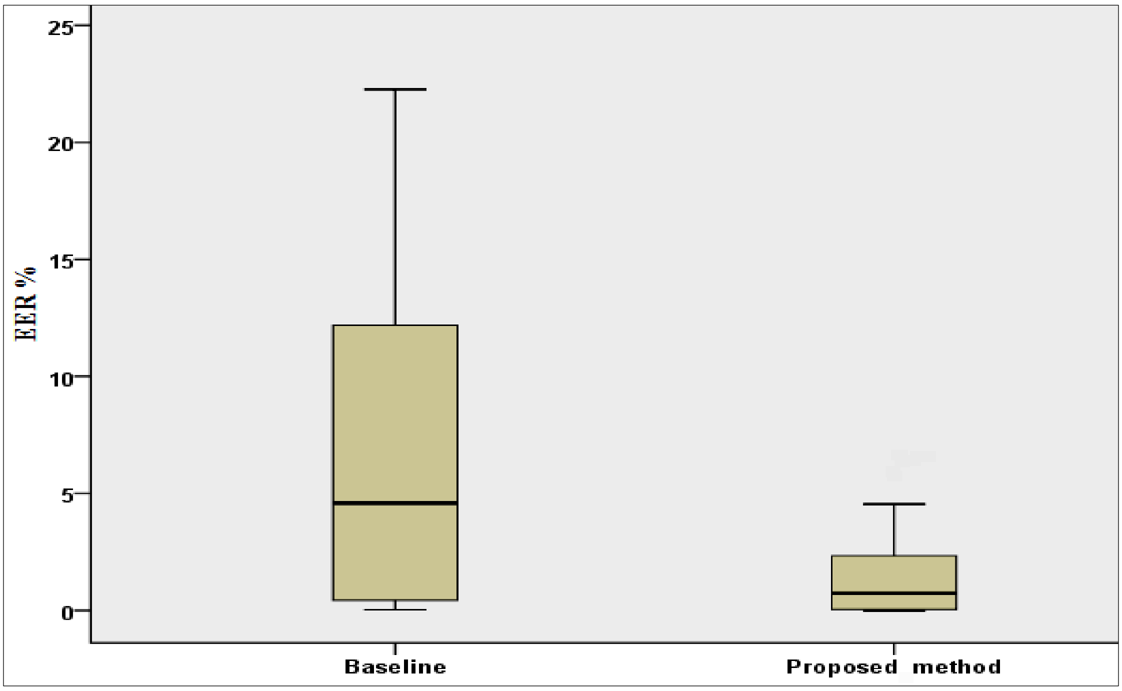
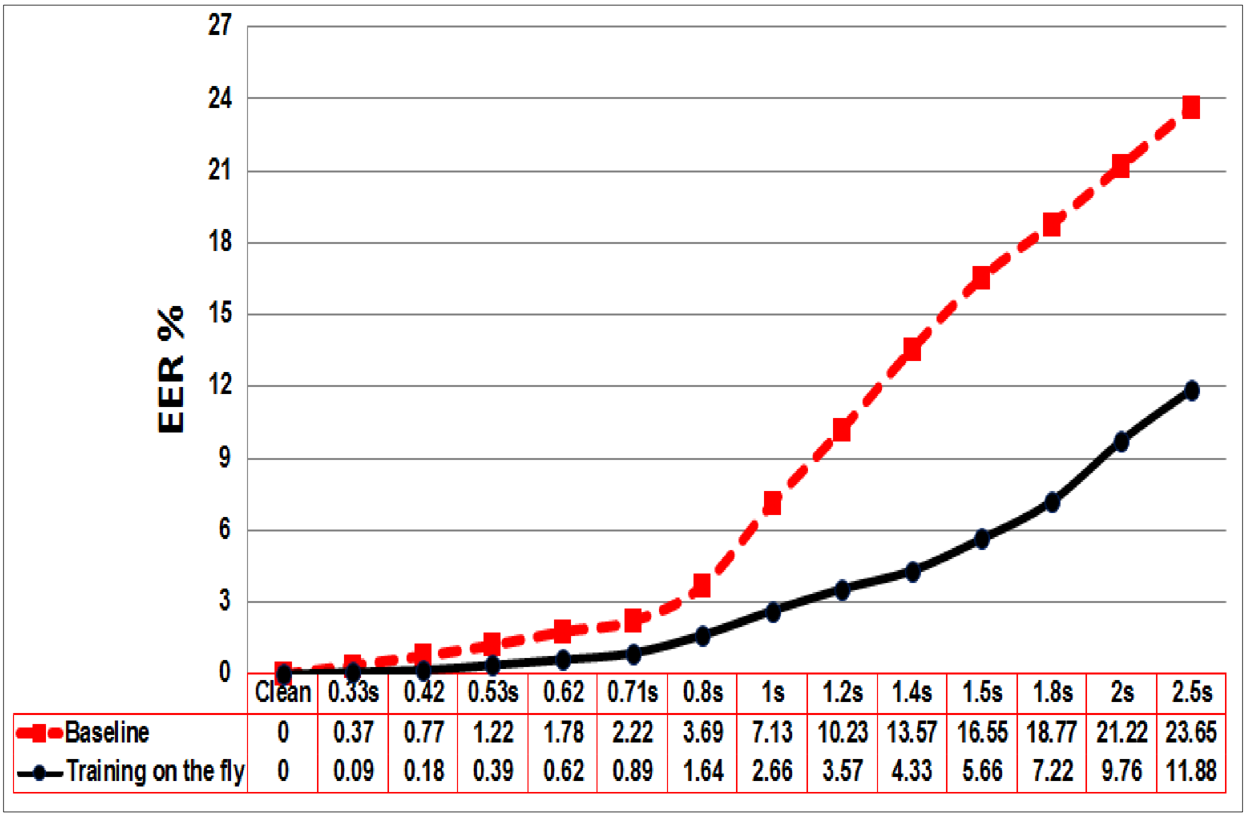


Figure 6.5 Boxplots of system performance for both methods corresponds to the standard deviation according to the percentage EER for each method performance

Figure 6.6 illustrates the result of authenticating any one speaker against the remaining 99 speakers, for baseline methods and the proposed method. Each speaker model has been tested 10 times, in which one represents the true speaker and the remaining nine are impostors. This figure corresponds to the standard deviation according to the percentage EER for each speaker model during 10 times testing using both methods with different reverberation time. In this figure, the x-axis represents the reverberation time level, and the y-axis represents the percentage EER. Each set of speaker models characterises a unique reverberant condition and is used independently for speaker recognition. As illustrated in Figure 6.6, results show different trends, the training on the fly schema outperforming the baseline result. “The proposed method significantly improves the results, especially in the cases when the reverberation time is longer”. The baseline result provided the highest EER (23.65%) with RT=2.5s, while the training on the fly method was lower (11.88%). Moreover, the baseline provided the lowest EER (0.37%) with RT=0.33 (against 0.09% of the training on the fly method). “The proposed method significantly improves the results especially in the case when the reverberation time is longer than 1.0s”. The improvement in the system performance with different reverberation times is shown in Table 6.3. Overall, the variation between the baseline result and the training on the fly method result is high. “Therefore, this indicates that there is a significant improvement in the verification performance”. “Furthermore, detection error trade-off curves plotted in Figure 6.7, 6.8, 6.9, 6.10, 6.11, 6.12 and 6.13 shows the false negative (rejection) rate and false positive (acceptance) rate for the proposed method are better with different reverberation time values”. In some cases, the false negative (rejection) rate for both methods is close. It can be seen that the accuracy of false positive rate (FPR) for the training-on-the-fly system (the solid blue line) shows significant improvement compared to the traditional system, especially when reverberation time tends to be longer. More results depending on the DET curves with different reverberation times are shown in Appendix C.



	Clean	0.33s	0.42	0.53s	0.62	0.71s	0.8s	1s	1.2s	1.4s	1.5s	1.8s	2s	2.5s
Baseline	0	0.37	0.77	1.22	1.78	2.22	3.69	7.13	10.23	13.57	16.55	18.77	21.22	23.65
Training on the fly	0	0.09	0.18	0.39	0.62	0.89	1.64	2.66	3.57	4.33	5.66	7.22	9.76	11.88

Figure 6.6 The system performance using simulated impulse response

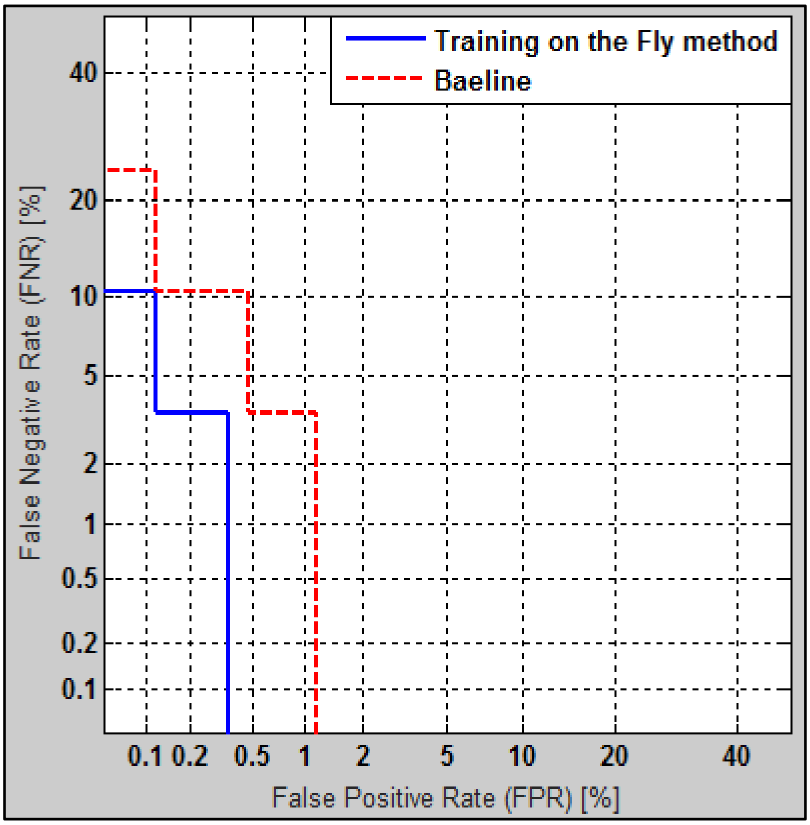


Figure 6.7 The DET curves for reverberation time 0.53s

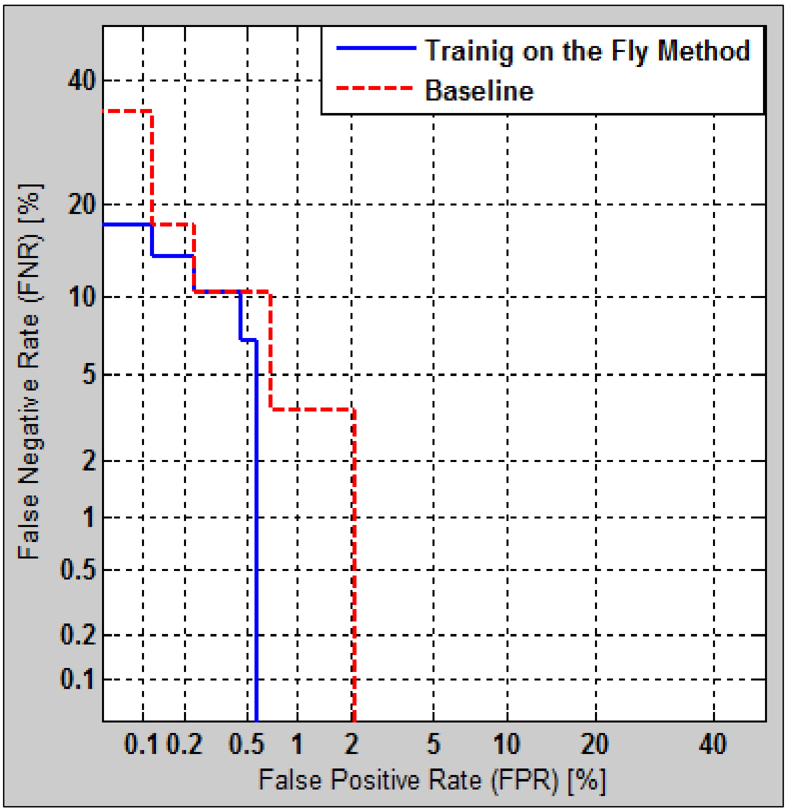


Figure 6.8 The DET curves for reverberation time 0.61s

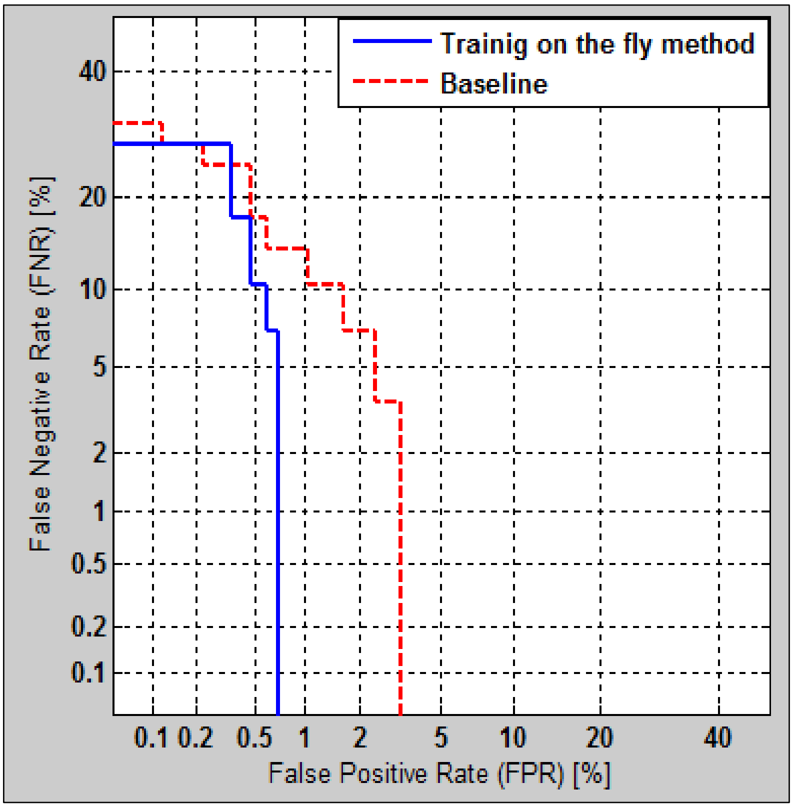


Figure 6.9 The DET curves for reverberation time 0.71s

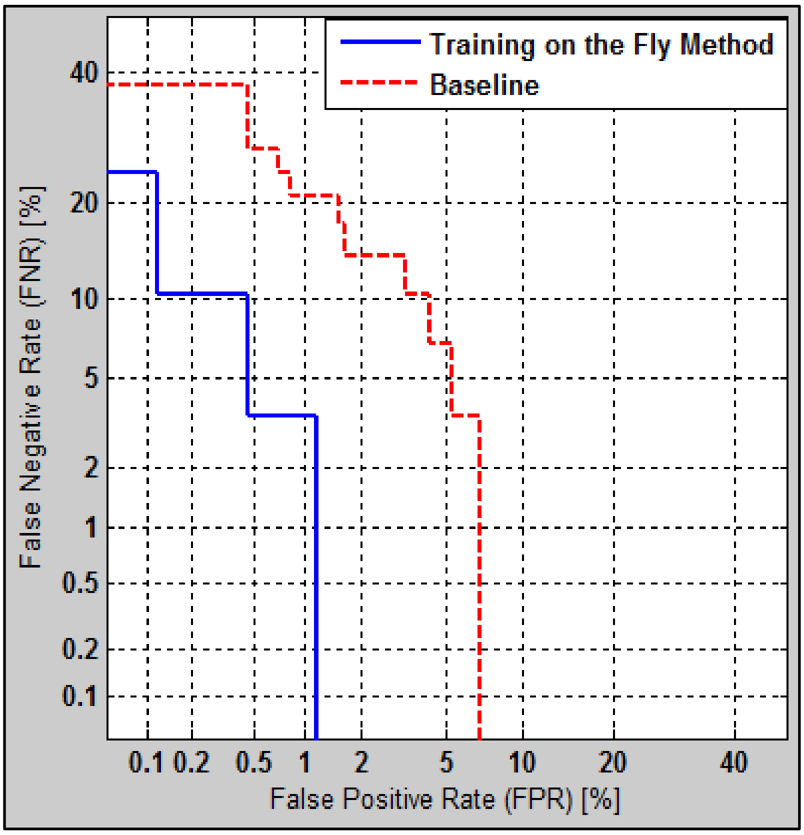


Figure 6.10 The DET curves for reverberation time 0.83s

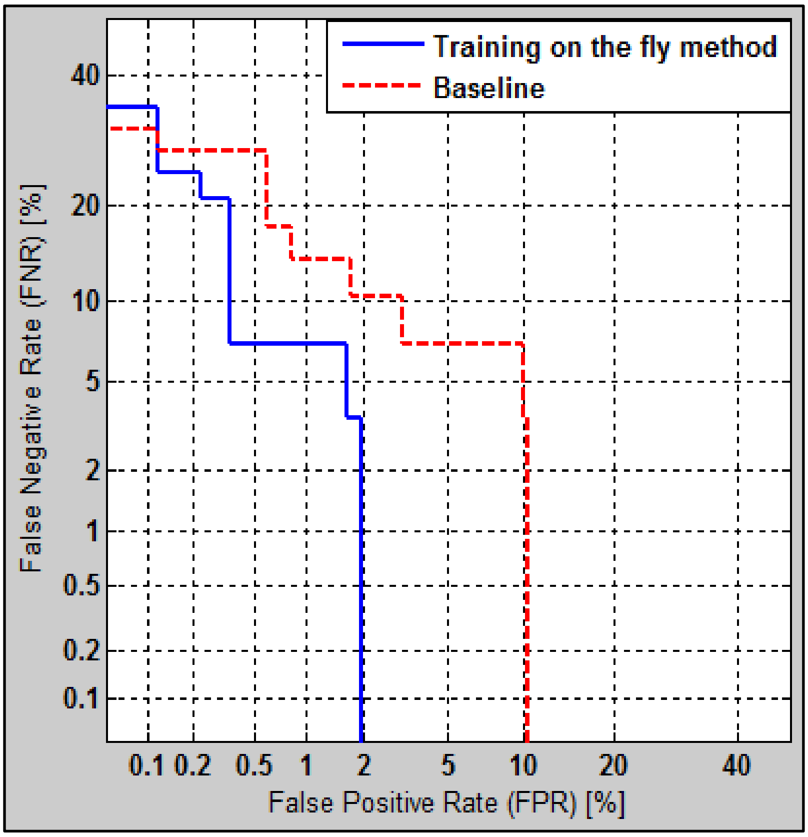


Figure 6.11 The DET curves for reverberation time 1.0s

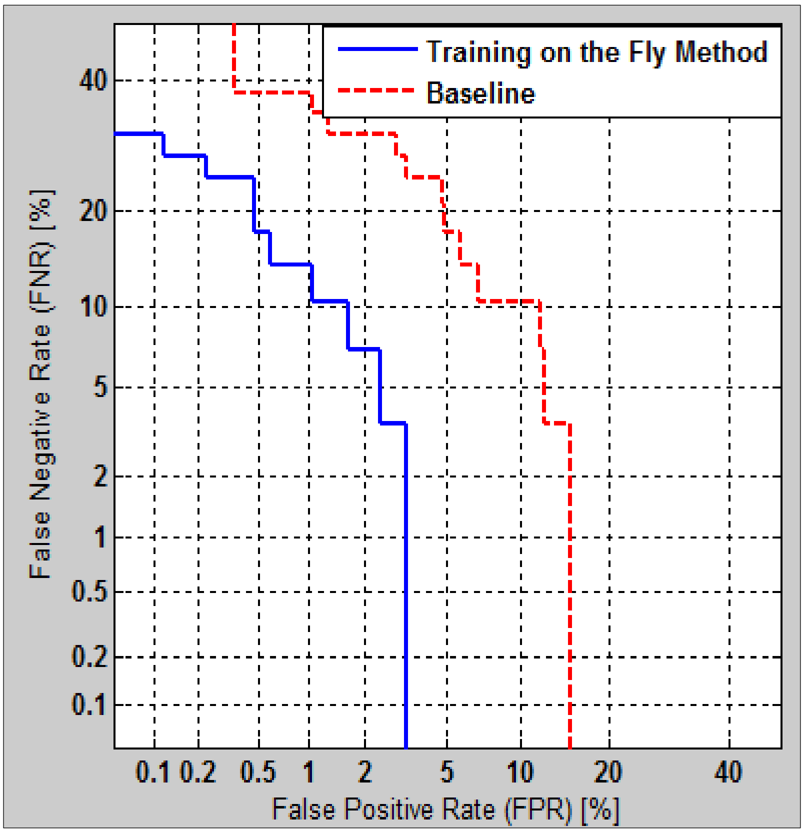


Figure 6.12 The DET curves for reverberation time 1.2s

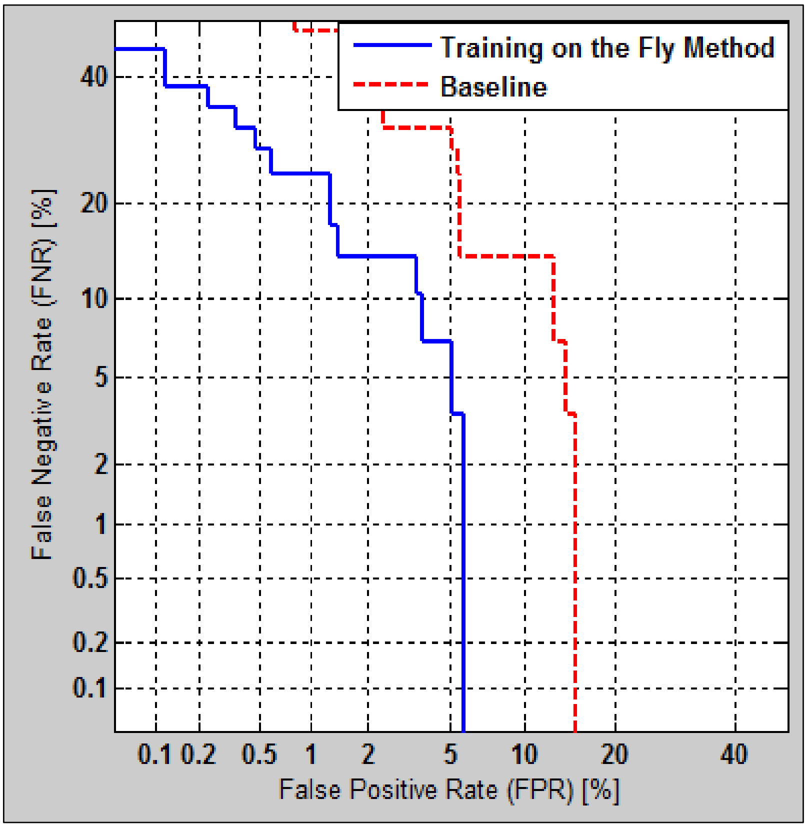


Figure 6.13 The DET curves for reverberation time 1.5s

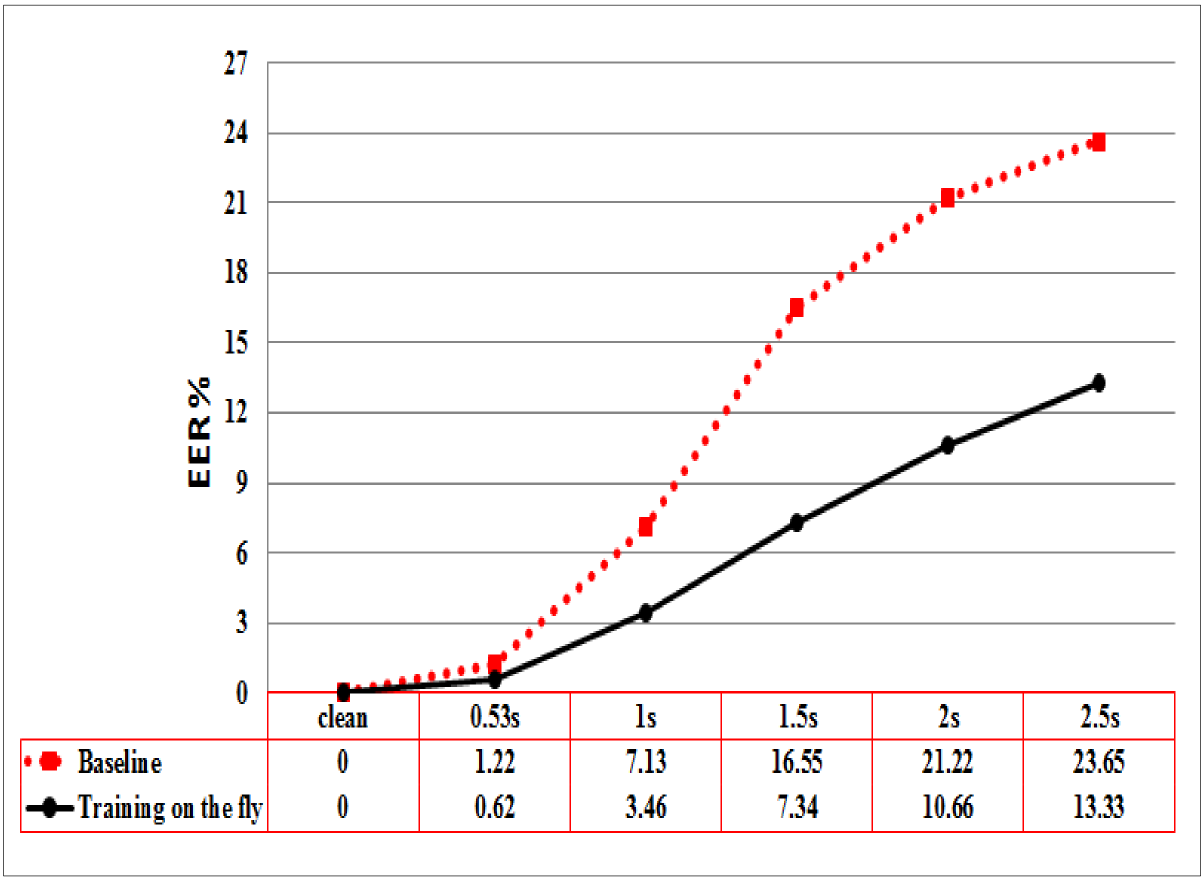
Table 6.3 The performance improvement with different reverberation times

Methods	**Reverberation Time(s)**										
	0.33	**0.53**	**0.7**	**0.8**	**1**	**1.2**	**1.4**	**1.5**	**1.8**	**2**	**2.5**
Baseline	99.6	98.7	97.7	96.3	92.8	89.7	86.4	83.4	81.2	78.7	76.3
Proposed method	99.9	99.6	99.1	98.3	97.3	96.4	95.6	94.3	92.7	90.2	88.1
Improvement %	0.28	0.83	1.3	2.0	4.4	6.6	9.2	10.8	11.5	11.4	11.7

6.4.7 Evaluation the Method with Real Impulse Responses

The results reported so far are generated using simulated RIRs. For cross-validation to the proposed method, we now test our system using RIRs recorded in real rooms to assess its utilities in real environments. We use the RIRs obtained from the Aachen Impulse Response (AIR) database which, offers a wide range of reverberant conditions (168 RIRs with T60 ranging from 0.1 s to about 4 s) (Jeub et al., 2009).There are five RT60 (0.53, 1, 1.5, 2 and 2.5 second) and 20 RIRs are collected from each room corresponding to different microphone positions. We use the 10-impulse response in training stage while the 10-impulse response was used in a testing stage. Clean speech samples were obtained from SALU-AC dataset. Speech samples acquired in the enrolment phase were convolved with these impulse responses.

Comparing the results of the proposed system using simulated impulse response and real impulse response, “ Figure 6.14”, the performance of the proposed method has been reduced when real impulse responses used. This indicates that the real acoustic environments are more challenging than simulated ones for speaker recognition. Overall, the proposed system outperforms the traditional systems in all the test conditions.



	clean	0.53s	1s	1.5s	2s	2.5s
Baseline	0	1.22	7.13	16.55	21.22	23.65
Training on the fly	0	0.62	3.46	7.34	10.66	13.33

Figure 6.14 The system performance using real impulse response

6.5 Early Reflections sound to Improve the Speaker Recognition Performance

In this part, use of the autocorrelation function (ACF) is proposed to find the early reflections from speech signals submitted for verification in the first stage of this experiment. The estimates are convolved with an anechoic signal for the use in the training of the system in the second stage. For channel matching, matching the channels with ones used in testing phases is adopted in this experiment. “Experimental results have shown significant improvement in system performance regarding reduced equal error rate and detection error trade-off”. This work was published by the IEEE conference (International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing EECCMC/ 2018). The proposed system is illustrated in Figure 6.15

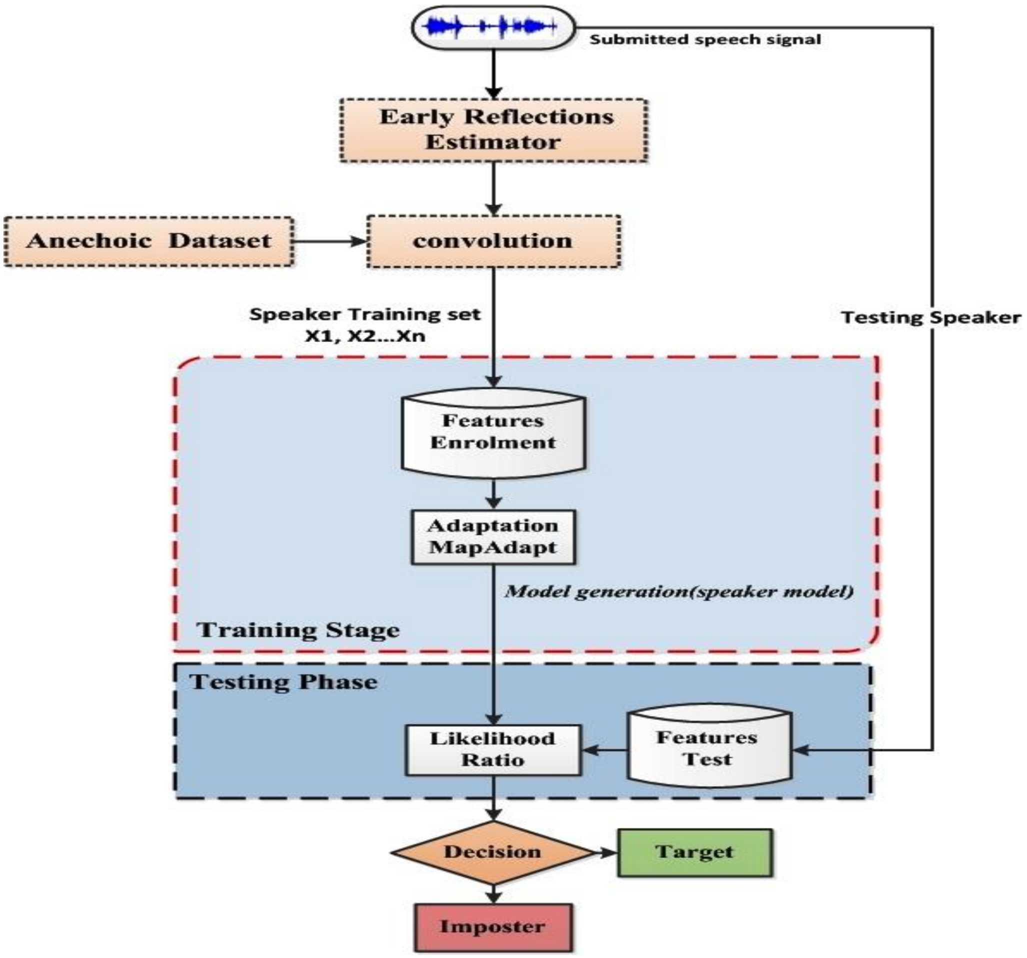


Figure 6.15 The proposed system

6.5.1 Early Reflection

The early reflections denote the sounds that arrive at the listener after being reflected once or twice from surfaces within the listening space, such as walls, ceilings and floor. They arrive later than the direct sound, often in a range from 5 to 100 milliseconds, but arrive before the onset of diffuse reverberation (H. Kuttruff, 1979). Early reflections can be easily recognised by their intensity and lower density because they are mainly reflections that rebounded just once or twice from the walls or ceiling of the room (Ristić, Pavlović, Pavlović, & Reljin, 2013). Early reflections should be modeled accurately, and it provides spatial information about the room acoustic characteristics and has a significant effect on our experience of sound in. In literature (Bork, Goerne, & Potratz, 2005; Havelock, Kuwano, & Vorländer, 2008; Noxon, 1992) it was noted that the position of the reflection in the impulse response is associated with a particular perception of sound in the room. Reflections arriving within the first few milliseconds immediately after the direct sound are responsible for the perception of the arrival direction of the sound, i.e. the position of the sound source. Early reflections are in general not harmful in speech/speaker recognition and can have a positive impact on the intelligibility (Bradley et al., 2003; H. Kuttruff, 1979; Omologo et al., 1998). Received speech typically benefits from the energy boost produced by replicas of the same signal arriving at the microphone within a limited time delay (Petrick et al., 2007). "Conversely, the reverberation tail critically affects the speech and speaker recognition behaviour (Sehr & Kellermann, 2010), due to the resulting time smearing, phonemes are mixed up with the preceding ones". Figure 6.17 shows this component of the reverberant signal. Hitherto, a number of works (Arweiler & Buchholz, 2011) have investigated the impact of early reflections on intelligibility, but a corresponding analysis used this technique to mitigate reverberation effects and improve the robustness of speaker recognition is still missing.

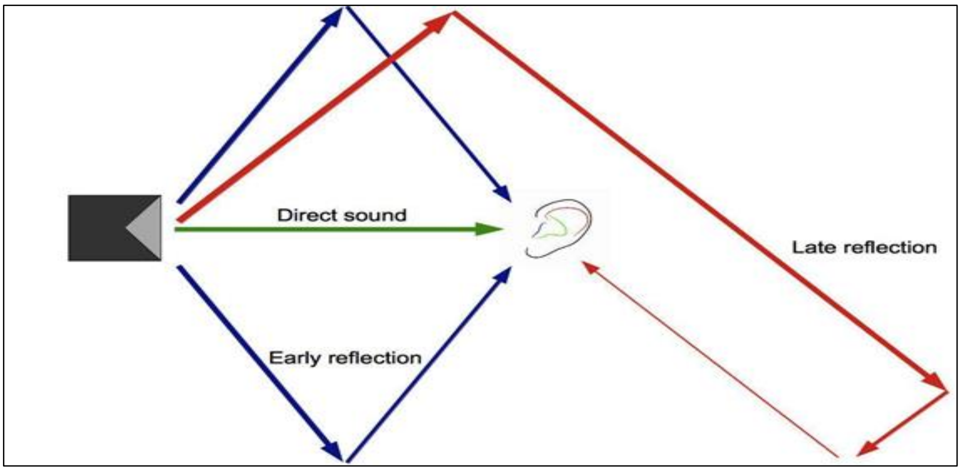


Figure 6.16 Different types of reflections

6.5.2 Autocorrelation Function

Correlation is a matching process, autocorrelation, also known as serial correlation, it refers to the matching of a signal with a delayed version of itself as a function of delay. It is utilised to compare a signal with a time-delayed version of itself. If a signal is periodic, then the signal will be perfectly correlated with a version of itself if the time-delay is an integer number of periods. Informally, it is the similarity between observations as a function of the time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analysing functions or series of values, such as time domain signals. Mathematically, the autocorrelation corresponding to a delay time τ is calculated by

- finding the value of the signal at a time t,
- finding the value of the signal at a time $t + \tau$,
- multiplying those two values together,
- repeating the process for all possible times, t, and then computing the average of all those products.

The process can be repeated for (all) other values of τ, resulting in an autocorrelation, which is a function of the delay time τ. Mathematically, for a continuous signal, $s(t)$, the autocorrelation, $R_x(\tau)$ is calculated using (Suits, 2015)

$$R_x(\tau) = \int_{-\infty}^{\infty} s(t)s(t+\tau)dt \qquad -\infty < \tau < \infty$$

6.8

The autocorrelation function $R_x(\tau)$ provides a measure of how closely the signal matches a copy of itself as the copy is shifted τ units in time. $R_x(\tau)$ is not a function of time, it is the only function of time differences τ between the waveform and its shifted copy. Sometimes it is convenient if the overall amplitude of the result is scaled so that the amplitude of the autocorrelation for $\tau = 0$ is 1, e.g. $R(0) = 1$. For that choice, then when $\tau = 0$ the signal must be perfectly correlated because the signal is compared with an exact copy of itself. If for any larger values of τ, the value of autocorrelation was also equal to 1, then that means that the signal delayed by a time τ is identical to the signal with no delay. In that case, the signal must be periodic.

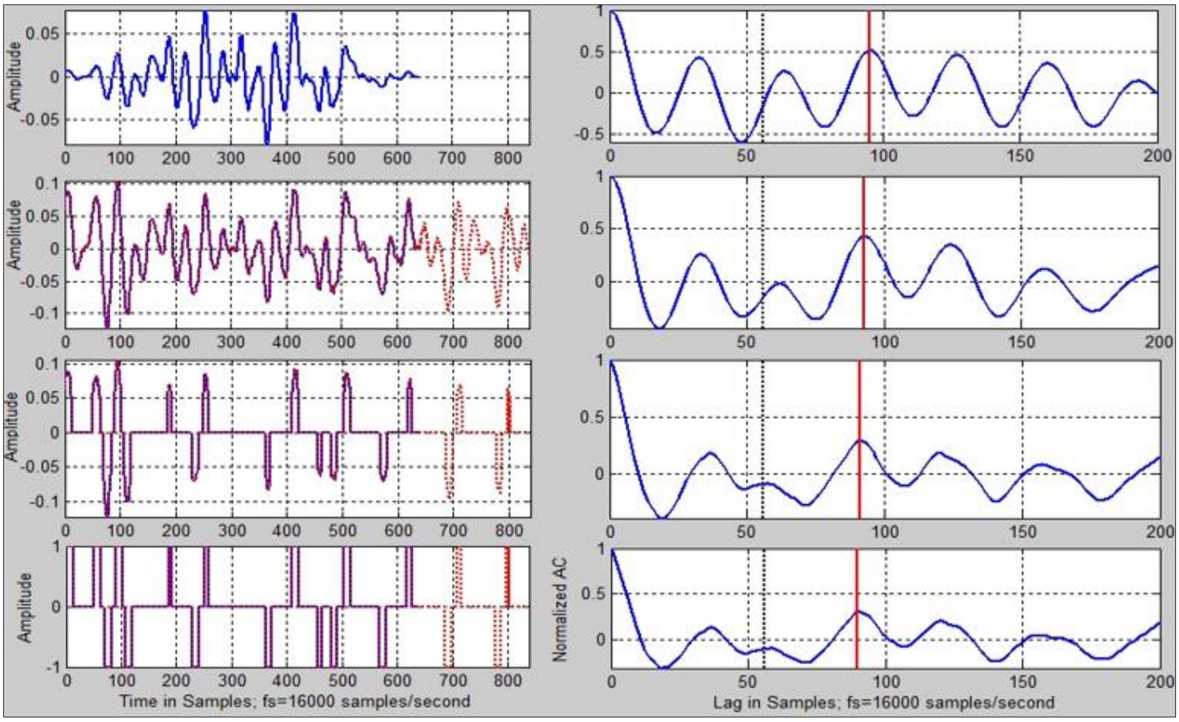


Figure 6.17 The original signal and the signal with applying the autocorrelation function

6.5.3 Detection of Early Reflections using Autocorrelation Function

To summarise, the following steps have been followed to calculate the early reflections:

- A signal contaminated with reverberation is firstly framed at about 50ms (800 samples at a sampling rate of 16000 kHz) as framed signal $x(n)$. Then, a Hamming window $w_t(n)$ is applied to the framed signal $x(n)$, which is expressed by

$$y(n) = w_t(n)x(n) \qquad 6.9$$

- Apply the aforementioned autocorrelation Equation (6.9) on the signal and a delayed copy of itself as a function of delay. Figures 6.18 and 6.19 demonstrated the ACF of speech samples with two different size window (50 ms and 100 ms) respectively, both of them are contaminated by RT equal to nearly (0.71s).
- Then the early reflections sound can be detected by determining the lag times that reflect the highest peaks as highlighted by the spikes in Figure 6.21 (d)
- Finally, in this study, the computed early reflection vector is convoluted later on with the anechoic samples to generate a reference model for training stage, more details in next section.

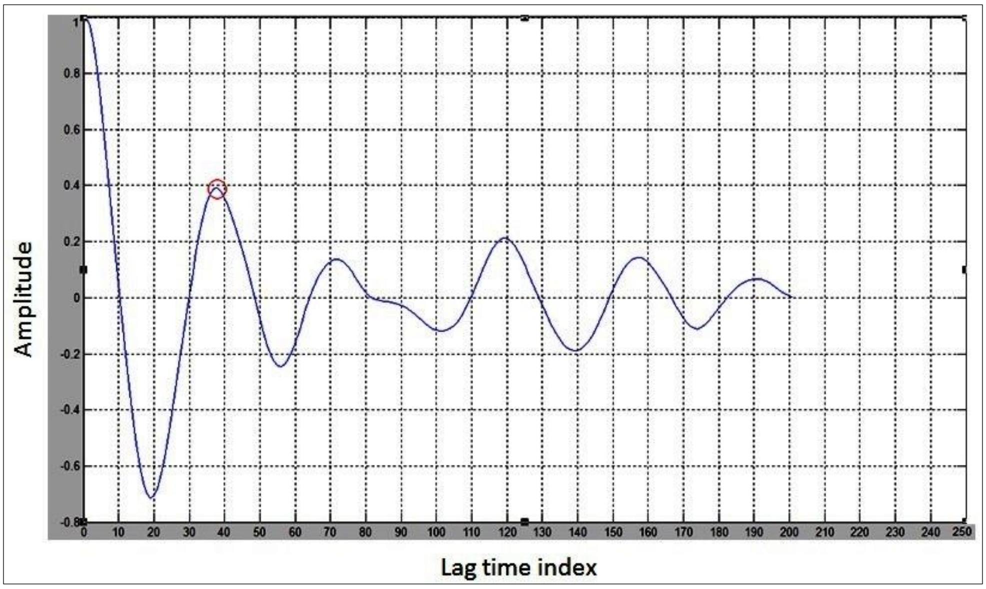


Figure 6.18 The ACF of speech samples with window size (50 ms)

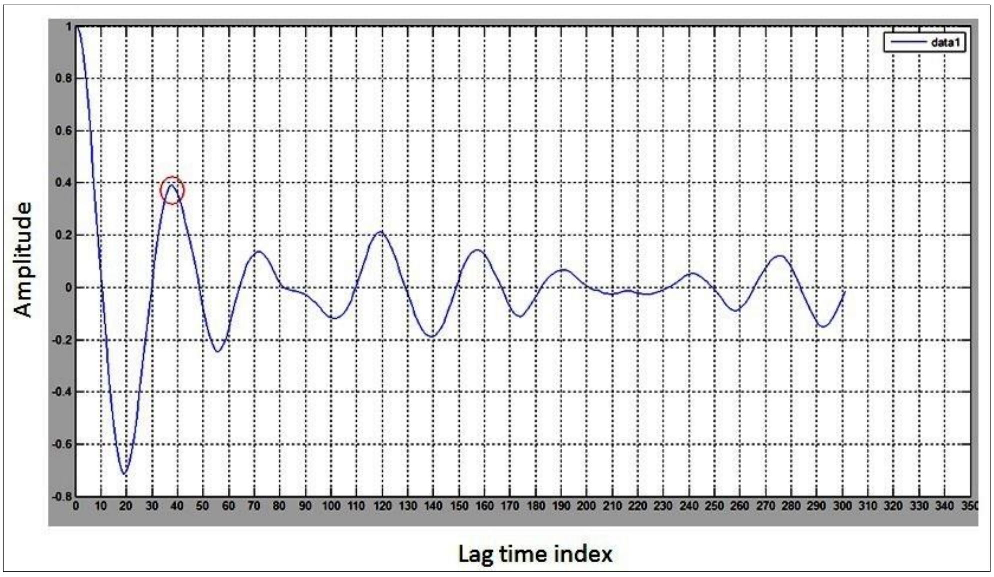


Figure 6.19 The ACF of speech samples with window size (100 ms).

Furthermore, Figure 6.20 demonstrates the process of estimating the early reflection times. Where (a) shows the waveform of a framed input signal *x* (*n*), (b) represents the estimated reflected sound reference using ACF and (c) shows part of the frame with its corresponding early reflection time reference (RF).

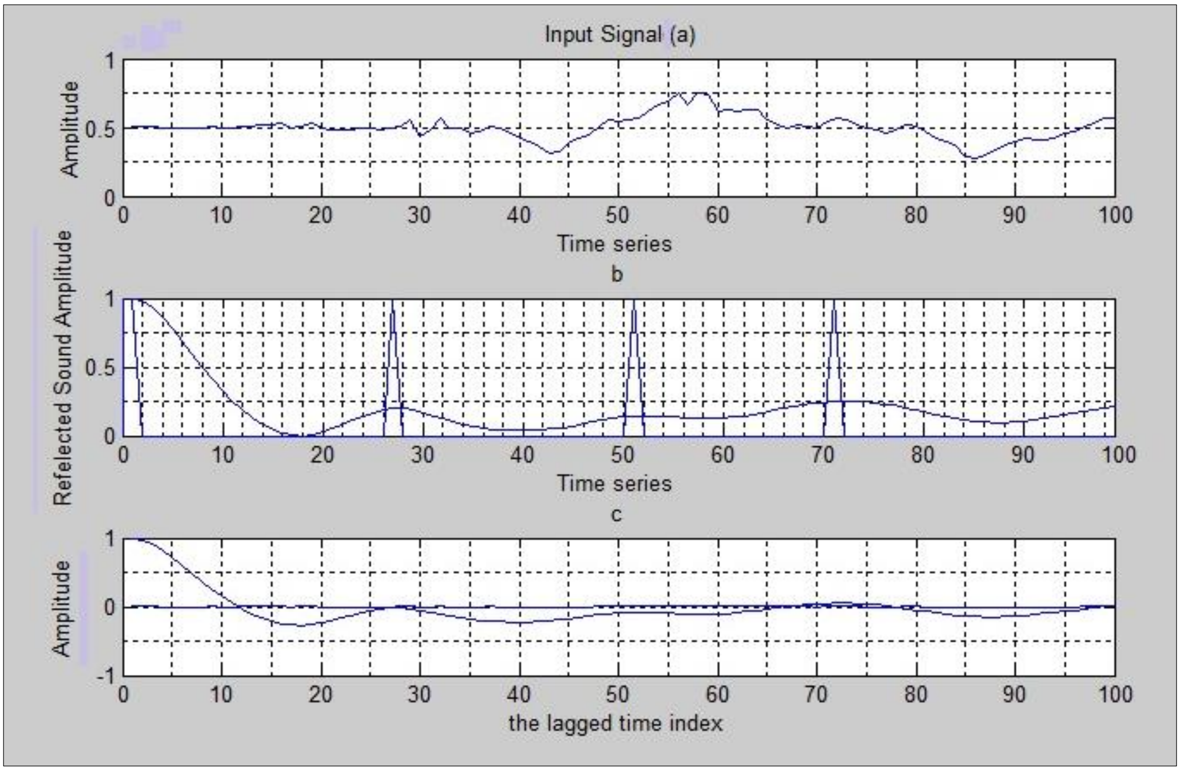


Figure 6.20 The process of estimated the early reflections from different frames

6.5.4 Experiment Setup

The current method involves an investigation of the effects of using the early reflection in the training stage to mitigate the impact of reverberation on the speaker verification performance. Experiments were carried out to validate the proposed method. The clean speech samples were obtained from 100 speakers, (50 male, 50 female) from the SALU-AC dataset. Each speaker provided 10 utterances; each utterance has a 5s duration of approximately 50s in total duration; 8 utterances (40s) were selected for the training phase with the remaining 2 (10s) used for the testing phase to produce exclusive training and test data sets. Each speaker model has been tested 10 times, in which one represents the true speaker and the remaining nine are impostors. The test audio is built by taking, for each speaker, the remaining two utterances. As common List data, the structure has been built, associating each speaker model with its correct test and with nine impostor tests; the companion key file has also been built. The impulse response used in this work was obtained from the Aachen Impulse Response (AIR) database (and down-sampled to 16 kHz) (Jeub et al., 2009). From this data set, four different rooms, comprising studio booth, meeting room, office, and lecture room are used. This gives a useful database of the impulse response and represents a range of different spaces with reverberation times from 0.2 to 2.5 seconds. Anechoic speech samples acquired in enrolment phase are convolved with the estimated early reflection time. The performance of the proposed method was evaluated using equal error rate (EER) values and the detection trade-off (DET) curve.

6.5.5 Experiment and Results Discussion

“In this method, the objective is to identify if the use of estimated early reflection can usefully determine a matched virtual channel for training and improve the system reliability”. The performance of the verification system was compared with the result of the baseline. In this method, early reflections sound is estimated from the submitted speech signal and then convoluted with anechoic speech signals to train the system and create a reference model for each speaker. Figure 6.21 depicts the box plot obtained with the baseline result and the proposed method. Each speaker model has been tested 10 times, in which one represents the correct speaker and the remaining nine are impostors. This simplest possible box plot displays the full range of variation (from min to max), and it corresponds to the standard deviation according to the percentage EER for each speaker model during 10 times testing using both methods with different reverberation time. The benchmark appears to have a larger variety than the proposed method. Moreover, it is indicated that the proposed method produces the best performance compared with the baseline.

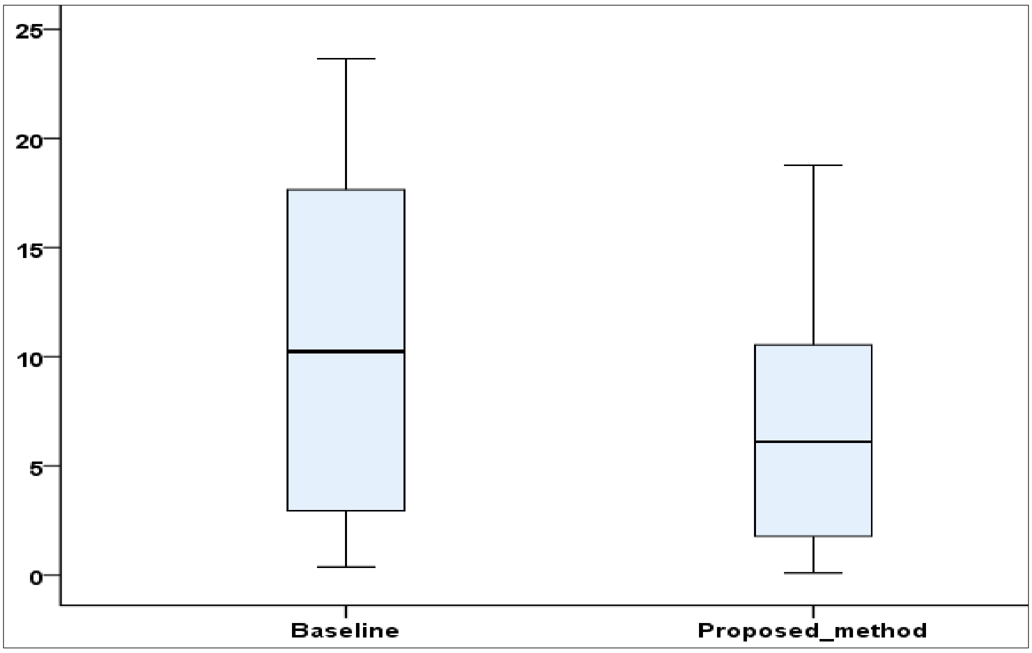
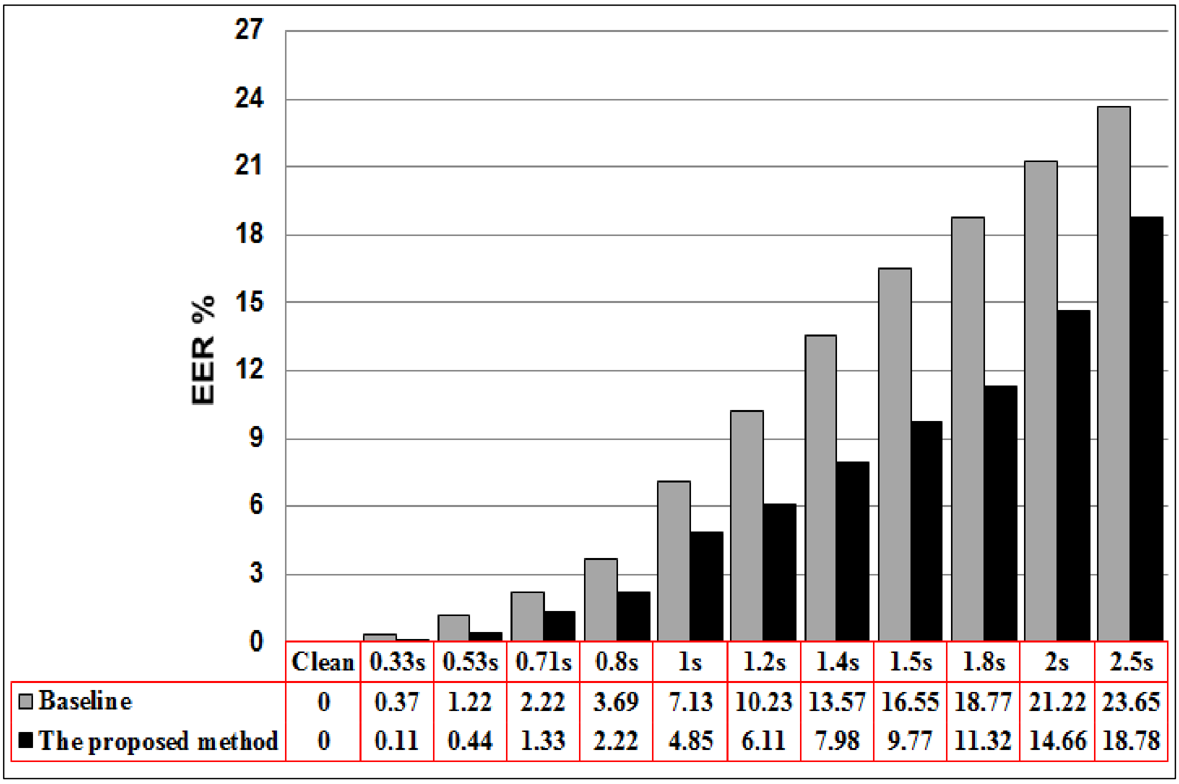


Figure 6.21 Boxplots of system performance for both methods corresponds to the standard deviation according to the percentage EER for each method.

Furthermore, “Figure 6.22 illustrates the result of authenticating any one speaker against the remaining 99 speakers, for baseline methods and the proposed method. In this figure, the x-axis represents the reverberation time level, and the y-axis represents the percentage EER”. As illustrated in Figure 7.22, results show different trends, the proposed method outperforming the baseline result. Overall, the variation between the baseline result and the training on the fly method result is significant. “Therefore, this indicates that there is a significant improvement in the verification performance. Furthermore, the detection error trade-off curves plotted in Figures 6.23, 6.24, and 6.25 shows the false negative (rejection) rate and false positive (acceptance) rate for the proposed method are better with different reverberation time values. Even in some cases, the false negative (rejection) rate for both methods is closed. The improvement in the system performance with different reverberation times is shown in Table 6.4”.



	Clean	0.33s	0.53s	0.71s	0.8s	1s	1.2s	1.4s	1.5s	1.8s	2s	2.5s
▨ Baseline	0	0.37	1.22	2.22	3.69	7.13	10.23	13.57	16.55	18.77	21.22	23.65
■ The proposed method	0	0.11	0.44	1.33	2.22	4.85	6.11	7.98	9.77	11.32	14.66	18.78

Figure 6.22 System performance using both methods with different reverberation time

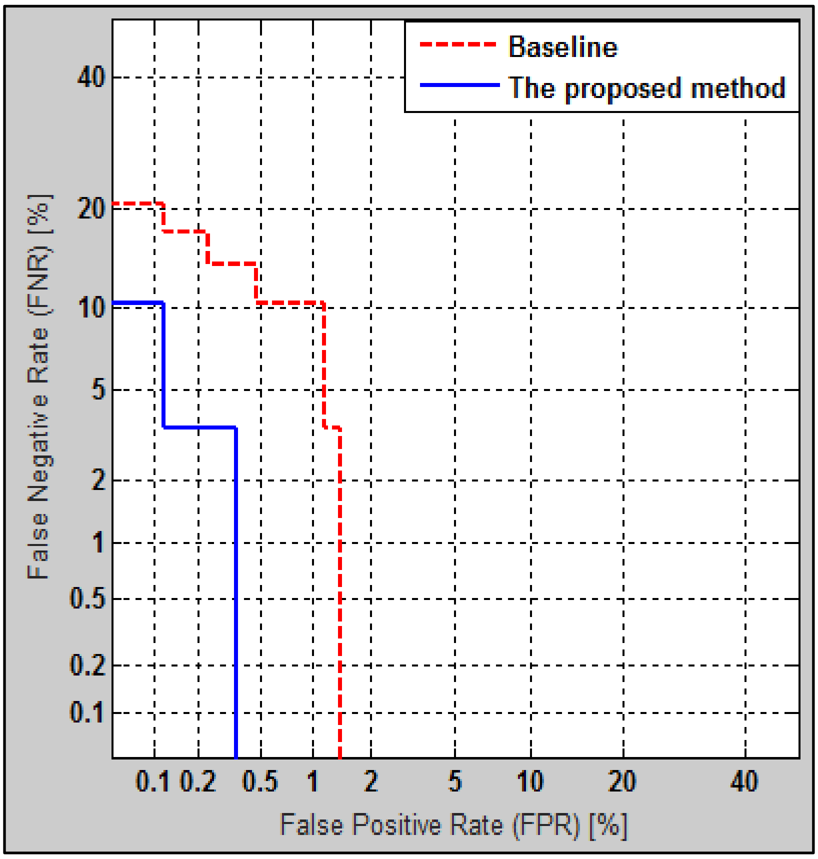


Figure 6.23 The DET curves for reverberation time 0.53s

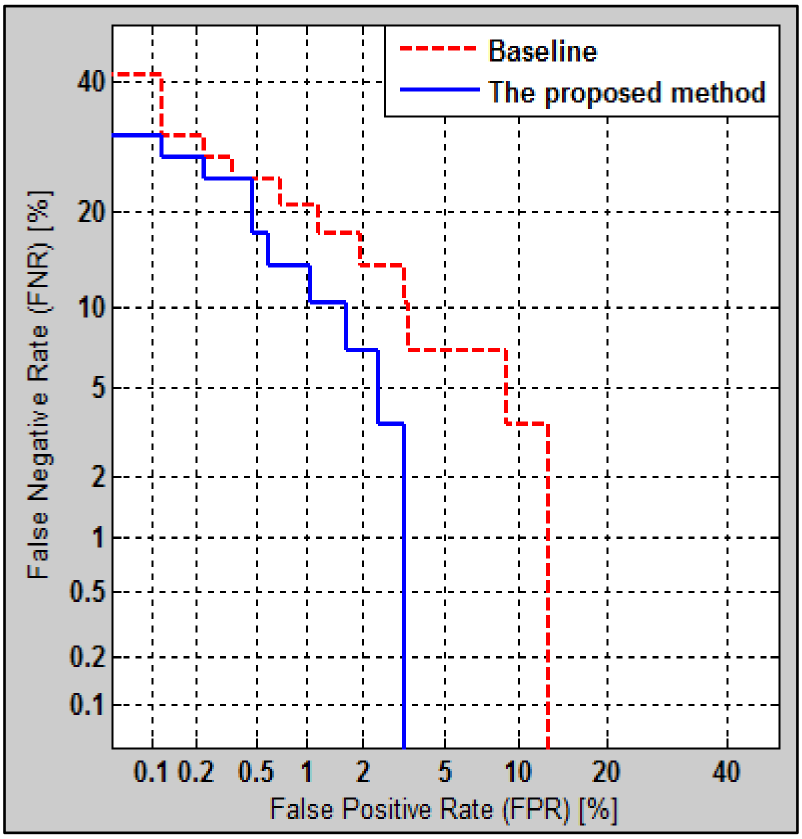


Figure 6.24 The DET curves for reverberation time 1.0s

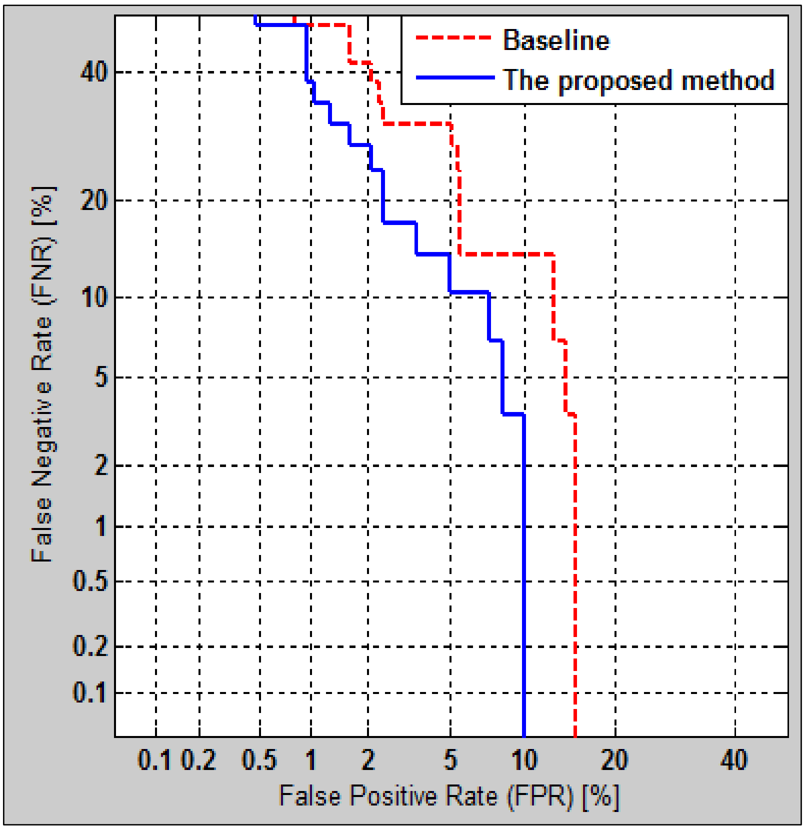


Figure 6.25 The DET curves for reverberation time1.5s

Table 6.4 The performance improvement with different reverberation time

Methods	**Reverberation Time(s)**										
	0.33	**0.53**	**0.7**	**0.8**	**1**	**1.2**	**1.4**	**1.5**	**1.8**	**2**	**2.5**
Baseline	99.6	98.7	97.7	96.3	92.8	89.7	86.4	83.4	81.2	78.7	76.3
Proposed method	99.8	99.5	98.6	97.7	95.1	93.8	92.0	90.2	88.6	85.3	81.2
Improvement %	0.2	0.7	0.8	1.4	2.2	4.1	5.5	6.7	7.4	6.5	4.8

From the results vary between the baseline system and the proposed method; the conclusion can be listed as follows:

- The early reflection sound has a significant improvement in the verification performance.
- The suggested method has shown promising results and can be considered as an applicable solution for mitigating the impact of the reverberation on the system performance.

6.6 The estimated RT and ELR with Training on the Fly Method to improve the Speaker Recognition Performance

In the past, speaker recognition performance in reverberant environments has been mainly associated with the reverberation time T_{60} or the distance between the source and the microphones (Kingsbury, 1998), keeping all the other factors affecting the room impulse response fixed (source directivity and orientation, room dimensions and wall absorption coefficients).

The combined effects of reverberation time and early to late ratio have been studied in in this work, and the results indicate that they pose a greater challenge than individual effects. This study addresses the combined effects in the domain of robust speaker verification. “This study proposes to train the system on the fly during its recognition operation”. For each submitted speech signal for recognition, the acoustic transmission channels that the signal has passed through are estimated by two critical parameters, in this case, “reverberation time (RT)” and “early to the late ratio (ELR)”, and these parameters are used to create a channel model or select a channel model from pre-stored one in a bank. “The clean speech sample collected through the enrolment phase is passed through the virtual channel model are used as examples to train the system on the fly”. The proposed system is illustrated in Figure 6.26. The dotted line in the first part indicates the training on the fly part. On receiving a submitted speech for recognition, the ELR and RT are first estimated, these are used to synthesise a model of the virtual channel or to choose a closet matched one from a channel model bank. Thus, a matched virtual channel is created. The proposed method provides training samples through an estimated and matched virtual channel. Simple Euclidean distance criteria are used to select the best match from the database. In this study, the objective is to identify if the use of estimated ELR and RT can usefully determine a matched virtual channel for training on the fly and improve the system performance.it would be ideal to avoid the use of same channel models for the training and testing.

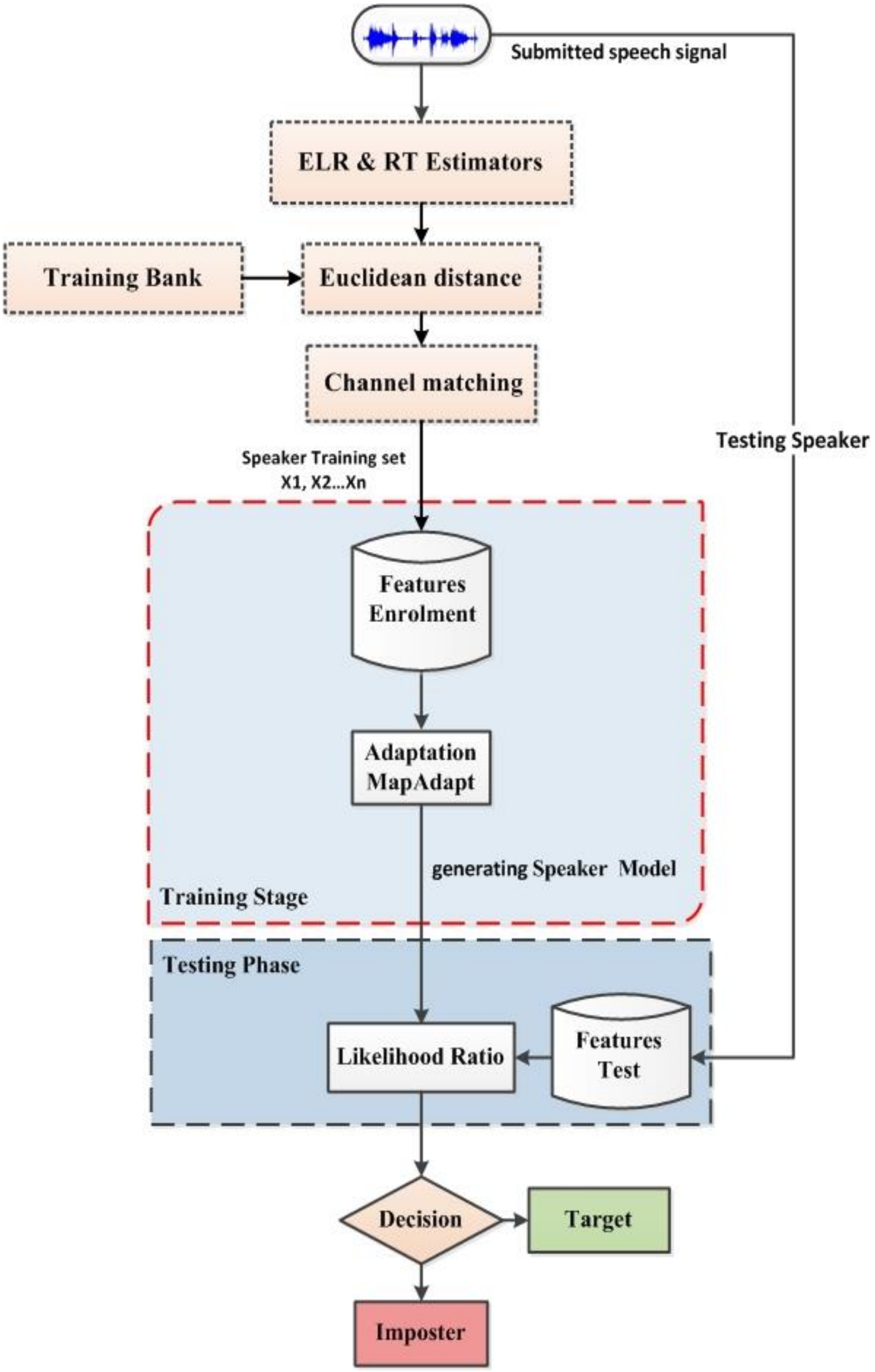


Figure 6.26 The framework of the proposed system

6.7 Virtual Acoustic Channel Creation

With the estimated early to late ratio and the reverberation times in sub-bands, it is possible to create a matched virtual channel for the training of the system on the fly.

6.7.1 Estimated Early-to-Late Ratio

A representation based on the combination of the early reflections and tail reflection is hence proposed to estimate the early to late ratio. In enclosures, acoustic waves propagate from a source to the acquisition device through multiple paths due to the presence of reflecting surfaces (e.g., walls, furniture). It is often convenient to split the impulse response into three parts (H. Kuttruff, 1979) each of them affecting the emitted signal in different ways:

$$h(\tau) = h_d(\tau) + h_e(\tau) + h_r(\tau) \qquad 6.10$$

where $h_d(\tau)$ is the anechoic direct propagation path, $h_e(\tau)$ describes the early arrivals up to some tens of milliseconds and $h_r(\tau)$ denotes the late diffuse reverberation typical of the impulse response tail. Ideally, the best propagation channel consists only of the direct path, which just introduces attenuation A_d and a delay τ_d to the submitted signals:

$$h_d(\tau) = A_d \delta(\tau - \tau_d) \qquad 6.11$$

Early reflections can have a positive impact on the intelligibility of speech (Bradley et al., 2003; H. Kuttruff, 1979; Omologo et al., 1998), typically benefit from the energy boost produced by replicas of the same signal arriving at the microphone within a limited time delay. The ratio between the energy associated with $h_d(\tau) + h_e(\tau)$ and $h_r(\tau)$ becomes a possible way to characterise the RIR influence on speaker recognition performance.

Lately, the Direct-to-Reverberant Ratio (DRR) has become a common method to measure the amount of distortion presented by a given room impulse response, independently of the specific environment and experimental setup.

The Direct-to-Reverberant Ratio measures the ratio between the energy propagating along the direct path (i.e. without reflections) and the reverberant energy (Naylor & Gaubitch, 2010).

$$DDR = \frac{\int_{\tau} h_d(\tau)^2 d\tau}{\int_{\tau} h_e(\tau) + h_r(\tau))^2 d\tau} \qquad 6.12$$

The metric is mostly utilised in dereverberation or speech enhancement, either to measure the performance or to characterise the experimental environments. Nevertheless, since the speaker recognition to benefit from early arrivals, therefore we consider a generalised Early-to-Late Reverberation Ratio (ELR), defined as follows:

$$ELR = 10\log_{10} \frac{\int_{\tau=0}^{T} h(\tau)^2 d\tau}{\int_{\tau=T}^{\infty} h(\tau)^2 d\tau} \qquad 6.13$$

where *T* defined the time instant when we divided between early and late arrival. Essentially, it is a generalisation of the clarity C_{80} utilised to characterise the music transparency in concert halls (Heinrich Kuttruff, 1991). Note that the DRR is a particular case of the ELR.

6.7.2 Reverberation

A maximum likelihood estimation algorithm is proposed for blind-estimation of reverberation time from speech signals submitted for verification. A cluster of impulse responses with the estimated reverberation time is selected. The reverberation times involved in the submitted speech samples in the same octave bands were estimated. Simple Euclidean distance criteria are used to select the best match from the database. The channel matched speech signal $s'(t)$ acquired by

$$s'(t) = s(t) \otimes h(t) + n(t) \qquad 6.14$$

Where $s(t)$ is the clean speech, $h(t)$ is the impulse responses chosen from the database and $n(t)$ is the estimated noise. More details on the estimation reverberation time in section 7.1

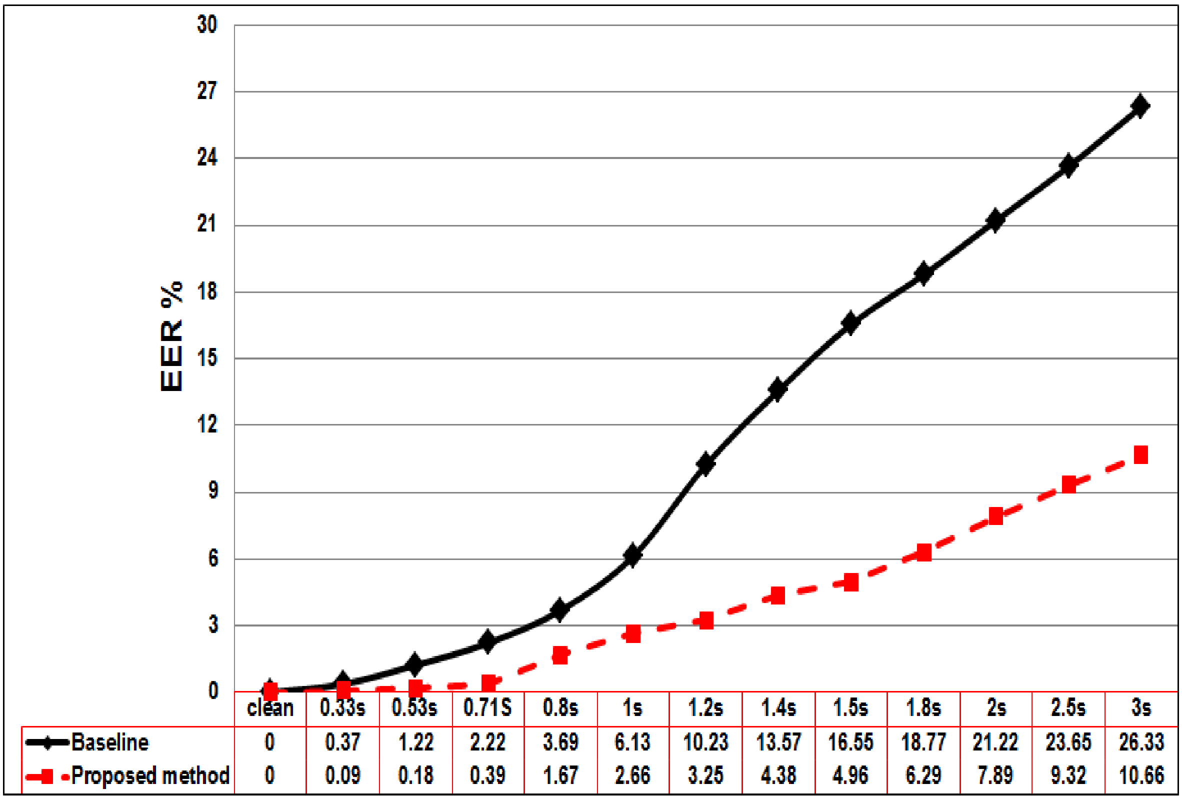
6.7.3 Experiment Setup

Experiments were carried out to validate the proposed method. Clean speech samples were obtained from 100 speakers, (50 male, 50 female) from the SALU-AC dataset. Each speaker provided 20 utterances; each utterance has a 5s duration of approximately 100s in total duration; 10 utterances (50s) were selected for the training phase with the remaining 10 (50s) used for the testing phase to produce exclusive training and test data sets. The commercial software CATT-Acoustic was employed to generate synthetic impulse responses from 50 rooms with different dimensions and acoustic properties. This gives a useful database of the impulse response. Each impulse was labelled by its reverberation time in seven-octave bands: 125, 250, 500, 1K, 2K, 4K and 8 kHz. The reverberation times involved in the submitted speech samples in the same octave bands were estimated. These RT values of the selected rooms are representative of a wide range of typical acoustic scenarios, and taking into account that the ASR deteriorates for distant speech applications, seven different distances were used in each room. Source-receiver distances of 0.5, 1, 2, 3, 4, 5, 6 and 7 m were used. In this work, the impulse response dataset represents a range of different spaces with broadband reverberation times from 0.23 to 3 seconds. Note that the test utterances are different from utterances used for training. In addition, a simple percentage error rate is not adequate to indicate the performance of the system since false acceptance, and false rejection has different impacts on the system. In speaker recognition and other biometric security systems, the EER (equal error rate) and DET curve are often used as a combined single measure for error. Therefore, the performance of the proposed method was evaluated using equal error rate (EER) values and the detection trade-off (DET) curve.

6.7.4 Experiment and Results Discussion

To evaluate a universal method for reference matching is not an easy task. The lack of a standardised benchmark regime makes comparisons to other work difficult. In this experiment, the objective was to identify whether the use of estimated RT and ELR can usefully determine a matched virtual channel for re-training and improve the system reliability. It would be ideal to avoid the use of similar channel models for the training and testing. In this work, the proposed method was validated utilising text-independent speaker recognition testbed based on the Microsoft Research (MSR) identity toolbox (Sadjadi et al., 2013). The objective parameters were estimated from the reverberated signal. The baseline system was created using MSR toolbox and GFCC features. The estimation methods that the speech signal has passed through was estimated by the critical parameters called reverberation time (RT) and early to the late ratio (ELR), and these parameters were utilised to create a reference model or select a reference model from pre-stored ones in a bank. The closest match from the pre-stored data set of reverberation times ranging from 0.33s to 3.0s was selected. Figure 6.27 illustrates the result of authenticating any one speaker against the remaining 99 speakers, for baseline system and the proposed method. In this figure, the x-axis represents the reverberation time level, and the y-axis represents the percentage EER. Each set of speaker models characterises a unique reverberant condition and is used independently for speaker recognition. As illustrated in Figure 6.27, results show different trends, the training on the fly outperforming the baseline result. The proposed method significantly improves the results. The baseline result provided the highest EER (26.33%) with RT=3s, while the training on the fly method was lower

(10.66%). Moreover, the baseline provided the lowest EER (1.22%) with RT=0.53 (against 0.09% of the training on the fly method). The improvement in the system performance with different reverberation times is shown in Table 6.5. “Overall, the variation between the baseline result and the training on the fly method result is high. Therefore, this indicates that there is a significant improvement in the verification performance. Furthermore, detection error trade-off curves are plotted in Figures 6.28, 6.29, 6.30 and 6.31 show the false negative (rejection) rate and false positive (acceptance) rate for the proposed method are better with different reverberation time values. In some cases, the false negative (rejection) rate for both methods is close”.



	clean	0.33s	0.53s	0.71S	0.8s	1s	1.2s	1.4s	1.5s	1.8s	2s	2.5s	3s
Baseline	0	0.37	1.22	2.22	3.69	6.13	10.23	13.57	16.55	18.77	21.22	23.65	26.33
Proposed method	0	0.09	0.18	0.39	1.67	2.66	3.25	4.38	4.96	6.29	7.89	9.32	10.66

Figure 6.27 System performance with both method

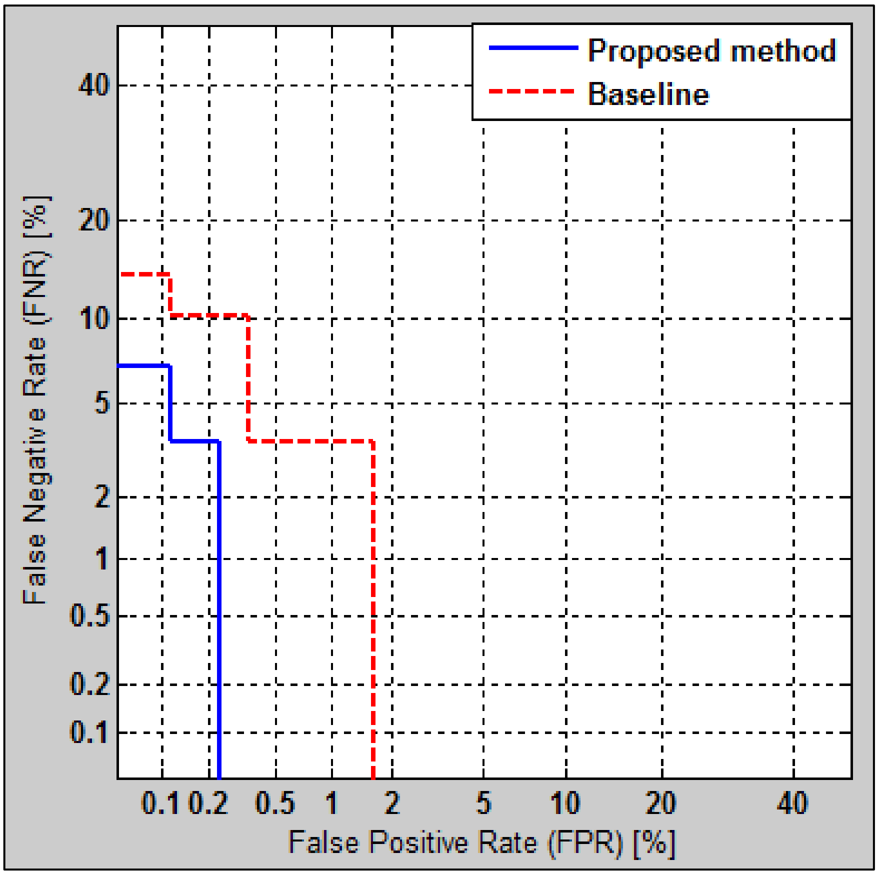


Figure 6.28 The DET curve with RT=0.53

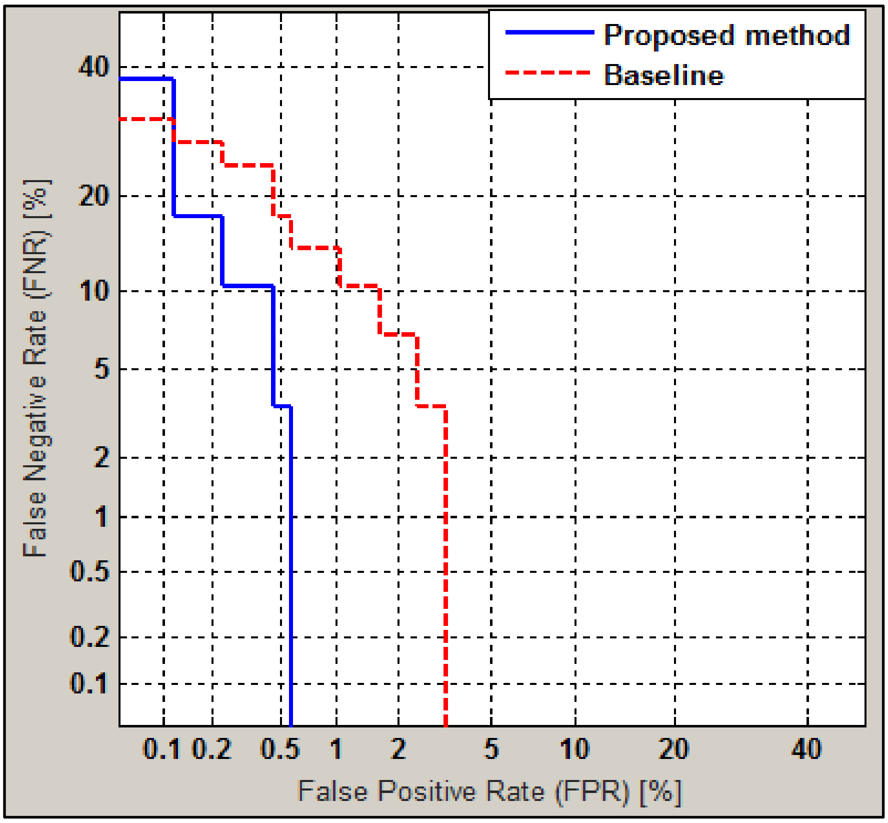


Figure 6.29 The DET curve with RT=1.0s

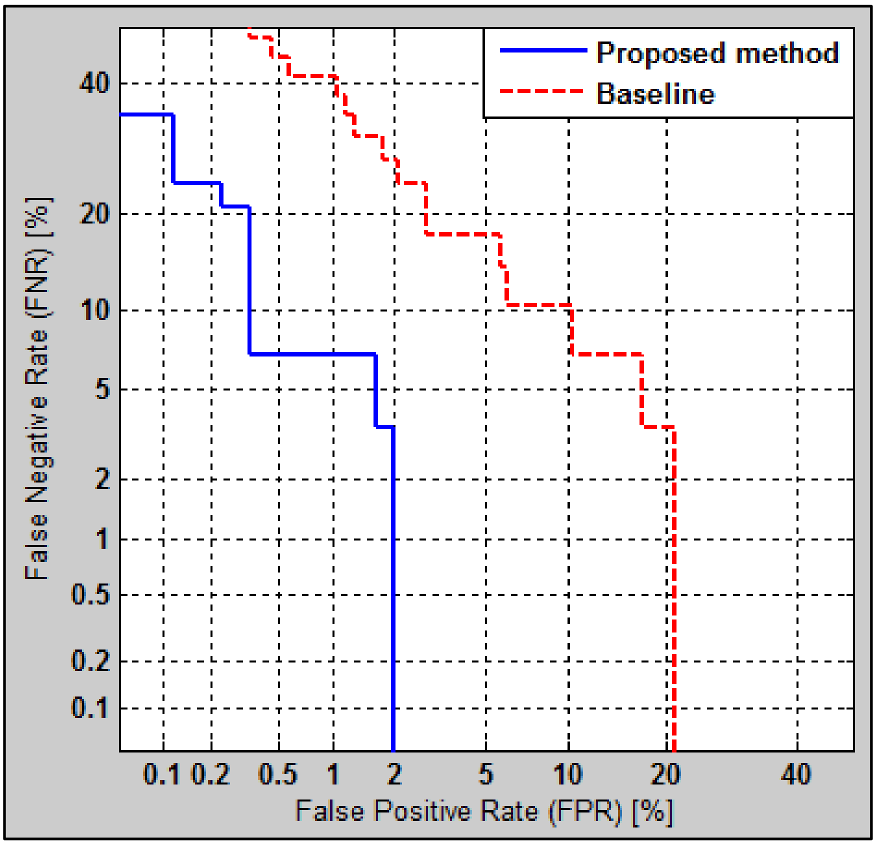


Figure 6.30 The DET curve with RT=1.5s

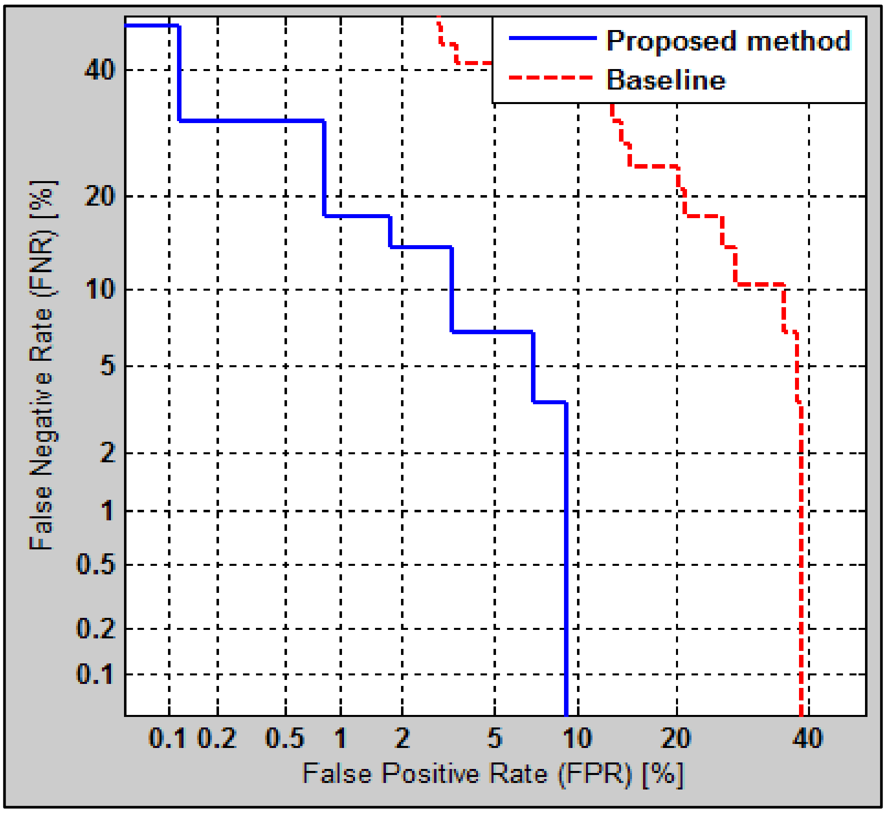


Figure 6.31The DET curve with RT=3s

Table 6.5 The performance improvement with different reverberation time

Methods	**Reverberation Time(s)**											
	0.33	**0.53**	**0.7**	**0.8**	**1**	**1.2**	**1.4**	**1.5**	**1.8**	**2**	**2.5**	**3**
Baseline	99.6	98.7	97.7	96.3	93.8	89.7	86.4	83.4	81.2	78.7	76.3	73.6
Proposed method	99.9	99.8	99.6	98.3	97.3	96.7	95.6	95.0	93.7	92.1	90.6	89.3
Improvement %	0.2	1.0	1.8	2.0	3.4	6.9	9.1	11.5	12.4	13.3	14.3	15.6

6.8 Chapter Summary

This chapter has provided a detailed background of the estimation methods employed in the literature to estimate or measure RTs. In addition, a detailed introduction of the Maximum Likelihood Estimation (MLE) based reverberation time (RT) estimation method has been provided. The concept is to use decay phases following speech utterances to estimate the decay curve using a model of sound decay. The method is inherently blind as it searches the signal for regions of free decay. Furthermore, this chapter has presented different methods to enable training the speaker recognition system with reverberant speech samples according to the estimated reverberant conditions. These are achieved by using “maximum likelihood estimation”, “early reflection estimation” and “early to late ratio estimation method”. In the first experiment, the maximum likelihood method was used to estimate the reverberation time from speech signal submitted for verification. The estimates are used to choose matched acoustic impulse responses or transfer functions for inclusion in the retraining or fine-tuning of the pattern recognition model on the fly. In the second experiment, the autocorrelation function (ACF) is proposed to find the early reflections from speech signals submitted for verification in the first stage of this experiment. The estimates are convolved with an anechoic signal for the use in the training of the system in the second stage. In the final experiment, the estimated reverberation time by maximum likelihood and the estimated early to late ratio from early and late reflection were used with training on the fly method to increase the robustness of a speaker verification system.

CHAPTER SEVEN: CONCLUSIONS AND FUTURE WORKS

In this final chapter of the thesis, the work and the results presented in the previous chapters are summarised. The overall conclusions of this study are made, and further work of academic and practical interest suggested.

7.1 Summary

The introduction of this thesis describes the general issues and applications of speaker recognition. This thesis is aimed to reduce the detrimental effects of reverberation on speaker recognition with the aim of improving the robustness of speaker verification for real-world applications in reverberant environments. “More specifically the thesis deals with methods that can help to reduce the detrimental effects of reverberation on a single microphone speech signal. Different methods have been suggested to solve the problem of reverberation effects of the speaker recognition”.

This study began with the generation and collection of anechoic speech samples (clean data set). The speech corpora can be considered of fundamental importance in testing the performance of speaker recognition techniques. The anechoic speech samples were collected from 110 (58 male, 52 female) English speakers in the anechoic chamber at Salford University and were used to conduct an experimental study of the suggested methods in this thesis. This data set was validated in different experiments.

The performance of a speaker verification system was more robust with the collected anechoic data set, which has background noise level equal to -12.4 dB.

Next, a pilot investigation of a well-developed system, namely MSR, was undertaken with the aim to evaluate its capability of verification of a speaker under clean and reverberant conditions.

The audio samples used were collected and generated by two different methods. In the first method, the samples were recorded in the reverberation room at Salford University with high reverberation time; while the image source method was used to generate samples with a different reverberation time and different room sizes. The accuracy of the system with the clean speech samples is 100%. However, the system performance with high reverberation time samples, recorded in the reverberation room, was only 5%. Clearly, system performance is degraded by an increase in reverberation time.

For example, system accuracy becomes 84.2% when RT=1.5s. This finding supports the literature, which concludes that the reverberation time and room dimensions have a significant effect on speaker recognition performance.

The study then proceeds to investigate two common features that used in the speaker recognition field with the reverberant condition, Mel Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC). It is worth noting that previous studies have primarily concentrated on using the MFCC and GFCC features in noisy conditions and there have been several studies reporting that the GFCC is more robust than MFCC in noisy conditions. The results demonstrate the superior robustness of GFCC relative to MFCC in reverberant environments. Deriving GFCC from the Cochleagram substantially improves its robustness, and produces better results than the MFCC and is presented as an effective feature representation for reverberant speaker recognition robustness. However, for speaker verification using the GFCC with reverberation times lower than 2s, a performance of around 81.3% accuracy is achievable, this is significantly reduced in the presence of a reverberation time higher than 2s. Therefore, higher efficiency in high reverberation time cases is sought.

The research then moved forward to investigate the effects of reverberation time and the microphone to source distance on the speaker recognition performance; this study has analysed the relationship between reverberation time (RT) and the source-receiver distance. One of the contributions of this research is a report on how the reverberation time and source-receiver distance can affect speaker recognition performance as well as investigating which factor has a more significant effect on system performance.

The relationship between the equal error rates EER of both factors was analysed using SPSS software, which revealed an interesting finding, which, the linear model demonstrated a strong relationship (Pearson's Correlation 0.926 for the RT and 0.399 for the distance). "Moreover, this study discovered that reverberation time is the dominant factor and has a negative relationship to the EER (%)". From the results in Figure 4.22, it is apparent that the combination of reverberation time and the source-receiver distance poses a more significant challenge than the individual parameters. In addition, the result also confirms that the effect of the source to receiver distance should be taken into account and not ignored. Moreover, the findings further confirm that degradation in the system performance in large rooms is greater than that of small and medium rooms.

The study then proceeded to investigate techniques to improve the robustness of speaker recognition and reduce the effect of reverberation via training. In this experiment, the speech samples were convolved with impulse responses with varying RT values, similar to those used in the training phase, to identify how the inclusion of reverberation can affect the performance of the system. The findings concluded that the effect of reverberation time could be combated by using reverberant speech in training. The best results are presented when the reverberation characteristics of trained and tested speech samples are as close as possible. These seem to suggest that, the inclusion of environmental conditions in the training stage can to some extent mitigate the performance degradation. Therefore, if acoustic conditions can be somehow estimated and suitably included in the pre-training of the models, the robustness of the system can be improved. However, such an approach, despite it potentially has a significant improvement and popularity has some disadvantages such as:

- Multi-conditioning training is known to be computationally expensive, and it is often difficult to get more utterances for each speaker in different environments. In addition, perfect reverberant matching is complicated in real-world applications.
- Since multi-training tends to make the system more tolerant to variations found in signals, this often results in high false acceptance error rates, making the system less secure.
- Multi-training requires the storage of several speaker models, thus may place a load on resource-constrained automatic speaker verification (ASV) applications.
- Multi training needs to estimate the reverberation time, and the estimated RT may be sensitive to additive ambient noise, thus generating erroneous RT estimates in practical everyday settings.

To conclude, the aforementioned methods, to some extent, mitigate the mismatching issue at the cost of added distortions to the vulnerable speech signals themselves.

This simultaneously distorts the transmitted signals that are crucial for speaker recognition. Furthermore, speech enhancement methods and dereverberation methods do not handle the reverberation issue well. Consequently, this study aims to estimate the acoustic parameters such as the reverberation time (RT) and early to late ratio. These parameters have been presented as to enable the training of a speaker recognition system with reverberant speech samples that are generated by convolution of clean speech with the estimated reverberant conditions (training on the fly). "The estimation process has achieved by two different methods. The first method represents “a maximum likelihood estimation” that is used to determine the reverberation time. It has been shown that the proposed method improves the reliability of speaker recognition. It is worth to note that the system improvement becomes more significant when reverberation time tends to be longer. Along with others, this method has proved that an exact match is not necessarily needed. This is supportive by significant improvement in reliability in terms of equal error rates, and detection trade-off plot.

Secondly, “autocorrelation function (ACF)” is proposed to find the early reflections from speech signals submitted for verification in the first stage. The estimates of the early reflection value were convolved with anechoic samples that can be used in the training stage. The designed method outperformed the baseline system. The conclusions from this method can be listed as follows:

- Early reflections provide a significant improvement in verification performance.
- The suggested method has shown promising results and can be considered as an applicable solution for mitigating the impact of the reverberation on the system performance.

The study continues to combine the previous two essential parameters together to estimate the early to late ratio and then use it with RT estimation simultaneously to enable the use of training on the fly techniques. The results indicate that the proposed method significantly improves the reliability in terms of equal error rates, and detection trade-off plot.

“All the documented objectives were positively met. Training on the fly methods can be considered an effective method to improve the robustness of speaker verification in reverberation condition”.

7.2 Future Work

Research work so far has indicated potential pathways to develop further the methods proposed in this thesis and extend their applications to other areas. The findings of this study also suggest some significantly new and interesting research topics:

1) The work in this thesis related to the single speaker recognition. Therefore, speaker recognition has been viewed as a problem of verifying or identifying a particular speaker in a speech segment containing only a single speaker. However, for some real applications, the problem is to verify or identify particular speakers in a speech segment containing multiple speakers. In multiple speaker scenarios, if the system cannot separate single speaker segments effectively, it will directly affect the system performance. The automatic system needs to be able to segment the speech containing multiple speakers into segments and determine whether the speech by a particular speaker is present and wherein the segment this speech occurs. In addition, the series of single speaker recognition approaches could be performed. Therefore, using speaker recognition with multiple speakers was suggested as a future work.
2) Another area that could be suggested for future study is the investigation of reverberation reduction techniques and finds suitable methods that can reduce or alleviate the impact of reverberation and increase the robustness of speaker recognition without added distortions to the vulnerable speech signals themselves.
3) Work is also required on the part of feature extraction. Another area for future research includes studies of the effects of reverberation environment linked to the development of speech processing, and speech enhancement methods compensating for the adverse effects of different environments. Therefore, An alternative method to mitigate the

impact of reverberation on speech signal can be suggested in future research using liftering (adaptive filter) in cepstrum domain embedding with the MFCC feature. The liftering operation is similar to filtering operation in the frequency domain where the desired quefrency region for analysis is selected by multiplying the whole cepstrum by the adaptive filter at the desired position. The main steps of calculating MFCC with liftering are:

- Apply pre-emphasis filter on the speech signal to address the overwhelming concentration of spectral energy in the low frequencies by emphasising or boosting the higher frequency content.
- Dividing the signal into windowed and overlapping frames
- Applying the Fast Fourier Transform.
- Taking the logarithm of the magnitude.
- Inverse Fast Fourier Transform now applies to the speech signal to compute complex
- Performing the liftering to the Cepstrum (Deconvolution)
- Converted to Mel frequency scale,
- Implement the discrete cosine transform (DCT)

Figure 7.1 shows the main steps of using liftering in cepstrum domain embedding with the MFCC feature

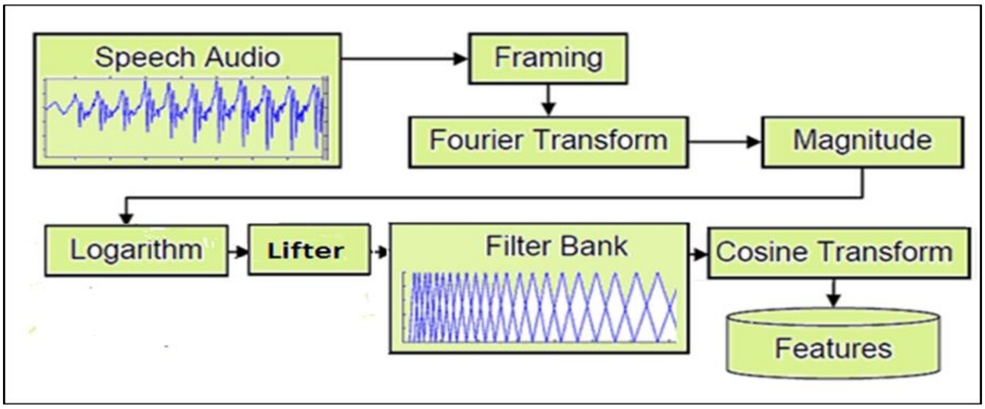


Figure 7.1 The MFCC framework with Liftering

4) Regarding estimation the reverberation time and early to late ratio, more investigation is needed using different types of the impulse response to evaluate the system. In this experiment, the concept of using both RT and ELR have been improved the system performance.

BIBLIOGRAPHY

Aachen Impulse Response Database. (2009). Retrieved from http://www.ind.rwth-aachen.de/AIR

Abdulla, W. H. (2002). Auditory based feature vectors for speech recognition systems. *Advances in Communications and Software Technologies*, 231-236.

Akula, A., Apsingekar, V. R., & De Leon, P. L. (2009). *Speaker identification in room reverberation using GMM-UBM.* Paper presented at the Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th.

Akula, A., & de Leon, P. (2008). *Compensation for room reverberation in speaker identification.* Paper presented at the Proc. Eur. Signal Process. Conf.

Al-Karawi, K. A., Al-Noori, A. H., Li, F. F., & Ritchings, T. (2015). Automatic Speaker Recognition System in Adverse Conditions--Implication of Noise and Reverberation on System Performance. *International Journal of Information and Electronics Engineering, 5*(6), 423.

Al-Noori, A. H., Al-Karawi, K. A., & Li, F. F. (2015). *Improving Robustness of Speaker Recognition in Noisy and Reverberant Conditions via Training.* Paper presented at the Intelligence and Security Informatics Conference (EISIC), 2015 European.

Aldrich, J., R. A. Fisher. (1997). The making of maximum likelihood 1912-1922. *Statistical Science, 12*(3), 162-176.

Allen, J. B., & Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, *65*(4), 943-950.

Arweiler, I., & Buchholz, J. M. (2011). The influence of spectral characteristics of early reflections on speech intelligibility. *The Journal of the Acoustical Society of America, 130*(2), 996-1005.

Assaleh, K. T. (1995). *Supplementary orthogonal cepstral features.* Paper presented at the Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on.

Assaleh, K. T., & Mammone, R. J. (1994a). New LP-derived features for speaker identification. *Speech and Audio Processing, IEEE Transactions on, 2*(4), 630-638.

Assaleh, K. T., & Mammone, R. J. (1994b). *Robust cepstral features for speaker identification.* Paper presented at the Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on.

Atal, B., Schroeder, M., & SEssLER, G. (1965). Subjective reverberation time and its relation to sound decay ICA-5. *G-32, Liittich*.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America, 52*(6B), 1687-1697.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America, 55*(6), 1304-1312.

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America, 50*(2B), 637-655.

Atlas, L., & Shamma, S. A. (2003). Joint acoustic and modulation frequency. *EURASIP journal on applied signal processing, 2003*, 668-675.

Avila, A. R., Sarria-Paja, M., Fraga, F. J., O'Shaughnessy, D., & Falk, T. H. (2014). *Improving the Performance of Far-Field Speaker Verification Using Multi-*

Condition Training: The Case of GMM-UBM and i-vector Systems. Paper presented at the Fifteenth Annual Conference of the International Speech Communication Association.

Becker, T., Jessen, M., & Grigoras, C. (2008). *Forensic speaker verification using formant features and Gaussian mixture models.* Paper presented at the Interspeech.

Beigi, H. (2009). *Effects of time lapse on speaker recognition results.* Paper presented at the Digital Signal Processing, 2009 16th International Conference on.

Beigi, H. (2011). *Fundamentals of speaker recognition*: Springer Science & Business Media.

Beigi, H. (2012). Speaker Recognition: Advancements and Challenges.

Benesty, J., Sondhi, M. M., & Huang, Y. (2008). *Springer handbook of speech processing*: Springer.

Berkley, D. A., & Mitchell, O. M. M. (1974). Seeking the ideal in "hands-free" telephony. *Bell Lab. Rec, 52*, 318-325.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute, 4*(510), 126.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., . . . Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP journal on applied signal processing, 2004*, 430-451.

Bloom, P. J. (1982). Evaluation of a dereverberation technique with normal and impaired listeners. *British journal of audiology, 16*(3), 167-176.

Borgstrom, B. J., & McCree, A. (2012). *The linear prediction inverse modulation transfer function (LP-IMTF) filter for spectral enhancement, with applications to speaker* *recognition.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.

Bork, I., Goerne, T., & Potratz, U. (2005). *Designing Early Reflection Patterns Suitable for Audio Recordings by Means of Acoustic Modeling.* Paper presented at the Audio Engineering Society Convention 118.

Botteldooren, D. (1995). Finite- difference time- domain simulation of low- frequency room acoustic problems. *The Journal of the Acoustical Society of America, 98*(6), 3302-3308.

Bradley, J., Sato, H., & Picard, M. (2003). On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America, 113*(6), 3233-3244.

Bricker, P., Gnanadesikan, R., Mathews, M., Pruzansky, S., Tukey, P., Wachter, K., & Warner, J. (1971). Statistical techniques for talker identification. *Bell System Technical Journal, 50*(4), 1427-1454.

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language, 20*(2), 230-275.

C-DAC. "Automatic Speaker Recognition using Voice Biometric "[Online]. Available: https://www.cdac.in/index.aspx?id=cs_bi_Speaker_Recognition.

Campbell, J. P. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE, 85*(9), 1437-1462. doi:10.1109/5.628714

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language, 20*(2), 210-229.

Castellano, P. J., Sradharan, S., & Cole, D. (1996). *Speaker recognition in reverberant enclosures.* Paper presented at the Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on.

CATT-Acoustic. (2010). v8.0c, Room acoustic modelling software. Retrieved from http://www.catt.se

Charoenruk, D. (2012). Communication research methodologies: Qualitative and quantitative methodology: Tailândia: University of Thai Chamber.

Chen, Y.-W., & Lin, C.-J. (2006). Combining SVMs with various feature selection strategies *Feature extraction* (pp. 315-324): Springer.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357-366.

De Leon, P. L., & Trevizo, A. L. (2007). *Speaker identification in the presence of room reverberation.* Paper presented at the Biometrics Symposium, 2007.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on, 19*(4), 788-798.

Deller Jr, J. R., Proakis, J. G., & Hansen, J. H. (1993). *Discrete time processing of speech signals*: Prentice Hall PTR.

Doddington, G. R. (1971). A Method for Speaker Verification. *The Journal of the Acoustical Society of America, 49*(1A), 139-139.

Doddington, G. R., Przybocki, M. A., Martin, A. F., & Reynolds, D. A. (2000). The NIST speaker recognition evaluation–Overview, methodology, systems, results, perspective. *Speech communication, 31*(2), 225-254.

Falk, T. H., & Chan, W.-Y. (2010). Modulation spectral features for robust far-field speaker identification. *Audio, Speech, and Language Processing, IEEE Transactions on, 18*(1), 90-100.

Fant, G. (1971). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations* (Vol. 2): Walter de Gruyter.

Fekkai, S., & Shafik, M. (2013). *Speaker Independent Phoneme Recognition for Information Technology Access Application Based on Fractal Dimension and the mel Frequency Cepstral Coefficients Features*. Paper presented at the International Conference in Information Technology (ICIT 2005)At: American University of Cyprus

American University of Cyprus.

Finan, R., Sapeluk, A., & Damper, R. (1996). *Comparison of multilayer and radial basis function neural networks for text-dependent speaker recognition.* Paper presented at the Neural Networks, 1996., IEEE International Conference on.

Flanagan, J. L. (1972). Speech analysis: Synthesis and perception.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 29*(2), 254-272.

Furui, S. (1997). *Recent advances in speaker recognition.* Paper presented at the Audio- and Video-based Biometric Person Authentication.

Furui, S. (2009). Selected topics from 40 years of research on speech and speaker recognition. *INTERSPEECH 2009 BRIGHTON*, 1-8.

Gammal, J. (2004). *Speaker recognition in reverberant environments.* Carleton University.

Ganapathy, S., Pelecanos, J., & Omar, M. K. (2011). *Feature normalization for speaker verification in room reverberation.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005). *Comparative evaluation of various MFCC implementations on the speaker verification task.* Paper presented at the Proceedings of the SPECOM.

Garcia-Romero, D., Zhou, X., & Espy-Wilson, C. Y. (2012). *Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.

Gong, Y. (2002). *Noise-robust open-set speaker recognition using noise-dependent Gaussian mixture classifier.* Paper presented at the Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on.

González-Rodríguez, J., Ortega-García, J., Martín, C., & Hernández, L. (1996). *Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays.* Paper presented at the Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on.

Gupta, S., Jaafar, J., Ahmad, W. F. W., & Bansal, A. (2013). Feature extraction using MFCC. *Signal & Image Processing, 4*(4), 101.

Haas, H. (1972). The influence of a single echo on the audibility of speech. *Journal of the Audio Engineering Society, 20*(2), 146-159.

Havelock, D., Kuwano, S., & Vorländer, M. (2008). *Handbook of signal processing in acoustics*: Springer Science & Business Media.

Haykin, S. (1999). Self-organizing maps: Neural networks-A comprehensive foundation 2nd Edition: Prentice-Hall.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America, 87*(4), 1738-1752.

Higgins, A., Bahler, L., & Porter, J. (1991). Speaker verification using randomized phrase prompting. *Digital signal processing, 1*(2), 89-106.

Huang, X., Acero, A., Hon, H.-W., & Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*: Prentice Hall PTR.

ISO. (3382 : 1997). "Acoustics - Measurement of the reverberation time of rooms with reference to other acoustical parameters".

Jarrett, D. P., Habets, E. A., Thomas, M. R., Gaubitch, N. D., & Naylor, P. A. (2011). *Dereverberation performance of rigid and open spherical microphone arrays: Theory & simulation.* Paper presented at the Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on.

Jeub, M., Schafer, M., Esch, T., & Vary, P. (2010). Model-based dereverberation preserving binaural cues. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(7), 1732-1745.

Jeub, M., Schafer, M., & Vary, P. (2009). *A binaural room impulse response database for the evaluation of dereverberation algorithms.* Paper presented at the 2009 16th International Conference on Digital Signal Processing.

Jin, Q. (2007). *Robust speaker recognition.* Carnegie Mellon University.

Jin, Q., Schultz, T., & Waibel, A. (2007). Far-field speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on, 15*(7), 2023-2032.

Jo, T., & Koyasu, M. (1975). *Measurement of reverberation time based on the direct-reverberant sound energy ratio in steady state.* Paper presented at the INTER-NOISE and NOISE-CON Congress and Conference Proceedings.

Jordan, V. L. (1970). Acoustical criteria for auditoriums and their relation to model techniques. *The Journal of the Acoustical Society of America, 47*(2A), 408-412.

Juang, B.-H., & Rabiner, L. (1993). Fundamentals of speech recognition. *Signal Processing Series. Prentice Hall, Englewood Cliffs, NJ*.

Kang, S. J., Kang, T. G., Lee, K. H., Cho, K., & Kim, N. S. (2014). *Reverberation and noise robust feature enhancement using multiple inputs.* Paper presented at the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Kelly, F. (2014). *Automatic Recognition of Ageing Speakers.* (Ph.D).

Kendrick, P. (2009). *Blind estimation of room acoustic parameters from speech and music signals.* University of Salford.

Kendrick, P., Li, F. F., Cox, T. J., Zhang, Y., & Chambers, J. A. (2007). Blind estimation of reverberation parameters for non-diffuse rooms. *Acta Acustica united with Acustica, 93*(5), 760-770.

Kenny, P. (2005a). Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13*.

Kenny, P. (2005b). Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13, 215*.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on, 15*(4), 1435-1447.

Kingsbury, B. E. (1998). *Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments.* University of California, Berkeley.

Kinnunen, T. (2005). Optimizing Spectral Feature Based Text-independent Speaker Recognition. Dissertations/University of Joensuu. *Computer Science. University of Joensuu.*

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: from features to supervectors. *Speech communication, 52*(1), 12-40.

Krishnamoorthy, P., & Prasanna, S. M. (2009). Application of combined temporal and spectral processing methods for speaker recognition under noisy, reverberant or multi-speaker environments. *Sadhana, 34*(5), 729-754.

Krokstad, A., Strom, S., & Sørsdal, S. (1968). Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration, 8*(1), 118-125.

Kurtovió, H. (1975). The influence of reflected sound upon speech intelligibility. *Acta Acustica united with Acustica, 33*(1), 32-39.

Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., & Shrishrimal, P. P. (2014). A comparative study of feature extraction techniques for speech recognition system. *International Journal of Innovative Research in Science, Engineering and Technology, 3*(12), 18006-18016.

Kuttruff, H. (1979). *Room Acoustics* (Vol. 2nd ed). London: Applied Science Publishers Ltd.

Kuttruff, H. (1991). On the audibility of phase distortions in rooms and its significance for sound reproduction and digital simulation in room acoustics. *Acta Acustica united with Acustica, 74*(1), 3-5.

Kuttruff, H. (1995). *Sound field prediction in rooms.* Paper presented at the Proc. 15th Int. Congr. Acoust.

Kuttruff, H. (2000). Room acoustics, ed. 4th: Spon Press, London,[England] New York, NY.

Kuttruff, H. (2009). *Room acoustics*: CRC Press.

Laitinen, M.-V., & Pulkki, V. (2012). *Utilizing instantaneous direct-to-reverberant ratio in parametric spatial audio coding.* Paper presented at the Audio Engineering Society Convention 133.

Larsen, E., Schmitz, C. D., Lansing, C. R., O'Brien, W. D., Wheeler, B. C., & Feng, A. S. (2003). *Acoustic scene analysis using estimated impulse responses.* Paper presented at the Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on.

Lehmann, E. A. "Image-source method: Matlab code" [Online]. Available:. Retrieved from http://www.eric-lehmann.com/

Li, F. F. (2002). *Extracting Room Acoustic Parameters From Received Speech Signals Using Artificial Neural Networks.* (Doctor Philosophy), University of Salford.

Li, F. F. (2016). *Robust speaker recognition by means of acoustic transmission channel matching: An acoustic parameter estimation approach.* Paper presented at the Innovative Computing Technology (INTECH), 2016 Sixth International Conference on.

Li, K.-P., & Porter, J. E. (1988). *Normalizations and selection of speech segments for speaker recognition scoring.* Paper presented at the Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on.

Lin, Q., Jan, E.-E., & Flanagan, J. (1994). Microphone arrays and speaker identification. *Speech and Audio Processing, IEEE Transactions on, 2*(4), 622-629.

Lochner, J., & Burger, J. (1961). The intelligibility of speech under reverberant conditions. *Acta Acustica united with Acustica, 11*(4), 195-200.

Löllmann, H., Yilmaz, E., Jeub, M., & Vary, P. (2010). *An improved algorithm for blind reverberation time estimation.* Paper presented at the Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC).

Löllmann, H. W., & Vary, P. (2008). *Estimation of the reverberation time in noisy environments.* Paper presented at the Proceedings of International Workshop on Acoustic Echo and Noise Control.

Lu, Y.-C., & Cooke, M. (2010). Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(7), 1793-1805.

Lu, Y. (2010). *Production and perceptual analysis of speech produced in noise.* University of Sheffield.

Lyon, R. H., DeJong, R. G., & Heckl, M. (1995). Theory and application of statistical energy analysis. *The Journal of the Acoustical Society of America, 98*(6), 3021-3021.

Mammone, R. J., Zhang, X., & Ramachandran, R. P. (1996). Robust speaker recognition: A feature-based approach. *IEEE signal processing magazine, 13*(5), 58.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). *The DET curve in assessment of detection task performance*. Retrieved from

May, T., van de Par, S., & Kohlrausch, A. (2012). Noise-robust speaker recognition combining missing data techniques and universal background modeling. *Audio, Speech, and Language Processing, IEEE Transactions on, 20*(1), 108-121.

McCowan, I. A., Pelecanos, J., & Sridharan, S. (2001). *Robust speaker recognition using microphone arrays.* Paper presented at the 2001: A Speaker Odyssey-The Speaker Recognition Workshop.

Memon, S. (2010). *Automatic Speaker Recognition: Modelling, Feature Extraction and Effects of Clinical Environment.* RMIT University.

Ming, J., Hazen, T. J., Glass, J. R., & Reynolds, D. A. (2007). Robust speaker recognition in noisy conditions. *Audio, Speech, and Language Processing, IEEE Transactions on, 15*(5), 1711-1723.

Mohn, W. (1970). Statistical feature evaluation in speaker identification.

Murty, K., & Yegnanarayana, B. (2006). Combining evidence from residual phase and MFCC features for speaker recognition. *Signal Processing Letters, IEEE, 13*(1), 52-55.

Nabelek, A. K., & Pickett, J. (1974). Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners. *Journal of Speech, Language, and Hearing Research, 17*(4), 724-739.

Naik, J. M. (1990). Speaker verification: A tutorial. *IEEE Communications Magazine, 28*(1), 42-48.

Nakasone, H. (2003). *Automated speaker recognition in real world conditions: controlling the uncontrollable.* Paper presented at the Eighth European Conference on Speech Communication and Technology.

Nakatani, T., & Miyoshi, M. (2003). *Blind dereverberation of single channel speech signal based on harmonic structure.* Paper presented at the Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on.

Naylor, P. A., & Gaubitch, N. D. (2010). *Speech dereverberation*: Springer Science & Business Media.

Ning, W., Ching, P. C., Nengheng, Z., & Tan, L. (2011). Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features. *Audio, Speech, and*

Language Processing, IEEE Transactions on, 19(1), 196-205. doi:10.1109/TASL.2010.2045800

NIST. (2001). Speaker Recognition Evaluation

Retrieved from http://www.itl.nist.gov/iad/mig/tests/spk

NIST. (2002). Speaker Recognition Evaluation. Retrieved from http://www.itl.nist.gov/iad/mig/tests/spk/2002/

NIST. (2004). Speaker Recognition Evaluation. Retrieved from http://www.itl.nist.gov/iad/mig/tests/spk/2004/

Noxon, A. M. (1992). *Correlation detection of early reflections.* Paper presented at the Audio Engineering Society Conference: 11th International Conference: Test & Measurement.

Oglesby, J. (1995). What's in a number? Moving beyond the equal error rate. *Speech communication, 17*(1), 193-208.

Omologo, M., Svaizer, P., & Matassoni, M. (1998). Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech communication, 25*(1), 75-95.

Oppenheim, A. V. (1969). Speech Analysis- Synthesis System Based on Homomorphic Filtering. *The Journal of the Acoustical Society of America, 45*(2), 458-465.

Peer, I., Rafaely, B., & Zigel, Y. (2008). *Reverberation matching for speaker recognition.* Paper presented at the Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on.

Petrick, R., Lohde, K., Wolff, M., & Hoffmann, R. (2007). The harming part of room acoustics in automatic speech recognition.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE, 81*(9), 1215-1247.

Pillay, S. G., Ariyaeeinia, A., Sivakumaran, P., & Pawlewski, M. (2009). *Open-Set Speaker Identification under Mismatch Conditions.* Paper presented at the Tenth Annual Conference of the International Speech Communication Association.

Poonkuzhali, C., Karthiprakash, R., Valarmathy, S., & Kalamani, M. (2013). An approach to feature selection algorithm based on ant colony optimization for automatic speech recognition. *International journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 11 (2).*

Pulkki, V. (2007). Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society, 55*(6), 503-516.

Pullella, D., Kuhne, M., & Togneri, R. (2008). *Robust speaker identification using combined feature selection and missing data recognition.* Paper presented at the Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on.

Qi, J., Wang, D., Xu, J., & Tejedor, J. (2013). *Bottleneck Features based on Gammatone Frequency Cepstral Coefficients.* Paper presented at the Interspeech'13.

Quatieri, T. F. (2002). *Discrete-time speech signal processing: principles and practice*: Pearson Education India.

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of speech recognition* (Vol. 14): PTR Prentice Hall Englewood Cliffs.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals*: Prentice Hall.

Rajan, P., Kinnunen, T., & Hautamäki, V. (2013). *Effect of multicondition training on i-vector PLDA configurations for speaker recognition.* Paper presented at the Interspeech.

Ratnam, R., Jones, D. L., Wheeler, B. C., O’Brien Jr, W. D., Lansing, C. R., & Feng, A. S. (2003). Blind estimation of reverberation time. *The Journal of the Acoustical Society of America, 114*(5), 2877-2892.

Reynolds, D. (2002). *An overview of automatic speaker recognition.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)(S. 4072-4075).

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., . . . Mihaescu, R. (2003). *The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition.* Paper presented at the Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on.

Reynolds, D., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on, 3*(1), 72-83.

Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification. *Speech and Audio Processing, IEEE Transactions on, 2*(4), 639-643.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech communication, 17*(1), 91-108.

Reynolds, D. A., & Heck, L. P. (2000). *Automatic speaker recognition.* Paper presented at the AAAS 2000 Meeting, Humans, Computers and Speech Symposium.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing, 10*(1), 19-41.

Ristić, D. M., Pavlović, M., Pavlović, D. Š., & Reljin, I. (2013). Detection of early reflections using multifractals. *The Journal of the Acoustical Society of America, 133*(4), EL235-EL241.

Rose, P. (2003). *Forensic speaker identification*: CRC Press.

Sabine, W. C. (1922). *Collected papers on Acoustics, prepared by T. J. Lyman, (Reprinted by Dover, New York, 1964).*

Sadjadi, S. O., & Hansen, J. H. (2012). *Blind reverberation mitigation for robust speaker identification.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.

Sadjadi, S. O., & Hansen, J. H. (2014). Blind spectral weighting for robust speaker identification under reverberation mismatch. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 22*(5), 937-945.

Sadjadi, S. O., Slaney, M., & Heck, L. (2013). MSR Identity Toolbox v1. 0: A MATLAB toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter*.

Santon, F. (1976). Numerical prediction of echograms and of the intelligibility of speech in rooms. *The Journal of the Acoustical Society of America, 59*(6), 1399-1405.

Schroeder, M. R. (1965). New method of measuring reverberation time. *The Journal of the Acoustical Society of America, 37*(3), 409-412.

Sehr, A., & Kellermann, W. (2010). *On the statistical properties of reverberant speech feature vector sequences.* Paper presented at the IWAENC.

Sethuraman, R., & Gowdy, J. (1989). *A cepstral based speaker recognition system.* Paper presented at the System Theory, 1989. Proceedings., Twenty-First Southeastern Symposium on.

Shao, Y., Srinivasan, S., & Wang, D. (2007). *Incorporating auditory feature uncertainties in robust speaker identification.* Paper presented at the Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on.

Shao, Y., & Wang, D. (2006). Model-based sequential organization in cochannel speech. *Audio, Speech, and Language Processing, IEEE Transactions on, 14*(1), 289-298.

Shao, Y., & Wang, D. (2008). *Robust speaker identification using auditory features and computational auditory scene analysis.* Paper presented at the Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on.

Shriberg, E. (2007). Higher-level features in speaker recognition *Speaker Classification I* (pp. 241-259): Springer.

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., & Stolcke, A. (2005). Modeling prosodic feature sequences for speaker recognition. *Speech communication, 46*(3), 455-472.

Sivakumaran, P., Fortuna, J., & Ariyaeeinia, A. M. (2003). *Score normalisation applied to open-set, text-independent speaker identification.* Paper presented at the Eighth European Conference on Speech Communication and Technology.

Sokolov, M. (1997). *Speaker verification in the World Wide Web*. Paper presented at the Eurospeech.

Soong, F., Rosenberg, A., Rabiner, L., & Juang, B. H. (1985, Apr 1985). *A vector quantization approach to speaker recognition.* Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.

Standard, I. (1997). Acoustics–Measurement of the reverberation time of rooms with reference to other acoustical parameters. *International Standards Organization*.

Stevens, K. N. (1971). *Sources of inter-and intra-speaker variability in the acoustic properties of speech sounds.* Paper presented at the Proceedings of the 7th International Congress of Phonetic Sciences.

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30): MIT press.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America, 8*(3), 185-190.

Suits, B. H. (2015, 2015). Autocorrelation (for sound signals). Retrieved from http://pages.mtu.edu/~suits/autocorrelation.html

Vesa, S. (2007). *Sound source distance learning based on binaural signals.* Paper presented at the Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on.

Vesa, S. (2009). Binaural sound source distance learning in rooms. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(8), 1498-1507.

Wang, C., Xu, D., & Principe, J. C. (1997). *Speaker verification and identification using gamma neural networks.* Paper presented at the Neural Networks, 1997., International Conference on.

Wang, D., & Brown, G. J. (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*: Wiley-IEEE Press.

Wang, L., & Nakagawa, S. (2009). Speaker identification/verification for reverberant speech using phase information. *Proc. of WESPAC 2009*(0130).

Wang, N., Ching, P., Zheng, N., & Lee, T. (2011). Robust speaker recognition using denoised vocal source and vocal tract features. *Audio, Speech, and Language Processing, IEEE Transactions on, 19*(1), 196-205.

Webb, A. R., & Copsey, K. D. (2011). Linear Discriminant Analysis. *Statistical Pattern Recognition, Third Edition,* 221-273.

Wen, J. Y., Habets, E. A., & Naylor, P. A. (2008). *Blind estimation of reverberation time based on the distribution of signal decay rates.* Paper presented at the Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on.

Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America, 51*(6B), 2044-2056.

Wu, M., & Wang, D. (2006). A two-stage algorithm for one-microphone reverberant speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on, 14*(3), 774-784.

Yuk, C.-S. (1996). *An HMM approach to text independent speaker verification.* Paper presented at the IEEE international conference on Acoustics, Speech and signal processing.

Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica, 91*(3), 409-420.

Zhang, Z., Wang, L., & Kai, A. (2014). Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation. *EURASIP Journal on Audio, Speech, and Music Processing, 2014*(1), 15.

Zhao, X., Shao, Y., & Wang, D. (2012). CASA-based robust speaker identification. *Audio, Speech, and Language Processing, IEEE Transactions on, 20*(5), 1608-1616.

Zhao, X., & Wang, D. (2013). *Analyzing noise robustness of MFCC and GFCC features in speaker identification.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.

Zhao, X., Wang, Y., & Wang, D. (2014). Robust Speaker Identification in Noisy and Reverberant Conditions.

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Technology, 16*(6), 582-589.

Zilovic, M. S., Ramachandran, R. P., & Mammone, R. J. (1995). *The use of robust cepstral features obtained from pole-zero transfer functions for speaker identification.* Paper presented at the Electrical and Computer Engineering, 1995. Canadian Conference on.

APPENDICES

Appendix A: Supporting Figures

This appendix provides Figures, which support explanations of results, which are provided in chapter 4. The figures here represented Detection trade-off curve (DET) for a different reverberation time, which are obtained by the MFCC and GFCC features

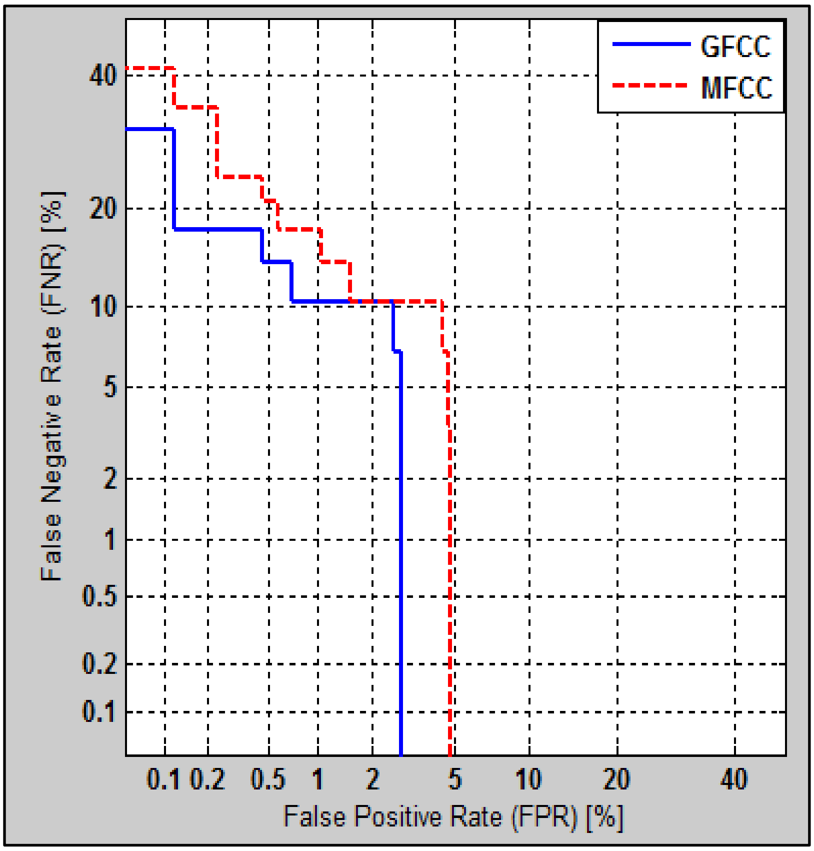


Figure A1 DET plot of both features with RT=0.53s

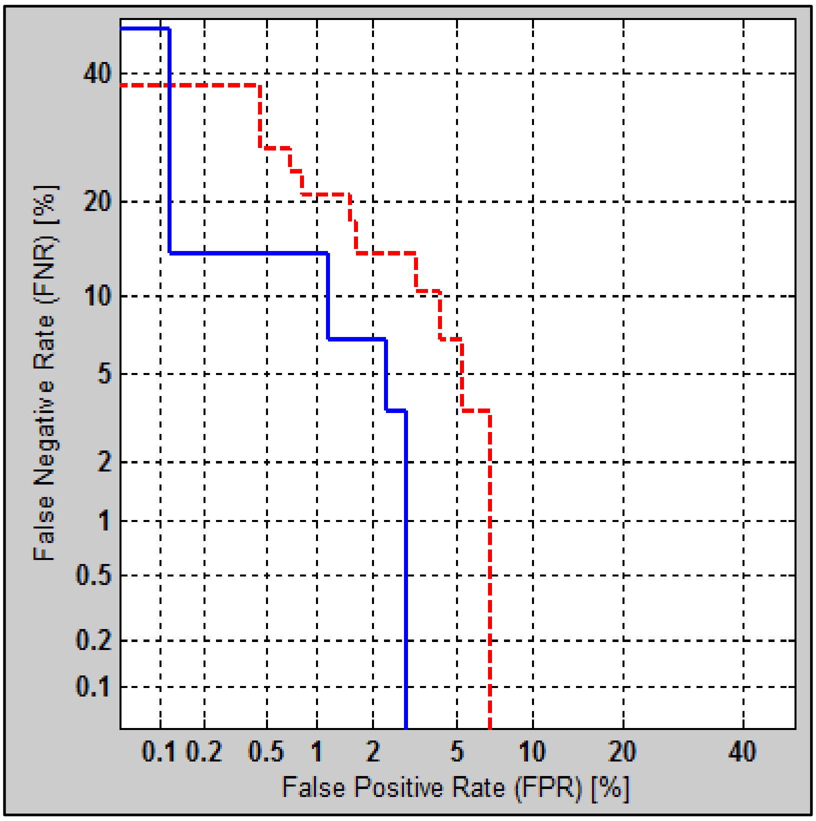


Figure A2 DET plot of both features with RT=0.84s

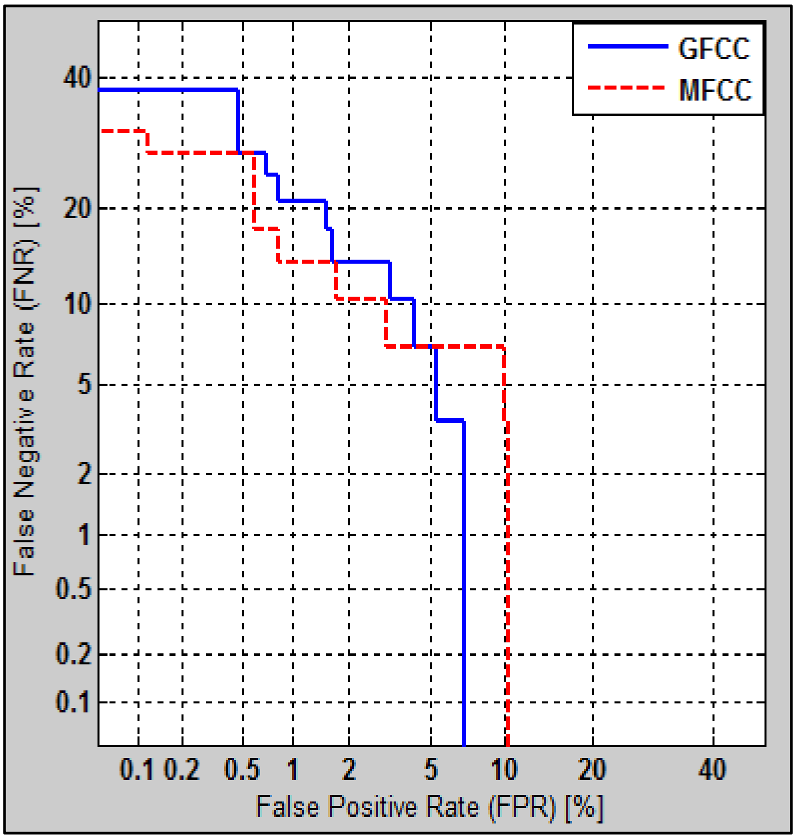


Figure A3 DET plot of both features with RT =1.0s

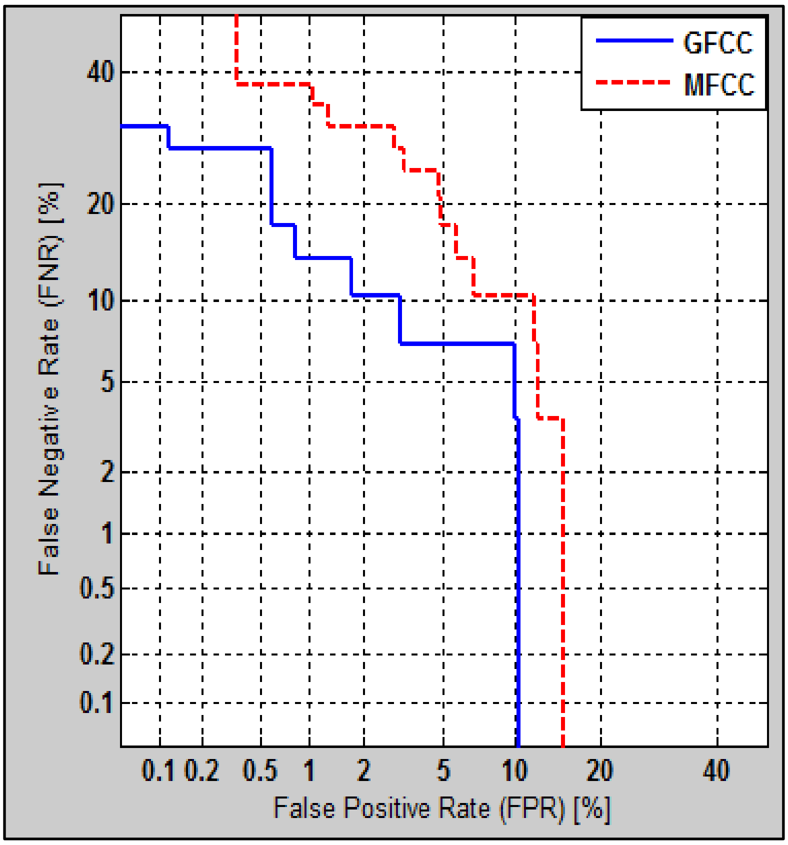


Figure A4 DET plot of both features with RT=1.2s

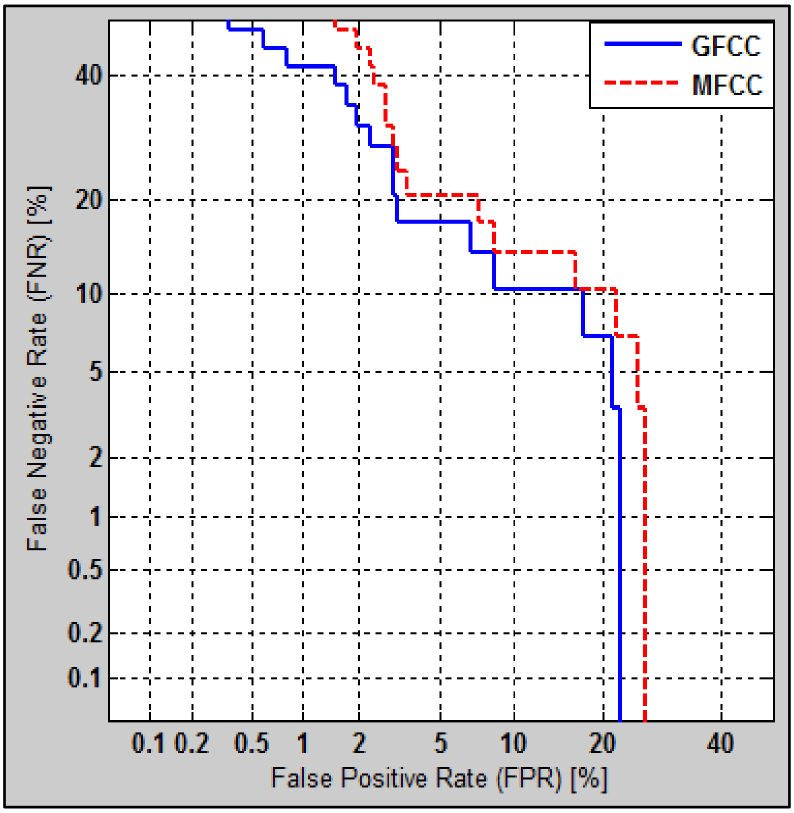


Figure A5 DET plot of both features with RT=1.4s

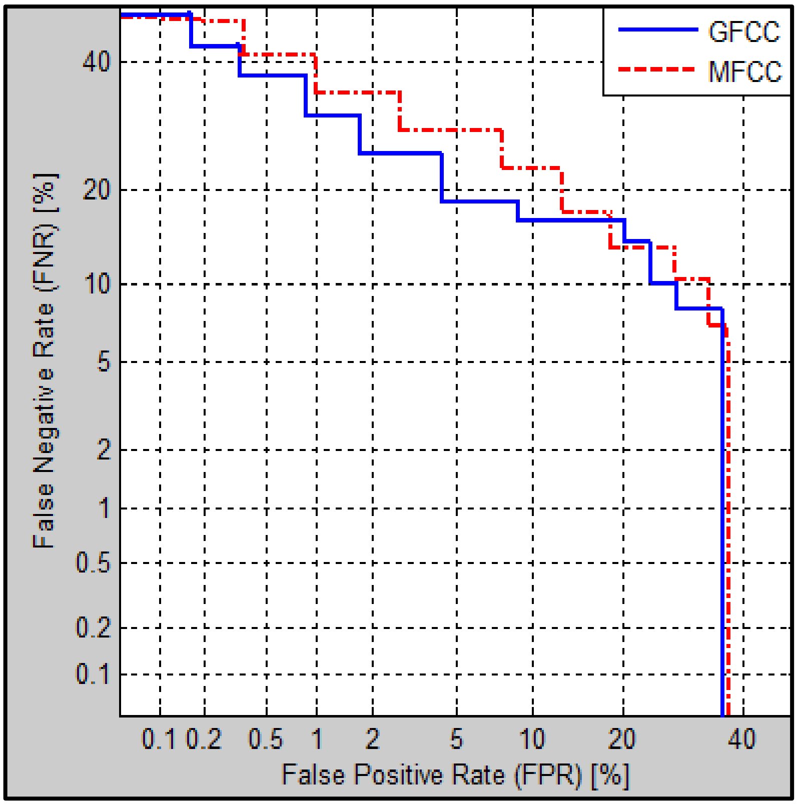


Figure A6 DET plot of both features with RT=1.8s

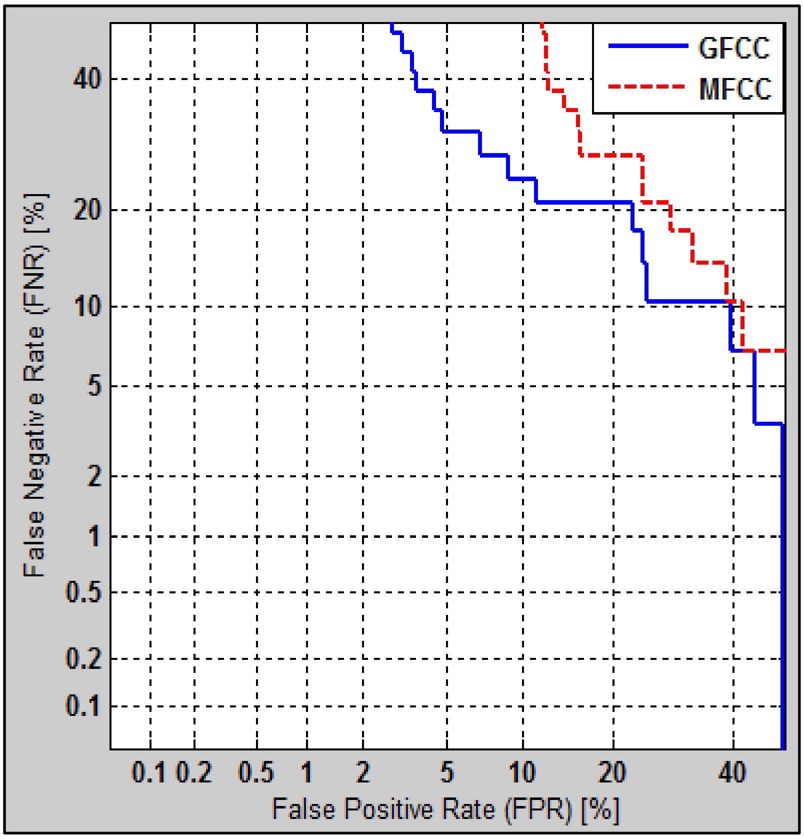


Figure A7 DET plot of both features with RT=2s

Appendix B: Supporting Figures

This appendix provides Figures, which support explanations of results, which are provided in chapter 6. The figures here represented Detection trade-off curve (DET) for a different reverberation time, which are obtained by applying different scenarios for training the system with reverberant speech.

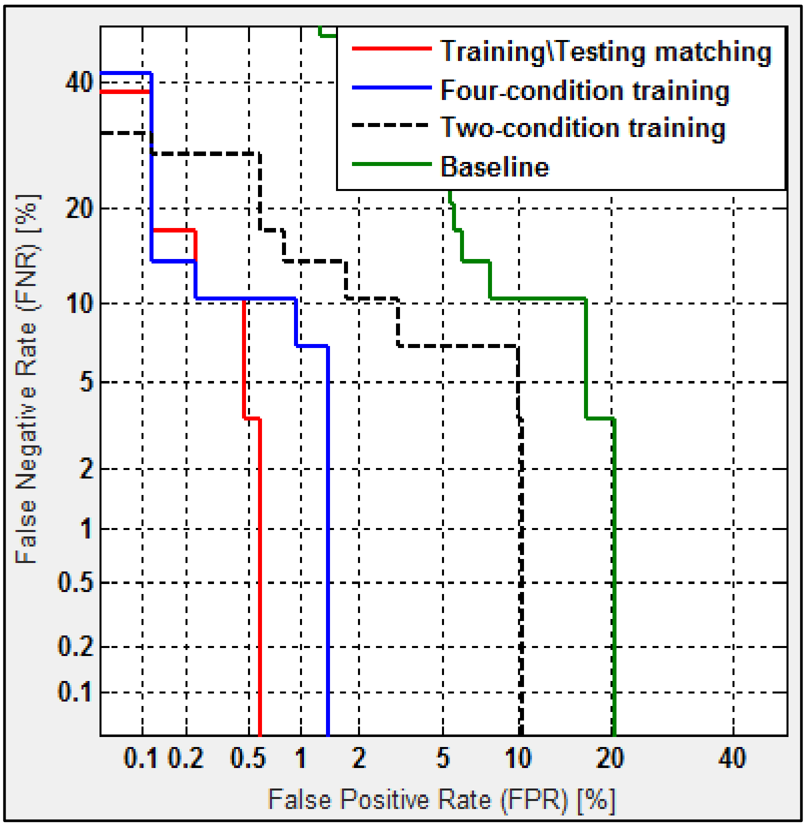


Figure B1 DET Curve for multi training scenarios with RT= 1.2s in testing phase

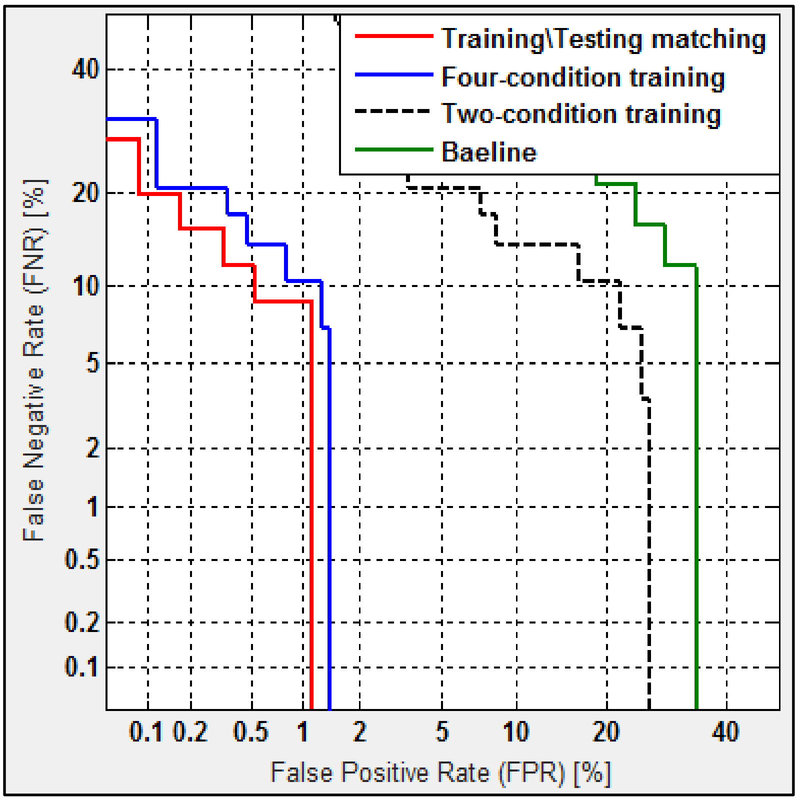


Figure B2 DET Curve for multi training scenarios with RT= 1.2s in testing phase

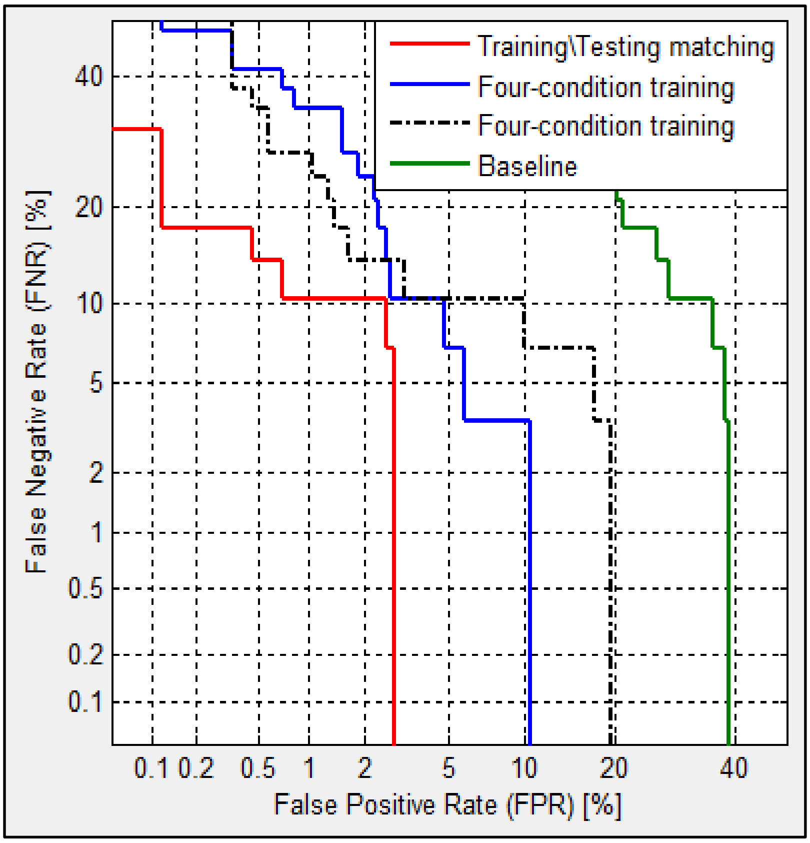


Figure B3 DET Curve for multi training scenarios with RT= 2s in testing phase

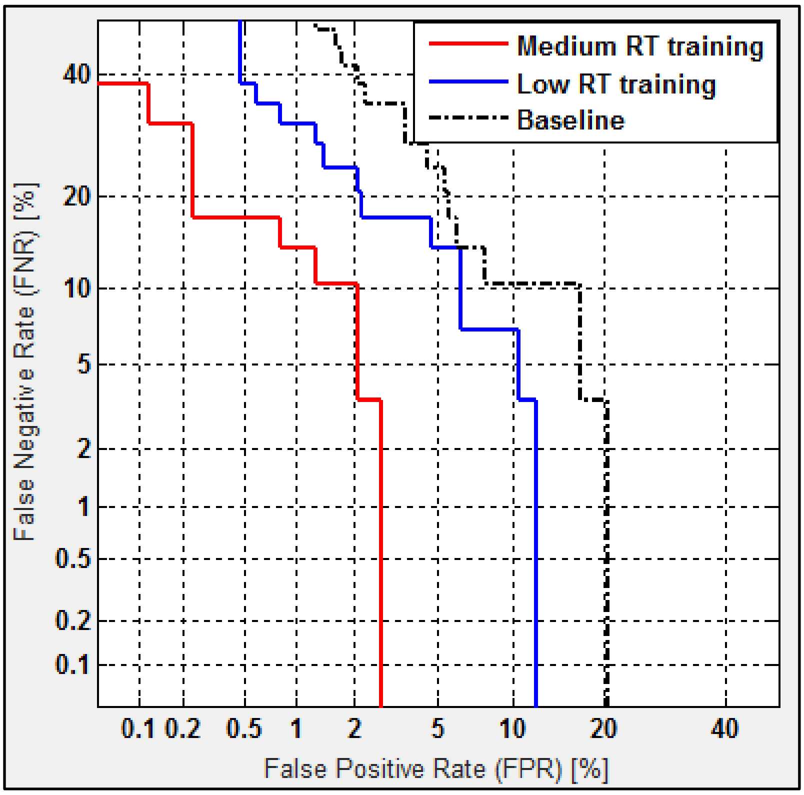


Figure B4 DET Curve plot with RT= 1.2s in testing phase

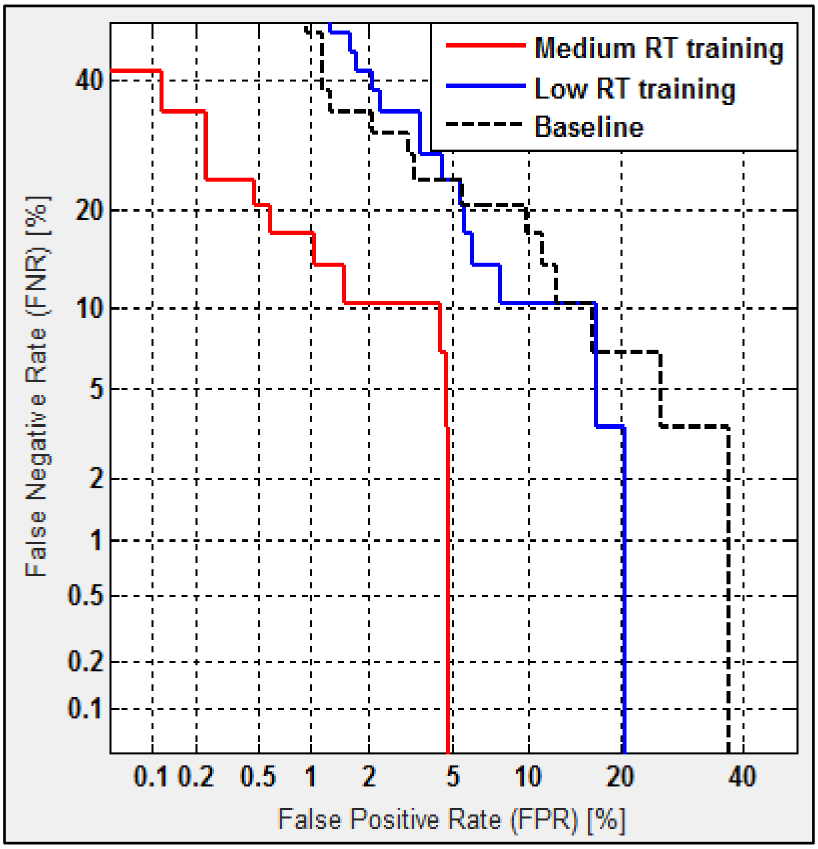


Figure B5 DET Curve plot with RT= 1.5s in testing phase

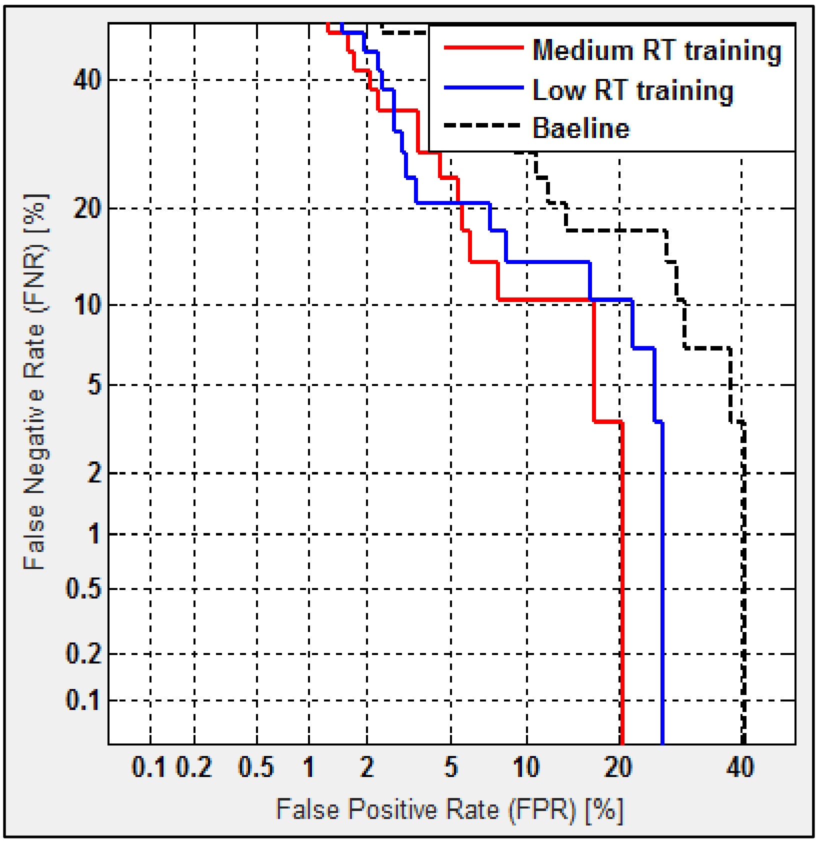


Figure B6 DET Curve plot with RT= 2s in testing phase

Appendix C: Supporting Figures

This appendix provides Figures, which support explanations of results, which are provided in chapter 8. The figures here represented Detection trade-off curve (DET) for different reverberation time.

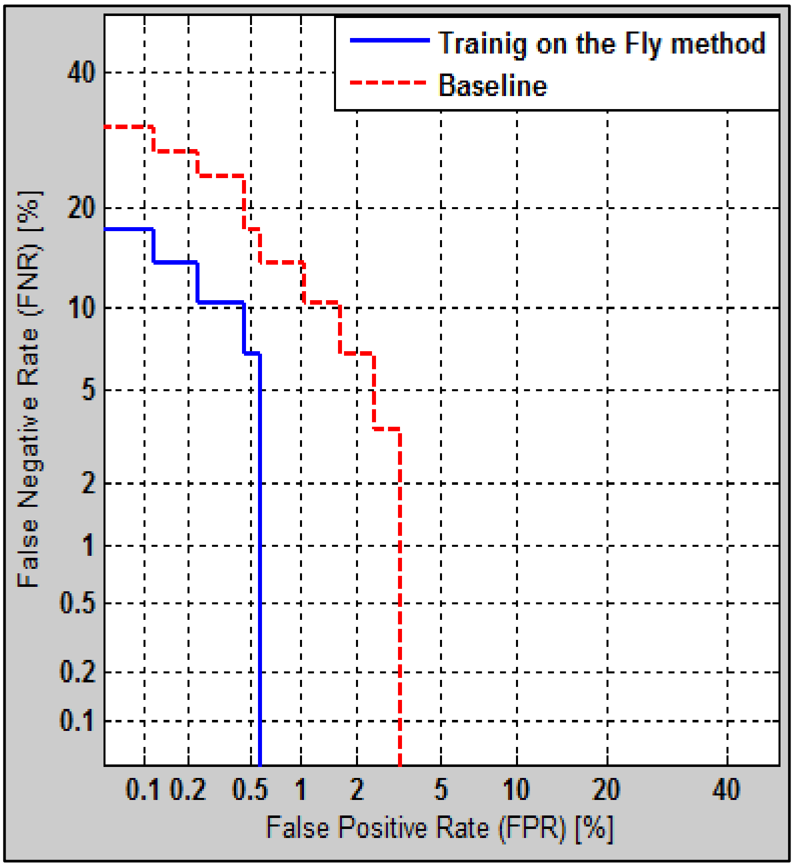


Figure C1 DET curve for reverberation time 0.71s

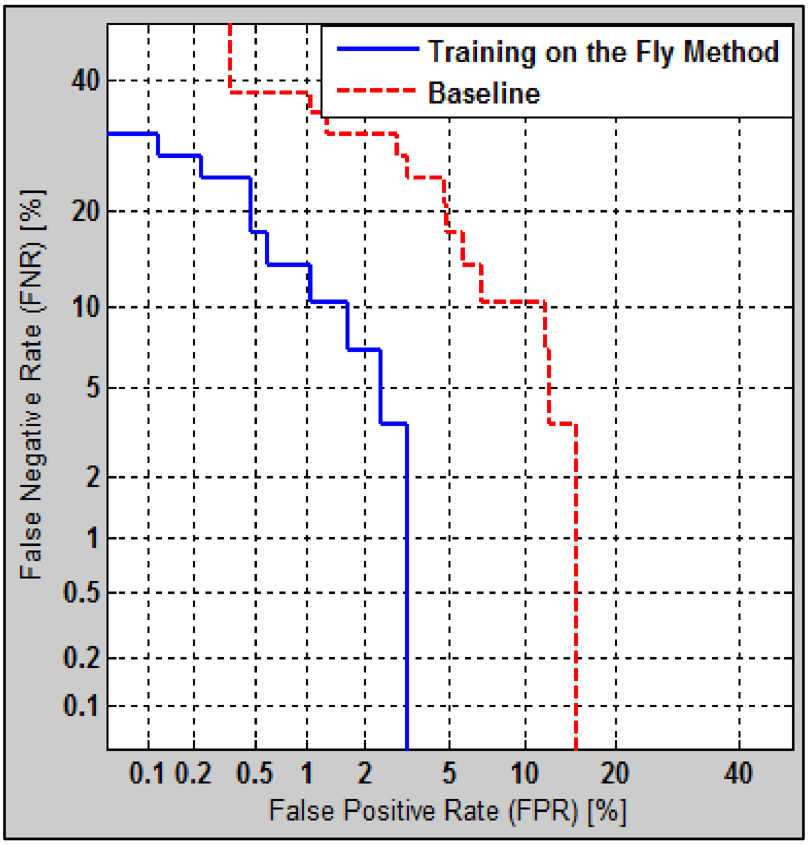


Figure C2 DET curve for reverberation time 1.2s

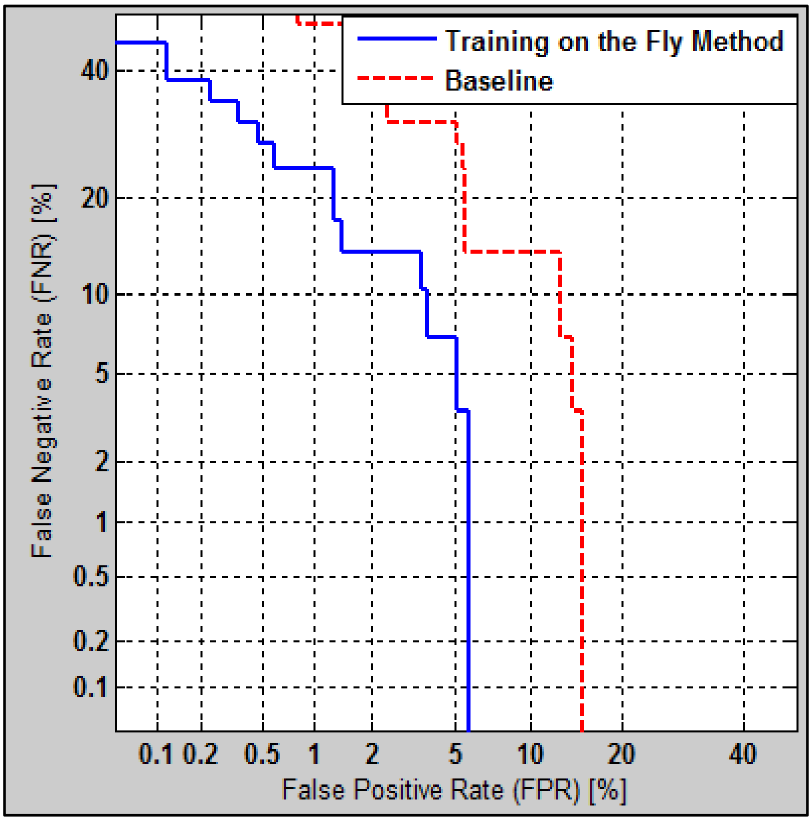


Figure C3 DET curve for reverberation time 1.4s

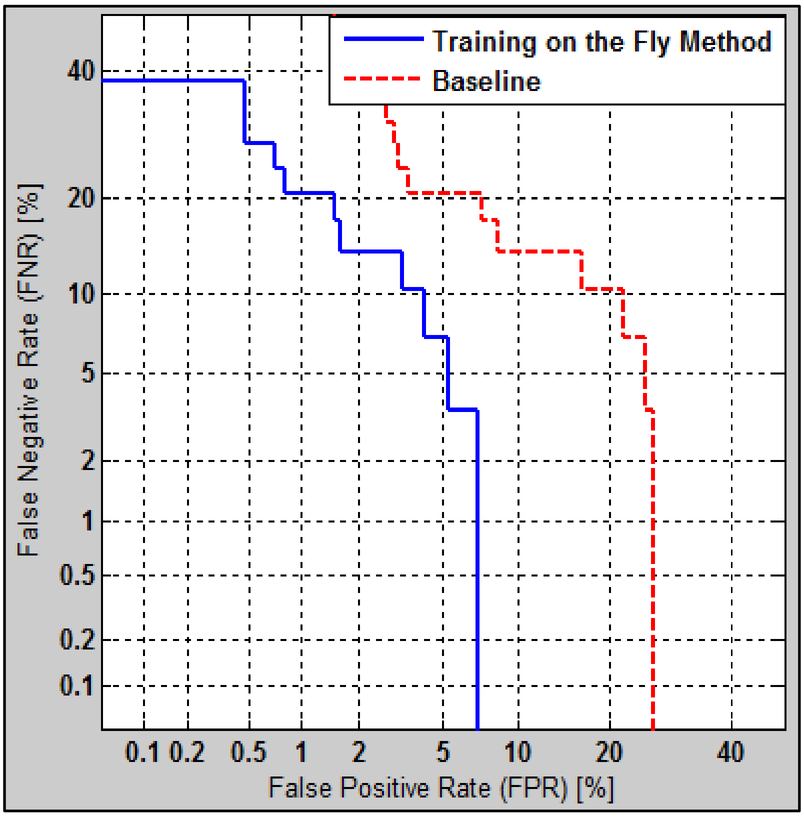


Figure C4 DET curve for reverberation time 1.5s

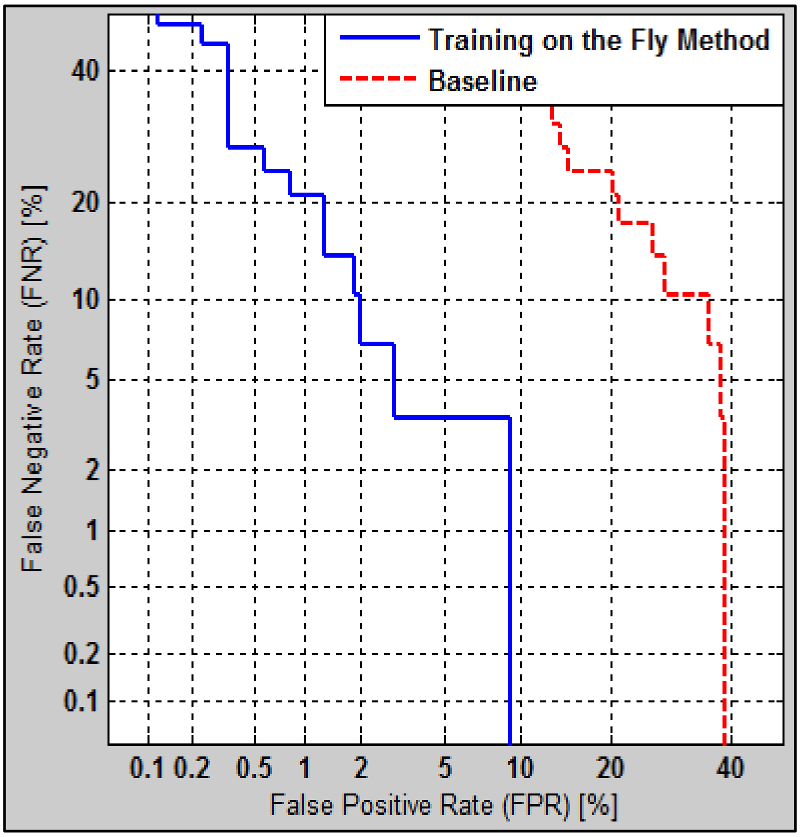


Figure C5 DET curve for reverberation time 1.8s

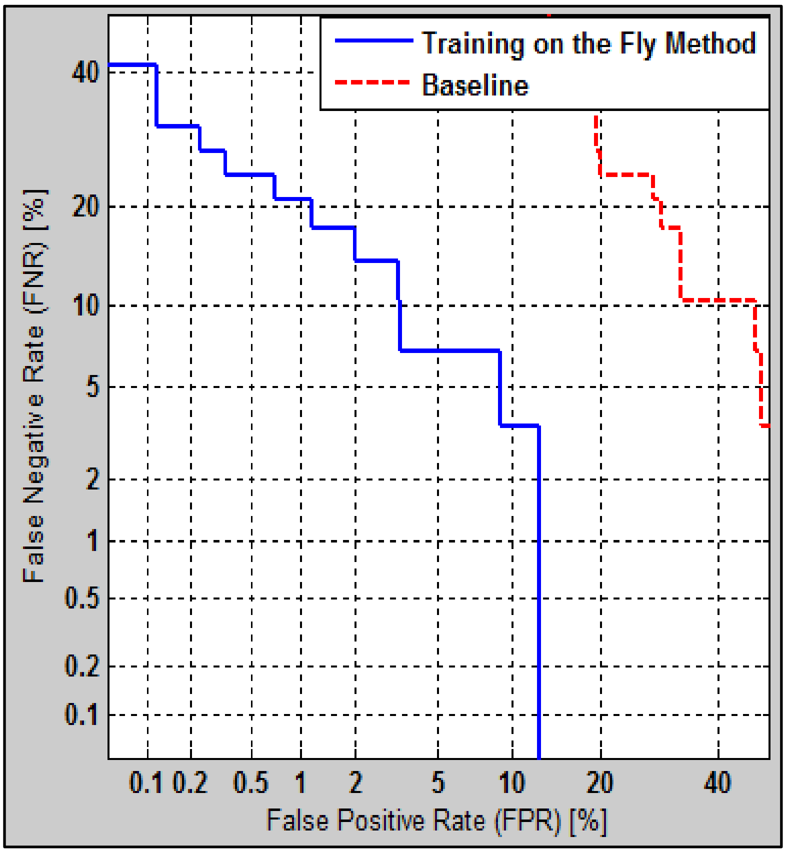


Figure C6 DET curve for reverberation time 2s

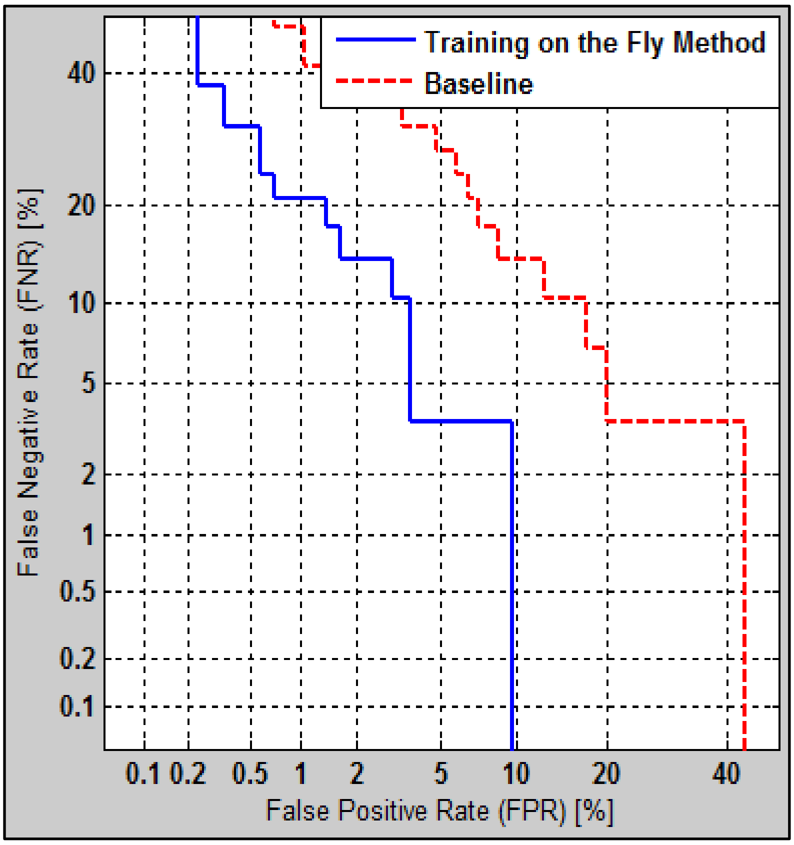


Figure C7 DET curve for reverberation time 2.5s