University of Salford-Manchester
School of Computing, Science &
Engineering

ROBUST SPEAKER RECOGNITION IN PRESENCE OF NON-TRIVIAL ENVIRONMENTAL NOISE

TOWARD GREATER BIOMETRIC SECURITY

By

Ahmed Hani Yousif Al-Noori

Supervised By
Dr. Phil Duncan
Dr. Francis Li

A thesis Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

June 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

All Praise is due to Allah Lord of the Universe, the most merciful.

Firstly I would like to express my gratitude to the Government of the Republic of Iraq including the Ministry of Higher Education and Scientific Research and The Iraqi Cultural Attaché-London for their help and support throughout my studies in the United Kingdom.

Immeasurable gratitude goes toward my supervisors, **Dr. Phil Duncan** and **Dr. Francis Li,** for their continuous support, unforgettable patience and motivation, which made this work possible. It was an honour to have them as my supervisors.

A special thanks must also go to all my colleagues at the acoustics research centre. The last four years as a PhD student would certainly have been more difficult without you all. But special thanks should go to James Massaglia and Will Bailey for their advice and guidance. I will never forget their help. It is an honour to know you as close friends.

In addition, I would like to acknowledge all my other friends, the Iraqis and the other nationalities at Newton building, at the University of Salford, and all my friends in other UK universities.

Finally, I really should thank my dearest Mum, my beloved wife Saba, my three sisters Haifa', Dina, and Siba, and my children Ibrahim and Shams for all the encouragement and supports through the time of this research. I dedicate this work to my late father, who has always wished for me to pursue a PhD, and I hope it makes him proud. Without their love, encouragement, and prayers I would certainly not be in the position I am today, and for that I am eternally grateful.

AUTHOR PUBLICATIONS

- Al-Karawi, K.A., Al-Noori, A.H., Li, F.F. and 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), p.423.
- Al-Noori, A.H, Li, F.F., 2015 Speaker Recognition of non-Trivial Noise. Salford Postgraduate Annual Research Conference (SPARC), University of Salford.
- Al-Noori, A.H., Al-Karawi, K.A. and Li, F.F., 2015, September. Improving robustness of speaker recognition in noisy and reverberant conditions via training. In Intelligence and Security Informatics Conference (EISIC), 2015 European (pp. 180-180). IEEE.
- Al-Noori, A.H, Li, F.F. and Duncan P.J., 2016. A comparative Study of Speech Enhancement Approaches for Speaker Recognition in Realistic Noisy Conditions. The CSE 2016 Annual PGR Symposium (CSE-PGR16).
- Al-Noori, A., Li, F.F. and Duncan, P.J., 2016, May. Robustness of speaker recognition from noisy speech samples and mismatched languages. In Audio Engineering Society Convention 140. Audio Engineering Society.
- Al-Noori, A., Li, F.F. and Duncan, P.J., 2017, Training 'on the fly' to improve the performance of Speaker Recognition in Noisy Environments. 2017 AES International Conference on Audio Forensics-Finding Signal in the Noise. Audio Engineering Society.

ABSTRACT

The aim of this thesis is to investigate speaker recognition in the presence of environmental noise, and to develop a robust speaker recognition method. Recently, Speaker Recognition has been the object of considerable research due to its wide use in various areas. Despite major developments in this field, there are still many limitations and challenges. Environmental noises and their variations are high up in the list of challenges since it impossible to provide a noise free environment.

A novel approach is proposed to address the issue of performance degradation in environmental noise. This approach is based on the estimation of signal-to-noise ratio (SNR) and detection of ambient noise from the recognition signal to re-train the reference model for the claimed speaker and to generate a new adapted noisy model to decrease the noise mismatch with recognition utterances. This approach is termed "**Training on the fly**" for robustness of speaker recognition under noisy environments.

To detect the noise in the recognition signal two different techniques are proposed: the first technique including generating an emulated noise depending on estimated power spectrum of the original noise using 1/3 octave band filter bank and white noise signal. This emulated noise become close enough to original one that includes in the input signal (recognition signal). The second technique deals with extracting the noise from the input signal using one of speech enhancement algorithm with spectral subtraction to find the noise in the signal.

Training on the fly approach (using both techniques) has been examined using two feature approaches and two different kinds of artificial clean and noisy speech databases collected in different environments. Furthermore, the speech samples were text independent. The training on the fly approach is a significant improvement in performance when compared with the performance of conventional speaker recognition (based on clean reference models). Moreover, the training on the fly based on noise extraction showed the best results for all types of noisy data.

LIST OF ABBREVIATIONS

AMFT	Advanced Missing Features Theory
ANN	Artificial Neural Network
BPC	Broad Phonetic Category
BPNN	Back Propagation Neural Network
BSL	Background Scaling Linear
CASA	Computational Auditory Scene
CFCC	Cochlear Filter Cepstral Coefficients
DCF	Dectection Cost Function
DCT	Discrete Cosine Transform
DET	Detection Erorr Trade-Off
DWT	Discrete Wavelet Transform
EER	Error Equal Rate
EFCC	Exponential Frequency Cepstral Coefficients
EM	Expectation Maximization
ERB	Equivalent Rectangular Bandwidth
FAR	False Acceptence Rate
FNR	False Negative Rate
FPR	False Positive Rate
FRR	False Rejection Rate
FVQ	Fuzzy Vector Quantization
GF	Gammatone Features
GFCC	Gammatone Frequency Cepstral Coefficients
GLDS	Generalized Linear Discriminant Sequence
GMM	Gaussian Mixture Model
GMM-UBM	Gaussian Mixture Model-Universal Background Model
GSV	Gaussian Supervector Support Vector Machine
HLDA	Hetero Linear Discriminant Analysis
HMFT-8	Hard Mask Missing Feature Theory
HMM	Hidden Markov Model
IER	Idetifcation Error Rate
JFA	Joint Factor Analysis
KL	Kullback-Leiber
KLD	Kullback-Leiber Distance
LBG	Lindw-Buzo-Gray
LDA	Linear Discriminant Analysis
LDC	Linguistic Data Consortium
LFCC	Linear Frequency Cepstral Coefficients
LLR	Log Likelihood Ratio
LogMMSE	Log Minimum Mean Square Error
LP	Linear Predictive
LPC	Linear Predictive Coding
LPCC	Linear Predictive Cepstral Coefficients
MAP	Maximum Posteriori
MFCC	Mel Frequency Cepstral Coefficients
MFDWC	Mel-Frequency Discrete Wavelet Coefficients
MFT	Missing Feature Theory
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Registration
MLP	Multilayer Perception
MMSE	Minimum Mean Square Error
MSR	Microsoft Speaker Recognition
MT-Mask	Psychoacoustical Motivated algorithm

OSI-EER	Open Set Error Equal Rate
PCA	Principal Component Analysis
PDFs	Probability Density Functions
PLP	Perceptual Linear Prediction
PLPC	Perceptual Linear Prediction Coefficients
PMC	Prallel Model Combination
ROC	Reciver Operating Chracteristics
SALU-AC	Salford University – Anechoic Chamber
SALU-CR	Salford University –Clean Room
SI	Speaker Identification
SNERF-gram	Syllable based Nonuniform Extraction Region Features
SNR	Signal/Speech to Noise Ration
SOM	Self Organization Map
SPIDRE	Speech Technology database
SR	Speaker Recogntion
SSA	Singular Spectral Analysis
STRF	Spectral-Temporal Receptive Fields
SV	Speaker Verification
SVM	Support Vector Machine
T-F	Time-Frequency
TIMIT	Texas Instruments Massachusetts Institute of Technology
T-Norm	Test Normalization
TOTF	Training on The Fly
UBM	Universal Background Model
VAD	Voice Activity Detection
VQ	Vector Quantization
WCCN	Within-Class Covariance Normalization
WOCOR	Wavelet Octave Coefficients Of Residues
Z-Norm	Zero Normalization

CHAPTER 1
INTRODUCTION

1.1 Speaker Recognition and other Biometrics

Speech signals include much information related to the speaker. This information has been used throughout recent decades by researchers in many fields, such as speech recognition, speech synthesis, and speech production. However, one of the most important usages of the speech signal, which is related to the security and forensic fields, is the biometric authentication called speaker recognition.

Speaker Recognition (or voice Biometric) is part of Biometric authentication, which means the automatic identity recognition of a person using specific intrinsic characteristics of the person concerned, such as fingerprint, hand, Iris, or DNA. To select a specific biometric, many factors should be taken into account: robustness, accessibility, acceptability, and flexibility.

Speaker Recognition is defined as the process of recognising a human depending on the unique characteristics that are included in his/her speech signal. This can be represented as a hybrid biometric because the speaker's voice is identified based on the structure of the vocal tract (physiological component) in addition to the way that the speaker talks (behaviour component)(Pillay, 2010).

The main advantages of Speaker Recognition are that it is easy to capture the samples from individuals compared with other kinds of biometric patterns, which usually require the presence of the person (Nengheng, 2005). For instance, in the application of telephone banking, the user needs to be remotely identified before verification to their bank account. Furthermore, systems based on speaker recognition are probably the most natural and economical among other biometric systems, since they do not need any special hardware to capture the human voice. For example, in telephone-based applications to recognise users only a telephone handset of a mobile phone or landline are required. These reasons make speaker recognition one of the most desirable types of biometric authentication (Pillay, 2010).

Speaker recognition is also widely used in various areas, such as access control for computer data, forensic voice sample, mobile communication, speaker classification and others. Table 1-1 illustrates a comparison of different types of biometrics.

Table 1-1: Comparison among different types of Biometric Authentications

Biometric	Flexible to use (Acceptability)	Accuracy	Robustness (long term stability	Error incidence	Required Security level
Fingerprints	Medium	High	High	Dryness, Dirt, age	High
Face	High	High	Medium	Age, Glasses, lighting, hair	Medium
Iris	Medium	High	High	Low lighting	High
Signature	High	High	Medium	Changing Signatures	Medium
Hand Geometry	Medium	High	Medium	Hand injury, age	Medium
Voice	High	High	Medium	Cold, Noise	Medium

There is a significant difference between speaker recognition and speech recognition, although they are related since both utilise speech signals. Speaker recognition tries to find **who** was speaking (identify or verify of the speaker) whereas the primary aim of speech recognition is to determine **what** was spoken.

1.2 Speaker Recognition Branches

Speaker recognition systems (SR) have various types which are either directly or indirectly related (Beigi, 2011). But in general, speaker recognition can be classified into two main branches:

Speaker Verification **SV** (or Speaker Authentication) and Speaker Identification **SI**. Speaker verification (SV) is defined as the process of determining whether this speaker’s voice belongs to the claimed person or not based on a given input utterance (e.g. username, or an identification number). Therefore, there are only two possible types of output for a verification system: either this voice sound is the sound of the claimed person (user verified) or it belongs to an imposter (user rejected) (Beigi, 2011, Rao and Sarkar, 2014).

The process of speaker verification includes taking the speech signal from a speaker to create a recognition model (test model) by extracting parametric speech features from this signal; these features represent the more stable, robust and compact coefficients of the test speech signal which are suitable for the next phases. This recognition model is then compared with another model that has already been stored in a database called the enrolled speaker model (training reference model) which is extracted in the same way in the recognition phase to obtain the match score between the two models. This score specifies the degree of similarity between the test speaker model and training speaker model (Figure 1.1).

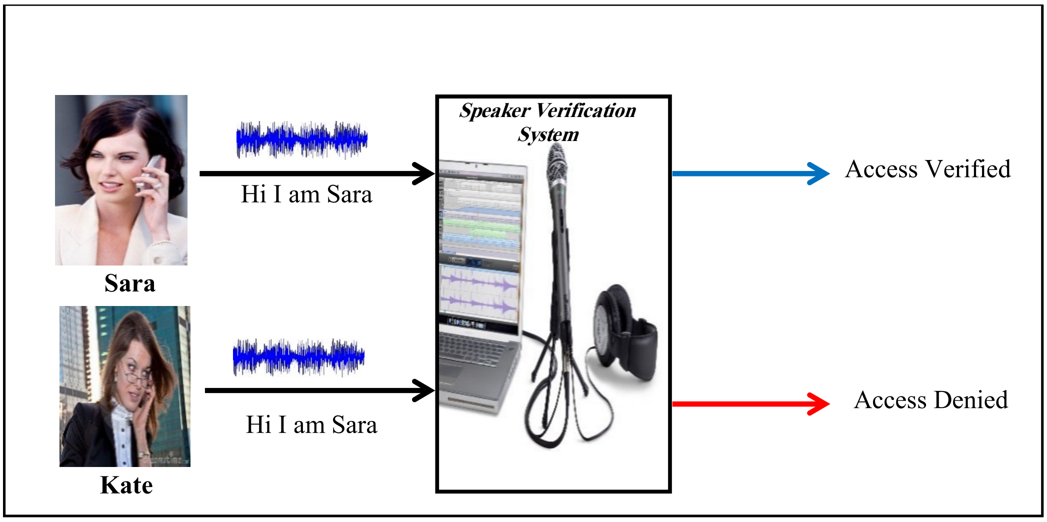


Figure 1.1: Speaker Verification System

Speaker identification, on the other hand, is known as the process of using a speech signal to determine the identity of the speaker from a population of registered speakers (Figure 1.2). In other words, the input speech signal of the unknown speaker is compared with a database of **M** reference speaker's models as a **1: M** process to find the closest matching speaker.

Speaker Identification can be divided into two types: Closed Set and Open set. In the closed set system, the speech of a target user is matched with all available registered speakers' models and the identity of the closest model (or sometimes models) is returned. Therefore, there is no rejection case in this type. Consequently, the number of alternative decisions is similar to population Size M (Figure 1.2.a). Thus, in closed set speaker identification, when the number of individual models in the system increases, the performance of the identification system decreases (Pillay, 2010, Beigi, 2011, Rao and Sarkar, 2014).

The open set model represents a combination of speaker verification and closed set speaker identification. In this kind of system, if the speech signal of the speaker is matched with one of the registered speaker models, the identity of that speaker will be returned. However, if the speech signal is not matched with any model in the system, the speaker may be rejected and registered as an unknown speaker (imposter). Therefore, the number of alternative decisions, in this case, is M+1, with the addition of the possibility that the speech belongs to an anonymous user (Figure 1.2.b).

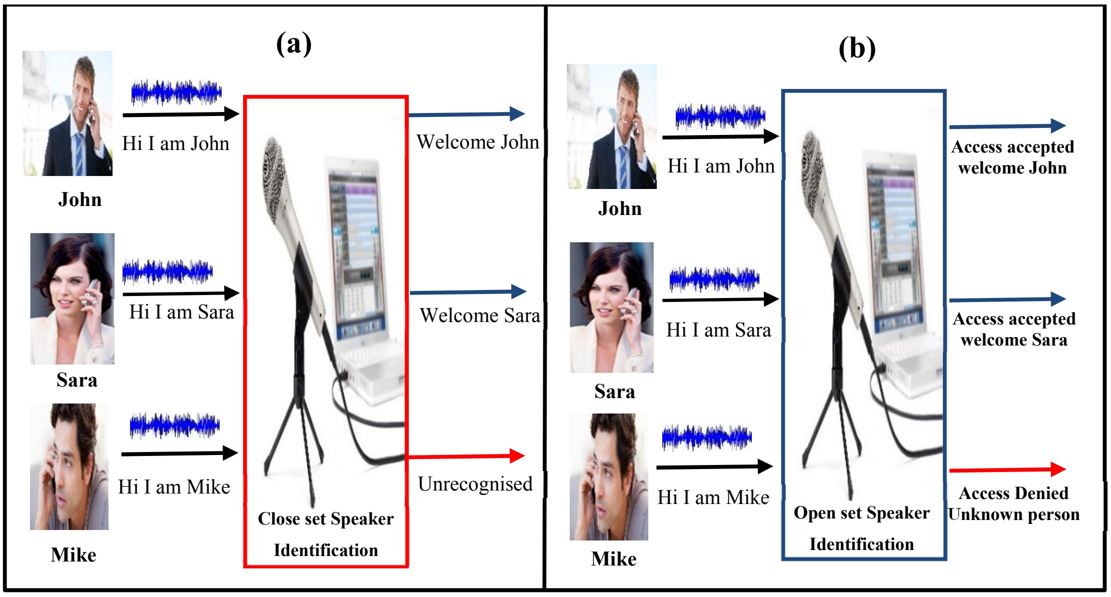

(a)
John
Hi I am John
Sara
Hi I am Sara
Mike
Hi I am Mike
Close set Speaker Identification
Welcome John
Welcome Sara
Unrecognised
(b)
John
Hi I am John
Sara
Hi I am Sara
Mike
Hi I am Mike
Open set Speaker Identification
Access accepted welcome John
Access accepted welcome Sara
Access Denied Unknown person

Figure 1.2: Closed-Open Set Speaker Identification Systems

In Addition to the previous classification, Speaker recognition can further be classified according to the speech modalities into text-dependent, text prompt, and text independent (Beigi, 2011). In the first type, which is usually used in speaker verification, the utterance used for accessing the system is limited to a specific phrase or phrases. So, the recognition phrases are constant and known in advance (Rao and Sarkar, 2014). One of the major problems of this type of recognition is what is called ***the liveness challenge*** (Beigi, 2011) which means that the imposter can easily record the voice of a target speaker and use it to access the recognizer system. So, the usage of this kind of recognition is limited to particular applications.

The text prompted system represents a remedy to the text dependent system. In this case, the system prompts the user to utter particular phrases at the time of recognition. If the user cannot supply the prompted phrase (the phrase is provided from the recognition system at the verification time), he/she cannot access the system. This kind of recognition is also commonly used with speaker verification.

In the latter type (i.e Text independent), the speaker has the flexibility to say anything he/she wishes, so there is no specification of the phrase or words that the users are allowed to speak, since this type depends mainly on the characteristics of speech voice of the user (source /vocal tract characteristics) instead of focusing on the content of speech. Thus, the enrolment model and recognition utterance may not have the same content, and the recognition system needs to take the phonetic differences into account. This type is more flexible and applicable. It is also used in all branches of speaker recognition (Verification or Identification), but it suffers from difficulty in achieving high performance (Rao and Sarkar, 2014). Figure 1.3 illustrates the classification of speaker recognition systems.

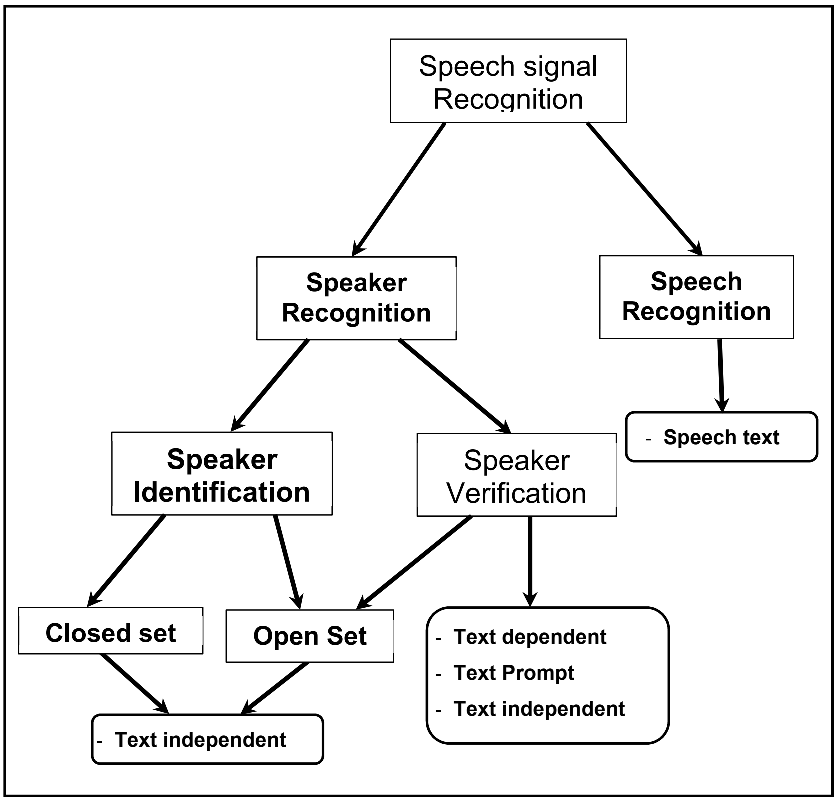


Figure 1.3: Classification of Speech signal recognition systems(Rao and Sarkar, 2014)

1.3 Review of State of the Art

Research studies in the fields of speaker recognition have been conducted for more than five decades, and this field is still an active area of speech signal processing (Furui, 2005, Nengheng, 2005). The primary motivation of such research was to study the way that humans can recognise talkers and the reliability of the listener in recognising the Talker (Stevens et al., 1968, Nengheng, 2005). The development in this field is mainly related with the evolution of speech and signal processing and computer technology (especially machine learning). One of the most significant studies in speaker recognition has used a spectrogram to identify the voice print of the speech (Nengheng, 2005). The next three decades have witnessed enormous technological progress in the speaker recognition fields, particularly when new features such as Linear Predictive Cepstral Coefficients (LPCC) (which represent a good implementation for differences of biological structure of human vocal system) and Mel-Frequency Cepstral Coefficients (MFCC) (which are commonly used to represent a human auditory system) have been used as features parameters in that field. Moreover, the development of statistical and artificial machine learning (such as Gaussian Mixture Model GMM, Hidden Markov Model HMM, and Support vector machine SVM) and various score normalisation techniques have been applied for speaker recognition systems. Recently, most of the systems in speaker recognition have evolved to adapt Gaussian mixture model-Universal Background model (GMM-UBM) solutions, Super vector methods (Campbell et al., 2006b), and factor analysis based engine voice (i-vector) framework (Sadjadi et al., 2013, Khoury et al., 2014).

1.4 Speaker Recognition Challenges

In spite of the developments in the field of speaker recognition, there are still more limitations and challenges in that field. The most accurate applications in speaker recognition systems in the real world still have limitations in becoming robust and reliable if we compare them with the recognition accuracy of the human ear. One of the primary challenges of speaker recognition is related to developing robustness of systems under mismatch conditions (Beigi, 2012). Many research studies deal with these cases, depending on finding new kinds of features, or propose using a new modelling instead of those classic models that were being used. In general, we can classify these challenges into two main categories: -

1. Intra-speaker variation (Time lapse effect), which occurs as a result of changes in phonation of the speaker due to change in the environment. The variations in intrinsic personal factors of the vocal system of the speaker have adverse effects on the performance of speaker recognition (e.g. Illness, whispering, speech under stress, ageing, surgery, and cold). This type can also be classified into short term (stress, illness, whispering, and cold) and long term (ageing). In addition, short utterances can also represent one of the intra-speaker variation challenges since relatively little information can be extracted from these types of signals.
2. Speaker-independent (Channel Mismatch): this refers to the variation between the reference models of the speech signals (enrolment models) and a given recognition speech signal for the same individual due to physical factors (e.g. change in the handset) or environmental factors (additive background noise, reverberation, and echo). Channel mismatch occurs as a result of the changes in characteristics of the recognized speech signal. In the past, it was thought that the mismatch was related to change of handset; however, in addition to handset mismatch, changes in environmental noise, acoustic properties of ambiance (e.g. echo and reverberation), microphone distance and angle (Far-field), and many other factors cause this kind of mismatch (Beigi, 2009). So far, different research efforts have been concerned with improving techniques for dealing with this kind of challenges. This has led to approaches that have the capabilities of explicitly modelling the channel variability of the input speech utterances. These techniques permit the effect of channel mismatching to be compensated for during the enrolment and recognition phases, causing significant improvement in accuracy when operating under adverse conditions (Pillay, 2010).

One of these fundamental challenges related to speaker recognition is any undesirable change in speech features due to environmental background noise. This kind of variation could cause a mismatch between the corresponding test (recognition) and the enrolment material of the same speaker, which would adversely affect the performance of the speaker recognition in terms of accuracy. Environmental noise represents one of greatest challenges for speaker recognition (especially for speaker verification) in matched and mismatched conditions, particularly for the latter case, which represents the greatest difficulty since it limits access to the authorised user(i.e. increase False rejection cases) (Ming et al., 2007) . Moreover, this problem has been aggravated by nature of speaker verification applications, due to the increasing probability that the input speech signal may be contaminated by different changeable and/or time-varying background noises, for instance, when a user attempts to log in to a device (PC , Handheld phone) using his/her voice print. In this case, background noises (speech babble, fan noise) which are usually not taken into account during the enrolment phase can adversely affect recognition of utterances produced from target speakers.

1.5 Research Motivation

As previously mentioned, the implementation of speaker recognition systems still has many limitations and challenges, since most of the recent applications of speaker recognition technologies are commonly used for application in public spaces (stations, hospitals, airports, etc.) and the internet. Moreover, providing a quiet environment at the time of recognition is sometimes difficult, if not impossible. This makes the environmental noises one of the main problems in this area, especially when the noise sources are non-stationary and time-varying. This research is motivated to investigate and develop a speaker recognition system under different background noisy environments, even when there is no information available about the identity of the noise.

1.6 The Aim and Objective of the Research

The aim of this research is to investigate existing approaches to speaker recognition in noisy conditions, and to develop an efficient technique to improve speaker recognition performance for real-world applications.

For this purpose, text-independent speaker recognition has been used, since this type is more flexible and can be easily generalised to other types of speaker recognition (i.e. text-dependent and text-prompted). Various techniques and experiments are involved to achieve this target, based on specific key objectives:

- Investigate the effectiveness of speech samples contaminated with different types of noise with different signal to noise ratios (SNRs) on performance of Speaker Recognition in context of biometric security. In this step, the effect of various noises with different SNRs on the performance of Speaker verification was investigated.

- Investigate the techniques used to clean or at least reduce the noise components and the side-effects these techniques impose on speaker recognition. The main purpose of this objective is to discover, to what extent, speech enhancement algorithms can improve the performance of speaker recognition in noisy environments.
- Specify the relevant feature extraction approaches that are used in the SR, especially under noisy environments. This can be done by investigating the features spaces that are less sensitive and more robust to noisy environments by using different speech verification sets from different data sets recorded in English and various other languages, followed by comparing these features with baseline features. For this purpose, Conventional Mel-frequency Cepstral coefficients (MFCC) - including the Dynamic Mel cepstral - were adopted, together with Gamatone Cepstral Coefficients (GFCC). The former features represent the most commonly used features in SR fields, while the latter represent noise robust features in the field of speaker recognition.
- Attempt to improve system robustness by adding different kinds of noise with different SNRs in the enrolment phase (the phase where the reference models are created) on the baseline speaker recognition system. Validation tests have been carried out with emulated noisy conditions with controlled SNR.
- Propose a new approach to address the issue of performance degradation in noisy environments. Based on the estimation of the signal-to-noise ratio (SNR) and profile of the ambient noise from input signals, the proposed method re-trains the enrolment models to generate new noisy models that adapt to the noise profile. This technique is termed "**Training on the fly**".
- Investigate the different techniques used to detect the additive noise in input recognition signal to improve the performance of training of the fly approach that adopted for speaker recognition application. For this purpose, two proposed techniques are adopted: the first is based on estimated power spectrum of the original noise using 1/3 octave band filter bank and white noise signal, while the second technique deals with extracting the noise from the input signal using a speech enhancement algorithm with spectral subtraction to find the noise in the signal.
- Critically evaluate the recognition systems, based on different kinds of evaluation techniques. Error Equal Rate (EER) and Detection Error Tradeof (DET) has been

adopted to evaluate the speaker recognition system. The main reason to adopt these metrics is that EER is generally used in the literature to evaluate the performance of Speaker recognition systems, while the DET graph represents the best graphical demonstration of both False Positive Rate and False Negative rate in recognition systems.

The main question that should be asked in this research is to what extent the proposed “Training on the fly” approach can improve text independent speaker recognition systems for real world applications under various environmental noises.

1.7 Research Methodology

The methods of this research, developed and adopted during the research programme include:

1. Investigating the research concerned with speaker recognition in general, and more specifically with speaker recognition in the presence of stationary and non-stationary noise background.

   At this stage, a review was undertaken of previous work concerned with the study of the degradation that occurs in the quality performance of speaker recognition due to the lack of relevant information from the speech signal as a result of noise existence (as seen in the next chapter). This has helped the researcher to acquire sufficient background knowledge about the best techniques used in clearing speech signals from noise to increase the performance of the speaker recognition system.
2. Investigating the effect of adding different types of noise to the speech signal with different signal to noise ratios (SNRs) on the performance of the speaker recognition (Chapter 4).

   At this stage, the effect of the existence of different types and levels of environmental noise in the speech signal on the performance of state-of-the-art speaker recognition approaches were studied, in addition to studying the degradation that occurs in recognition systems. The results of this investigation show the effect of different types of noise on the accuracy (in a biometric security context) of speaker recognition. In addition, they show that the signal-to-noise ratio (SNR) plays a major role in the robustness of speaker recognition, especially in determining the threshold at which the speaker recognition performance is affected by added noise.

3. Investigating the possibility of improving the robustness of speaker recognition by using noisy speech samples with different signal to noise ratios in the training of reference models (Chapter 5).

   At this stage, the robustness of speaker recognition when using controlled noisy reference models contaminated with different types of noise and a different signal to noise ratios were investigated, to find out to what extent the performance of speaker recognition could be improved when compared with SR based on reference models built by a noise clean signal. The results from this investigation suggests that if acoustic conditions can be estimated and suitably pre-trained models chosen, the robustness of the system can be improved.
4. Investigating different types of feature extraction algorithms (Chapter 6).

   Two different types of features extraction algorithms were used as a baseline for speaker recognition.

   At this stage, two different types of features were examined(MFCC,and GFCC) , and suitable features were identified by investigating their sensitivity to environmental noise as well as to different languages, and their effectiveness for the speaker recognition system. The results from this chapter show that GFCC has better recogntion perfomance than conventional MFCC. However, the language mismatch of the recogntion set shows adrop in recognition performance of GFCC both in clean and noisy speech.
5. The impact of using speech cleaning algorithms was studied and they were compared for their robustness as a pre-processing stage for speaker recognition (Chapter7). For this purpose, seven types of speech enhancement approaches were employed. The results show that this kind of approach is unreliable for robustness of speaker recognition in different noisy environments.
6. A speaker recognition framework has been developed based on extracting relevant information from the noisy speech recognition signal, to create an adaptive reference model close enough to the input signal. This approach is called “Training on the Fly” for robust speaker recognition (Chapter 8).
7. The newly developed approach was evaluated by checking its performance and robustness to a noisy environment (Chapter 8 and Chapter 9). The proposed approach

based on the two detection noise techniques shows significant improvement in performance of the MFCC-GMM baseline when compared with the performance of conventional speaker recognition.

8. Making appropriate modifications to the developing approach in order to improve its performance, based on evaluating the results in step 7 and using the features in step 6. The evaluation results show that training on the fly using the MFCC-GMM baseline is more efficient than the GMM-GFCC baseline. However, the improvement of the GFCC-GMM baseline is limited to low SNR.
9. The final stage will be making final conclusions and recommendation and disseminating results. The main steps of the methodology are shown in Figure 1.4.

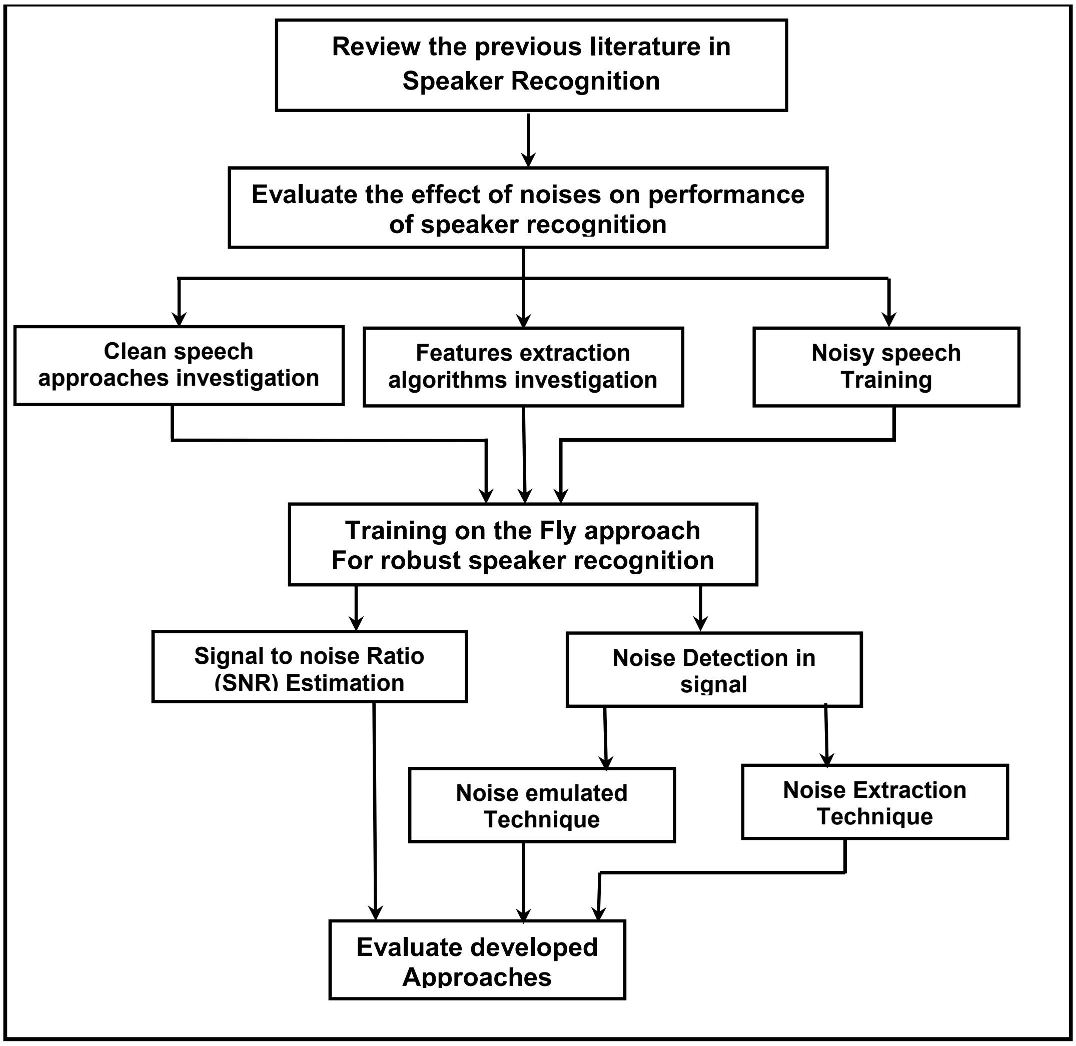

Review the previous literature in Speaker Recognition
Evaluate the effect of noises on performance of speaker recognition
Clean speech approaches investigation
Features extraction algorithms investigation
Noisy speech Training
Training on the Fly approach For robust speaker recognition
Signal to noise Ratio (SNR) Estimation
Noise Detection in signal
Noise emulated Technique
Noise Extraction Technique
Evaluate developed Approaches

Figure 1.4: Block Diagram of Research Methodology

1.8 Contribution of the Study

This study makes the following contribution to finding the best approaches that have been used in improving the performance of speaker recognition system in noisy environments and decreasing the mismatch between the enrolment reference models and input recognition speech signal.

1. It has implemented a text-language independent speaker verification tool box based on the state of the art Gaussian Mixture Model GMM and Gaussian Mixture Model Universal Background Model GMM-UBM and investigated the effect of artificial noisy speech signal contaminated with different types of noise (five types of noise were chosen) with various SNRs on the speaker verification.
2. It has investigated the effect of using noisy speech signals contaminated with various kinds of noise and with various SNRs to create reference models on the accuracy of the speaker verification toolbox.
3. In this research, all investigations and experiments were applied on two different types of speech databases. The first database is named TIMIT, one of the most commonly used databases in speech and speaker recognition fields since it is one of the standard databases used to evaluate speaker recognition in real-world environments. The second database includes the speech samples recorded by research in the anechoic chamber in the University of Salford (named SALU-AC). What makes this database unique is that the environment is used to record its speech samples, which represents the cleanest environment from adverse conditions since the chamber used has a measured background noise of -12.4 dBA. This makes the artificial noisy speech samples generated from this database free from other adverse conditions, making it easy to deal with the noise issue separately from other adverse conditions. In addition to the environment that it was collected in, the SALU-AC database includes speech samples from different languages.
4. Using speech samples from various languages, the robustness of the feature spaces used in speaker recognition in noisy environments has been investigated, and the result compared with the one obtained from the English samples set.

5. A new approach has been developed to deal with the degradation issue that occurs in speaker recognition under environmental noise. In contrast to other approaches, which focus on dealing with the input recognition signal, this approach is concerned with creating an instant adapted reference model based on SNR level and identity of noise that is compared with the input signal to decrease the mismatch between the reference model and input signal.
6. To detect the noise in the recognition signal, two techniques are proposed, the first one based on using 1/3 octave filter bank to estimate the profile of the noise and then creating an emulated noise close to the original that is contaminated with the recognition signal. The second is based on using a speech cleaning algorithm and spectral subtraction to extract the noise from a noisy recognition signal.

1.9 Thesis Outline

The thesis is organised into ten chapters. A summary description of each chapter is given below:

Chapter 2: Background and Literature Review

This chapter starts with a description of the process of human speech production. It then presents a review of the literature in the area of automatic speaker recognition. This includes a review of feature extraction algorithms, speaker modelling approaches and the state of the art of these approaches, and the evaluation metrics of speaker recognition. Furthermore, it describes the approaches that have been used to improve the performance of recognition systems under mismatch noise conditions.

Chapter 3: Speech and Noise databases

This chapter describes the speech databases used for the purpose of the investigations in this research and then focuses on presenting the new speech database SALU-AC, and the types of noise used in the experiments of this study. Finally, the process of obtaining noisy speech is then explained.

Chapter 4: GMM-UBM based Speaker Recognition

This chapter is concerned with speaker recognition based on GMM-UBM which is adopted in this research and how this system works, followed by an experiment to investigate the effect of various kinds of noise with different SNRs on the performance of MSR speaker recognition.

Chapter 5: Improving the Robustness of Speaker Recognition in Noisy Conditions via Training

In this chapter, an attempt is described to improve robustness by using noisy speech to create training reference models. First, the state-of-the art techniques for modelling and classification models are explained, including GMM and GMM-UBM, followed by an experiment that explains the difference between using GMM baseline and GMM-UBM baseline with limited data. Finally, details of an experiment using noisy reference models with various types of noise and SNRs are presented, and results of these experiments are then discussed.

Chapter 6: Comparison of Feature Spaces for Robust Speaker Recognition and Mismatched Languages

In this chapter, we focus on making a comparison between the robustness of two types of features extraction algorithms on the performance of speaker verification. Traditional Mel-frequency Cepstral coefficients (MFCC), including the dynamic features, together with Gammatone Cepstral coefficients (GFCC) have been adopted in this study. Furthermore, this comparison is extended to include investigations into language independency features in recognition phases. An explanation of a process for extraction of each feature is also presented.

Chapter 7: The Impact of Using Speech Enhancement techniques on Performance of Speaker Recognition.

Different types of speech enhancement algorithms have been investigated and compared for their robustness as a pre-processing stage for speaker verification system. Moreover, review descriptions of these approaches and their types have been presented.

Chapter 8: Training on the Fly for robust speaker recognition

This chapter includes a description in detail of the approach proposed in this thesis to improve the performance of speaker recognition in environmental noise, followed by the techniques used for signal to noise ratio estimation, and noise profile detection using 1/3 octave filter. Finally, results of an experiment to evaluate the proposed approaches are discussed.

Chapter 9: Training on the Fly using noise extraction technique.

Training of the fly based on different techniques to detect the noise in the recognition signal are presented in this chapter, followed by experiments to evaluate training on the fly using this technique and comparing results with those obtained from Chapter 8.

Chapter 10: Discussion, Conclusions and Recommendation for Future Works

The last chapter presents the overall conclusions in this research. Furthermore, some suggestions for future research are provided in this chapter.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

Chapter Overview

Over the last fifty years, research in the fields of speaker recognition has yielded many developments, especially in feature extraction and modelling approaches. These range from high level features (such as phonemes, dialects) and template classification (such as vector quantization, and time warping) to cepstral features (such as MFCC and GFCC) and state of the art modelling (such as GMM-UBM and i-Vector). This chapter presents a background review of the general structure of speaker recognition systems as well as some literature on the various techniques that have been used to implement the front end and back end of speaker recognition, in addition to the most common approaches used to improve the performance of speaker recognition in noisy environments. The chapter begins by explaining the process of human speech production. Details on the structure of speaker recognition are given in section 2.2, and the most conventional algorithms for feature extraction in section 2.3. The discussion in Section 2.4 is concerned with the main speaker modelling and classification approaches and what state of the art techniques are used. Section 2.5 then describes speaker recognition evaluation techniques. Finally, there is a detailed review in section 2.6 of the commonest methods in speaker recognition to improve performance in noisy environments.

2.1 Human Speech Production

To clarify the mechanism of human speech production (which is considered to be important in identifying the discriminative characteristics between different speakers), one should first understand the anatomy of the speech production system (Vocal system organs). In general, the vocal system consists of three main parts (Loizou, 2013, Beigi, 2011):

1. Lungs: The lungs represent the power supply or pressure system which is responsible for pushing the airflow through the trachea to the larynx. The role of the lungs in this process is no more than to provide the appropriate pressure, which plays a major role in speech production. Figure 2.1.a.
2. Larynx: also called the voice box since it represents the main section for generating the voice based on the laryngeal parts located on top of it called vocal cords (vocal folds). Vocal cords are two folds of tissue joined to the thyroid. The triangular opening located between the two folds is called the glottis, which represents a transition area between the larynx and pharynx Figure 2.1 (b). When airflow passes through the glottis, it will cause vibration of the vocal cords and modulate the air to generate the human voice in a process called *phonation* (Beigi, 2011).
3. Vocal Tract: this refers to the tube that starts from the top of the larynx and ends in the nasal and oral cavity. This section is responsible for shaping the modulated air coming from the larynx by changing the shape of different organs (pharynx, tongue, jaw, and lips) and it works as a filter for modulated air.

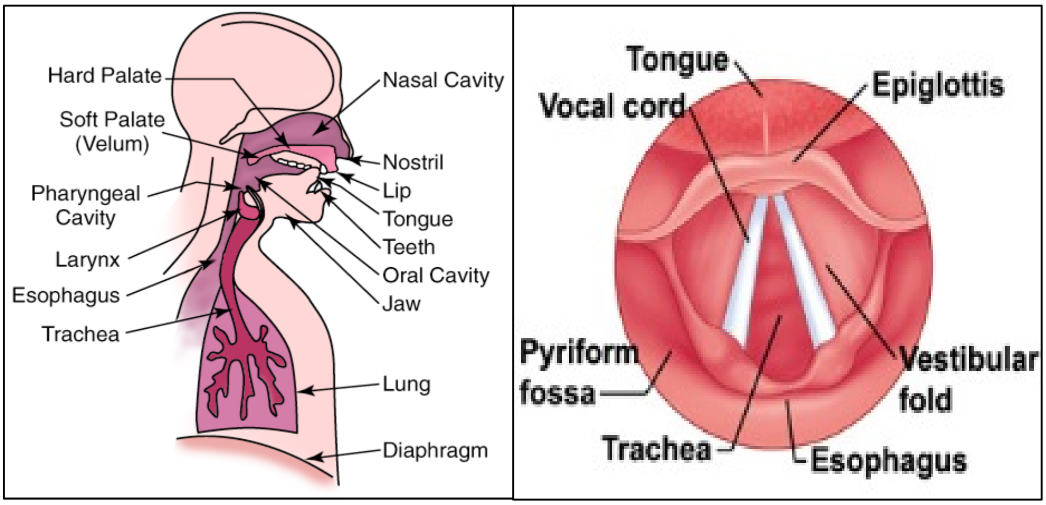


Figure 2.1: (a) The sagittal section of Speech production system (b) Vocal Cord

The process of generation of human speech involves pushing the inhaled air from lungs through the trachea to the larynx. The pressure of the air can depend on the mood of the speaker (where the pressure increases if the speaker is angry or tense). This process of pushing air from the lungs is called ***respiration*** (Nengheng, 2005). The pushing air is then passed through the glottal region to the larynx causing vibrating of the vocal cords at a rate based on their length, thickness and tension. This vibration depends mainly on the status of the vocal cords, such that when the vocal cords are tensed the airflow is transformed into a semi-periodic waveform known as voiced sound and this process of modulating the air flow is called ***phonation*** (Quatieri, 2002). Figure 2.2 presents the semi-periodic glottal waveform, which can be represented as a function of time. When the cords are in closed position, the speech signal starts from zero and increases to reach the peak which represents the loudness of the voice. It then swiftly decreases to zero when the vocal cords suddenly close. The time interval when the vocal cords are closed and no air flow is passing through the glottis is called the *closed phase,* and the time period from when the signal increases from zero until reaching the top of the pulse is referred to as the *open phase*, and finally the period from the maximum signal to when the vocal cords close again is referred to as the *return phase*. The shape of this signal can be varied depending on the speaker, the speaking style, and the specific speech sound. Sometimes the vocal cords do not completely close; in this case, the closed phase does not exist (Quatieri, 2002).

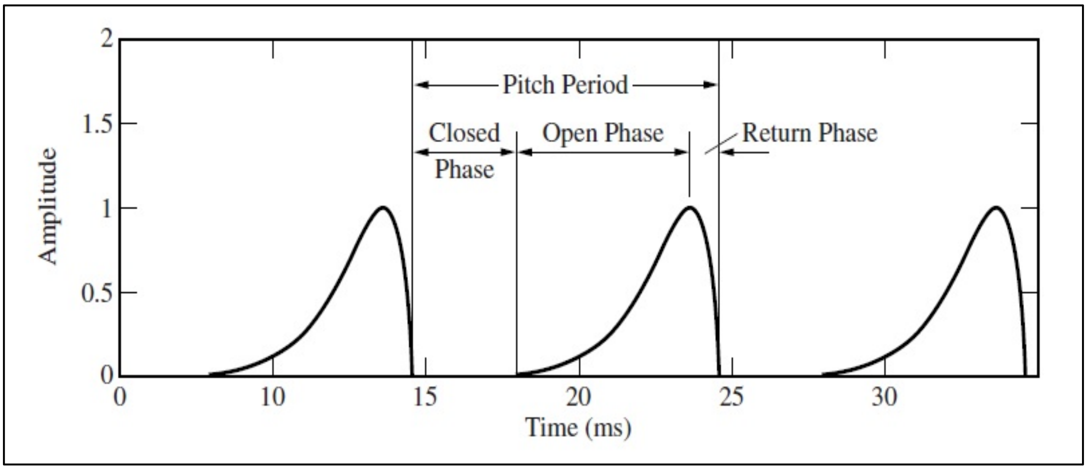


Figure 2.2: Periodic Glottal Waveform (Quatieri, 2002)

On the other hand, when the vocal cords are relaxed the airflow comes through causing turbulence in the vocal cords instead of vibration, which results in a non-periodic speech waveform known as unvoiced sound; and this turbulence at the vocal cords is called ***aspiration*** (Quatieri, 2002).The voiced sound is then filtered by the vocal tract system which acts as a variable acoustic filter. This filter plays a role in varying the amount of acoustic energy by changing the shape of the vocal tract organs (Jaw, tongue, lips and other internal parts) so that the natural resonances happen at different frequencies. This process is called ***articulation*** (Nengheng, 2005) and these resonant frequencies called the formants. Finally, the volume air flow that passes through the vocal tract is radiated as a pressure signal at oral or nasal cavity with a process called lips radiation (Jeevakumar, 1993, Holmes, 2001).

There are various factors, depending on physical and behavioural characteristics of the speaker, which make the speech waveform for each speaker unique. Thus, different levels of information can be extracted from this speech signal to recognise the differences between speakers. This extracted information is known as ***feature parameters*** which represent speaker identity (Pillay, 2010, Nengheng, 2005).In Section 2.3 we focus more on these features and the most common features used in speaker recognition.

2.2 The Process of Speaker Recognition

The process of speaker recognition in both types (Verification and Identification) can be divided mainly into two phases, firstly the ***Enrolment phase*** (***Training phase***) which is responsible for creating a reference speaker model (or models) from previous speech samples with some features extraction and modelling approaches. In the ***Recognition phase*** (***Testing phase***), on the other hand, the features of the speech signal are extracted in the same approach as that used in the enrolment phase. These features are then matched with the reference model to measure the similarity between them, based on one of the matching techniques. In general, the stages of automatic speaker recognition can be outlined as follows(Rao and Sarkar, 2014, Chauhan et al., 2013):

1. **Pre-processing**

   The essential objective of this stage is to digitise the analogue speech signal by sampling it at a standard sampling frequency. By using a high pass filter, this digital signal is usually pre-emphasized by emphasising higher frequency components and compensating for the human speech production mechanism which tends to attenuate them (Rao and Sarkar, 2014). Voice activity detection (VAD) has been used in this stage to split the speech segment from a given audio signal.

2. **Features Extraction**

   Also known as front-end stage, this stage is responsible for transforming the speech signal into a set of parameters, known as feature vectors, that contain the discriminative characteristics of the speaker. In the enrolment phase, the features are extracted from the speech signal and used to generate a reference model for the target speaker. In the recognition phase, these features are derived from unknown speaker utterances and compared with a reference model for the claimed speaker to give similarity scores (Kinnunen and Li, 2010, Rao and Sarkar, 2014).

3. **Speaker Modelling**

   The primary goal of this stage, which also known as back-end stage, is to create templates or models for each enrolled speaker based on statistical modelling techniques. These techniques are employed to capture the distribution of features extracted from the registered speaker. More details on these techniques are given in section 2.4.

4. **Pattern matching and classification**

   In this stage, the recognition (test) utterance is matched based on its statistical similarities with a reference model (or models) for the claimed speaker. This pattern match depends completely on the techniques used to create the speaker models. A decision is made based on the scores obtained from this match such that the recognition utterance is classified as reference model generating a maximum score (Rao and Sarkar, 2014). Figure 2.3 and Figure 2.4 illustrate the structure of both speaker identification and speaker verification.

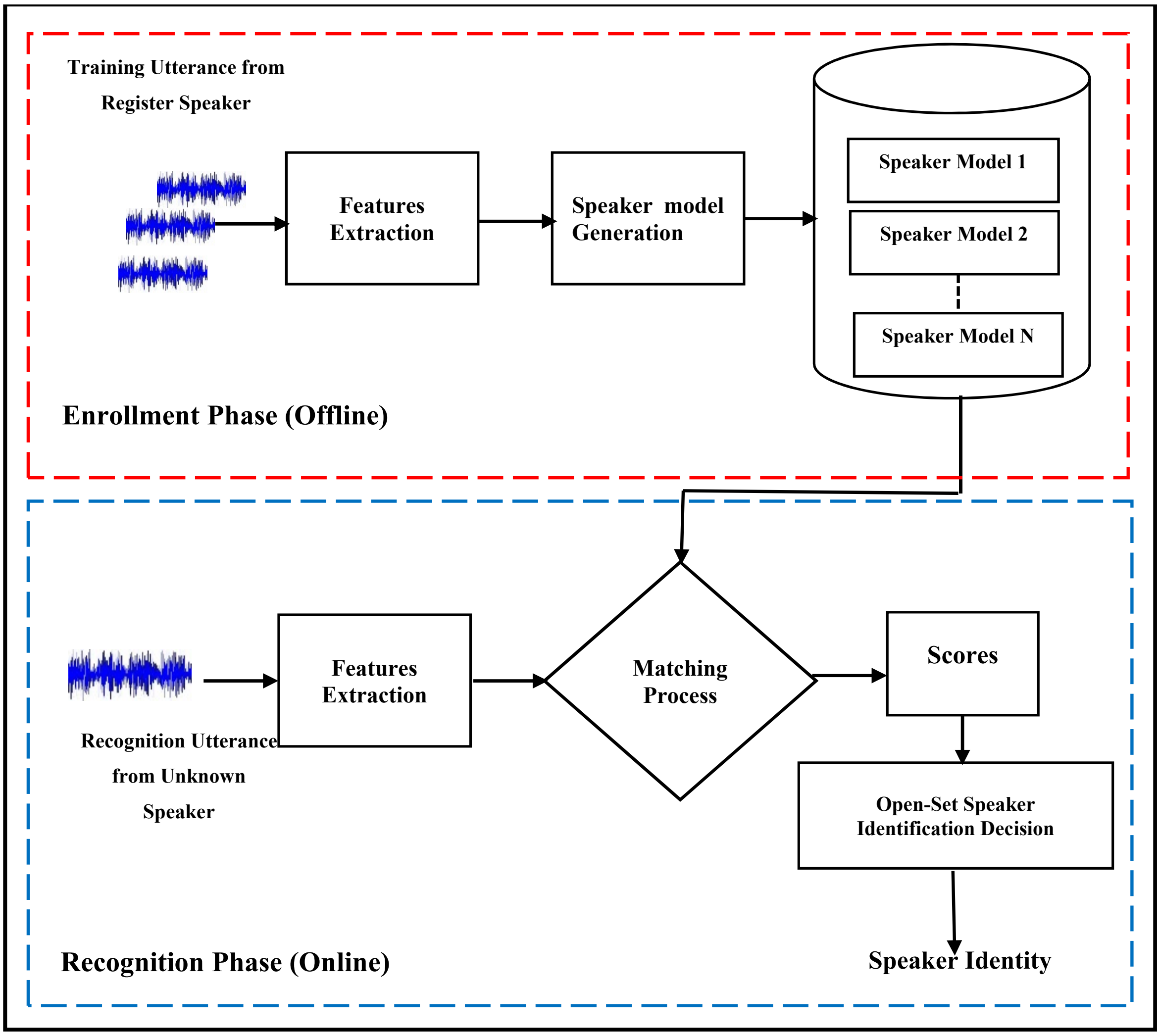


Figure 2.3: General approach to Speaker Identification

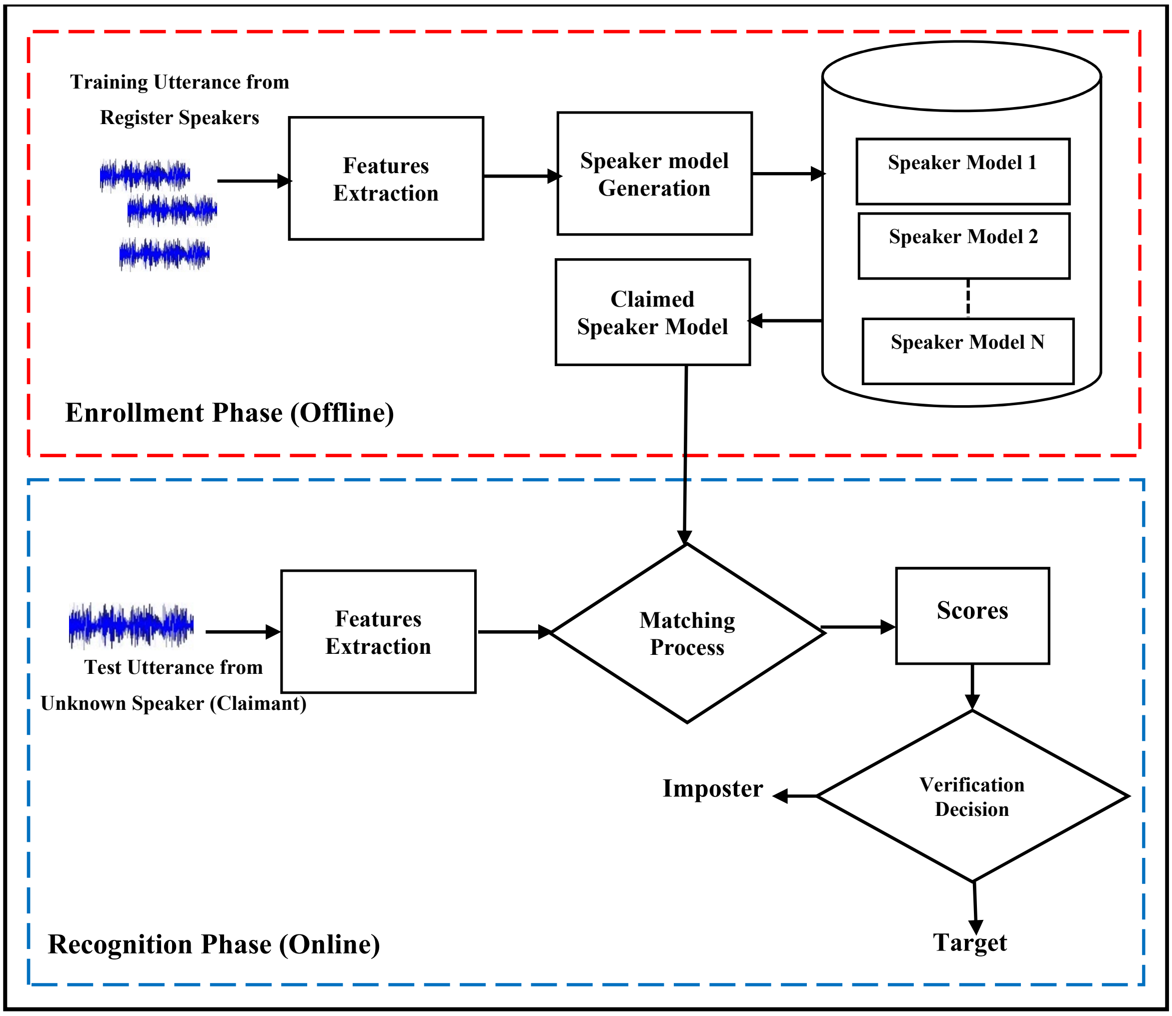


Figure 2.4: General approach to speaker verification

2.3 Review of Audio Features

As mentioned earlier, the primary objective of features extraction is to extract the unique characteristics of the speech signal which represents the speaker identity. These features represent the differences that exist in the vocal system (source /vocal tract) as well as the style of human speaking (Nengheng, 2005). There are different ways to classify the features. In general, the features can be classified into three hierarchical main types based on their physical interpretation (Kinnunen and Li, 2010, Nemati and Basiri, 2011) (Figure 2.5):

- Low-level spectra features: these kinds of features are related to the structure of the vocal system, which comprises short-term spectral features (vocal tract features) and source features. These features represent the descriptor of the short-term spectral envelope which is an acoustic correlation of timbre (colour of voice), in addition to the resonance characteristics of the supralaryngeal vocal tract (Kinnunen and Li, 2010).

- Prosodic and Spectro-temporal features: These refer to features that include intonation rhythm, and pitch. These types of features are mainly based on phoneme of the speaker.
- High-level features: These kinds of features are related to the way that the person speaks, including phonemes, accent, and pronunciation. These features usually try to capture properties included in the conversation level of speakers, such as the characteristics included in words (Doddington, 2001). High-level features are commonly used in speech recognition compared to speaker recognition.

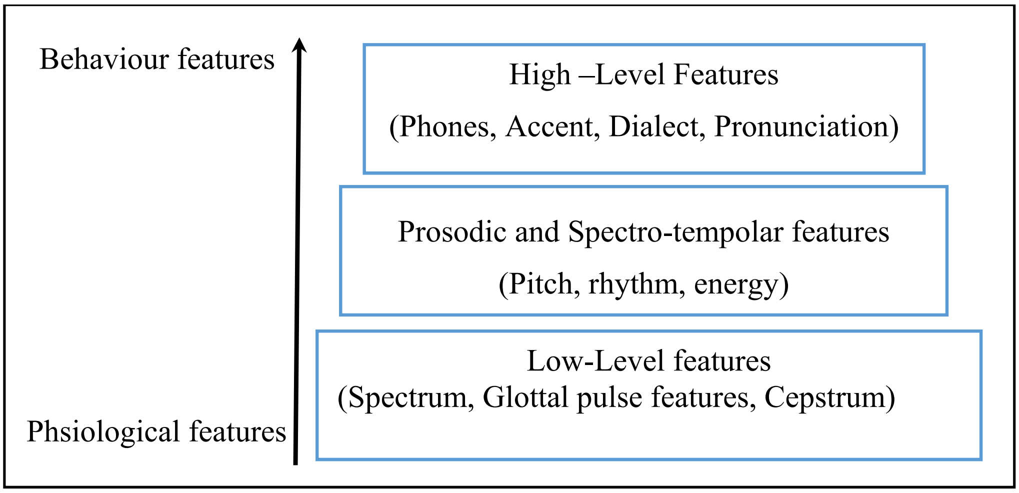

Behaviour features
High –Level Features
(Phones, Accent, Dialect, Pronunciation)
Prosodic and Spectro-tempolar features
(Pitch, rhythm, energy)
Low-Level features
(Spectrum, Glottal pulse features, Cepstrum)
Phsiological features

Figure 2.5: Hierarchy of Speech features

Much research has used a high-level and Prosodic features in speaker recognition systems, since they are more efficient and stronger in speaker recognition and more efficient against channel mismatch and noise. However, high-level features need more complex front end (i.e. more complex features extraction approaches), and require massive training data for modelling, and decision making delay. Furthermore, these kinds of features are less disseminated and easier to impersonate. Due to these limitations, high-level features are not commonly used with speaker recognition, compared with their usage in speech recognition (Pati and Prasanna SR, 2010, Kinnunen and Li, 2010).

One of the earliest methods based on using Prosodic and Spectro-temporal features in speaker recognition is presented by Atal (1972). The method depends on using a temporal variation of pitch in the speech signal for identifying a feature in text-dependent speaker recognition. These pitch data were obtained from 10 speakers with six utterances of the same sentences for each speaker. He used a 20-dimensional vector to represent the pitch feature and form the reference model for each speaker by averaging the transformed pitch of that speaker, while the recognition procedure was based on computing the Euclidian distance between the enrolment and recognition vectors. Based on these experiments the author recommended the possibility of using pitch contours for ASR. However, he also mentioned that pitch contours and short-time spectrum were not enough for automatic speaker recognition ASR.

Shriberg et al. (2005) presented a method to compute the various duration, pitch, and energy features for each estimation of a syllable in the output of speaker recognition. They referred to these features as SNERF-grams (N-grams of Syllablebased Nonuniform Extraction Region Features) and they used support vector machines (SVMs) to model a speaker's references. This research found that adding SNERF-grams improved the system from Error Equal Rate (EER) (which represents one biometrics metric as described later) 6.418 to 5.783 when compared with a system without this type of features. The authors, using SNERF, provide an analysis that can lead to better understanding of how speakers differ prosodically in a mostly voluntary way.

Mary and Yegnanarayana (2008) presented an approach for extracting prosodic features from the speech signal in both speaker and language verification systems. In this work, syllable units are selected as the main units for representing the prosodic characteristics. The continuous speech was segmented into syllable-like units by finding vowel onset points (VOP) automatically. The information from these VOPs plays a role in extracting prosodic features from the speech signal. The combination of prosodic features with spectral features shows a significant improvement in the performance of speaker recognition, with an EER of 9.3%.

Jin et al. (2007) explored using high-level linguistic features for robust speaker recognition in a far-field microphone situation. They applied n-gram models trained on multilingual phone strings to capture speaker idiosyncrasies. n-gram models are created by decoding speech by various phone recognisers and employing the relative frequencies of phone n-gram as features for training the models. This approach shows the high robustness of high-level features for mismatch conditions. However, one main drawback of this phonetic speaker recognition is the need for a huge amount of training data to reliably estimate phonetic n-gram models. Moreover, this system is sensitive to language variety, which adversely affects the recognition result.

Wang et al. (2017) proposed novel acoustic features that are based on spectral-temporal receptive fields (STRFs). The STRF is derived from physiological models of the auditory system in the spectral-temporal domain, in which the temporal response and temporal response in the signal had been specified by the rate (in Hz) and scale (in cycle/octave) respectively. To evaluate the proposed features, the experiments were conducted on 36 speakers using a support vector machine (SVM) to classify the speaker. The proposed feature set increased the speaker recognition accuracy by 3.85% and 18.49% on clean and noisy environments respectively, when compared with MFCC Baseline system.

Low-level features (especially the Cepstrum features), on the other hand, are easy to compute, show good performance and are more commonly used in speaker recognition than higher-level features. Besides, the usage of low-level features means including all information is retained and nothing is discarded. In the next sections, we focus mainly on the most commonly used features in speaker recognition and some of the modern features used in this field, reviewing some of the literature for each one.

2.3.A Mel Frequency Cepstral Coefficients (MFCC)

This kind of feature represents the most commonly used feature in speaker recognition. Most researchers use it as a baseline feature for constructing speaker recognition (Hasan et al., 2004, Muda et al., 2010). Moreover, these kinds of the feature are commonly used in conjunction with other types of features to make the speaker recognition more robust, since it is used to estimate a human auditory system (Beigi, 2011). In this extracting method, a filter bank is applied such that each filter applies to a different frequency band of the speech spectrum. The arrangement of these filters depends on the human perception of speech. This includes the spacing of band-pass filters based on the Mel scale (Pillay, 2010) Figure 2.6. The logarithm of energy from each filter is then computed and accumulated, before applying Discrete Cosine Transform (DCT) to obtain Cepstral coefficients. MFCC represents an ideal candidate for both speaker and speech recognition, since they contain significant information about the structure of the speech signal (Beigi, 2011). Mel Cepstral Dynamics, on the other hand, are the extracted dynamics of MFCC features. This type of Cepstral coefficient is driven by using first and second order differences which are somewhat independent of the actual MFCC. Hasan et al. (2004) applied MFCC as a feature for speaker identification with vector quantization VQ to minimise the data of the extracted feature. They found that the combination of Mel-frequency with a Hamming window gave the best performance. They also mentioned that the Mel scale is less vulnerable to the change of speaker vocal cords in the course of time. Sinith et al. (2010) adopted MFCC as a speaker speech parameter for a text-independent speaker identification system based on a Gaussian mixture model (GMM) to trained four speaker model with 4, 8, and 16 Gaussian mixture components. This experiment was performed on using different speech time durations and several languages. Depending on the maximum likelihood ratio algorithm for the process of making a decision, the recognition rate shows the result (98.8%) when speech utterance is of 60 seconds duration and a number of the Gaussian mixture is 16 components. This experiment depends on only four speakers.

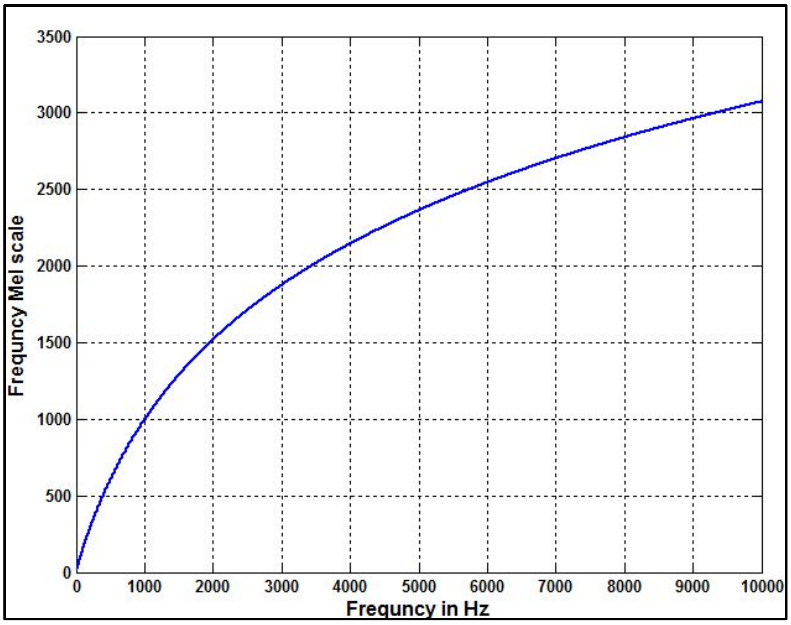


Figure 2.6: Mel-Scale representation

The model was proposed by Ezzaidi and Rouat (2004) to use fundamental frequency F0 together with MFCC to represent information of vocal source/vocal tract in a speaker identification system based on Gaussian Mixture Model GMM modelled by 32 and 4 Gaussian mixture respectively. This experiment was carried out on speakers from a speech telephony database (SPIDRE) recorded from the various handset telephones. The system based proposed features were compared with a baseline system functioning on all voiced and unvoiced speech segments and with another system based on voiced segments only. The results illustrate limited improvement in performance.

The conventional MFCC can characterise the voiceprint only, but it does not take into account the dynamic characteristics of speech (Weng et al., 2010). Furthermore, MFCC parameters are very susceptible to interfering noise (Wang et al., 2009). To deal with this limitation Wang et al. (2008 and 2009) proposed an algorithm to extract the dynamic of the Mel Cepstral coefficients (also known as the Delta and Delta-Delta Cepstral coefficients) which combines the information obtained by MFCC with pitch. This dynamic can be obtained by dynamically constructing a set of Mel-filters according to pitch detection. After that, the Mel-filters are applied to extract these feature coefficients which represent the speaker identity characteristics. The experimental results from the dynamic MFCC show greater robustness in a noisy environment. However, the drawback of this algorithm is that it increases the parameter dimensions and computational complexity of SR systems. An optimisation algorithm was proposed by Weng et al. (2010) to extract a new series of coefficients called weighted dynamic MFCC. This approach was based on combining MFCC and dynamic MFCC as a new series of coefficients and then using these coefficients as a parameter of the Gaussian Mixture Model. This work depended on speech samples obtained from the TIMIT and VOA databases. The results show this weighted dynamic MFCC could achieve a higher recognition rate over conventional MFCC and a combination of MFCC and dynamic MFCC. Since MFCC represents the most common feature spaces in the speaker recognition field and represents baseline features for most speaker recognition research, then the MFCC is adopted in this work. More Detail on the MFCC features extraction is given in chapter 6.

2.3.B Linear Predictive Cepstral Coefficients (LPCC)

LPCC are one of the important features used in speaker recognition since they represent the differences of biological structure of the human vocal tract (Yujin et al., 2010). LPCC is the result of recursive computing from Linear Predictive coding coefficients (LPC) to Linear Predictive Cepstral coefficients according to an all pole filter (Beigi, 2011, Yujin et al., 2010). In LPC analysis this all pole filter represents a model for the vocal tract depending on the assumption that for voiced sounds, the excitation (vocal cord vibration) can be modelled by an impulse train generator which represents the series of nearly periodic glottal pulses produced by the vocal cords. For unvoiced, a random noise generator is applied to represent turbulent air flowing through a shrinkage of a long vocal tract. After that, the speech sample approximates as a linear combination of previous samples. The output of the LPC analysis is then obtained by a vector of predictor coefficients, which represent the parameters of the vocal tract structure for each speech frame. These are obtained by minimising the predictor error. Cepstral analysis depends on the principle of homographic signal processing, offering an intuitive approach of changing the convolutive connection between the fast and slow varying aspects of the speech spectrum into summation(since the speech signal represents a convoluted combination of excitation of the vocal cords and impulse response of the vocal tract). This enables these components to be separated easily (Pillay, 2010).

LPCCs, just like MFCC, are commonly used for speaker recognition. One of the earliest studies done by Furui (1981) used LPCC features by means of LPC analysis of fixed sentence long utterances to form a reference model for a speaker verification system. Furui used a time warping method to build the training reference models of speakers, and the verification decision was made based on the overall distance. Three sets of utterances based on telephone speech were used to evaluate this system. The evaluation indicated mean error rates (which represent the average of False acceptance rate and False rejection rate) of 0.19%, 0.36, and 0.77% respectively for each utterance set. Yujin et al. (2010) made comparisons to measure the strengths and weaknesses of using LPCC, MFCC and the differential cepstrum of both features (named *LPCC,*MFCC) as extracted features for a text independent speaker recognition system. In addition, they studied the effect of combining both kinds of features on the performance of the recognition system. Vector quantization and Dynamic time warping were used in the system to recognise the speaker identity. This experiment was applied to a speech recording library that included 40 speakers, and the results of this experiment showed that the combination of these features improves recognition rates in the system over LPCC-*LPCC and MFCC-*MFCC baseline systems.

2.3.C Perceptual linear prediction (PLP)

PLP was proposed by Hermansky (1990). These kinds of features are extracted in a three step process: Firstly, the short-term spectrum is processed according to human perception of tones.

In this case, the centre frequencies of filters are spaced equally on a Bark scale (Pillay, 2010). Then, PLP analysis compensates for variations between the actual and perceived loudness of tones, which happen at different frequencies; finally using all pole modelling on the resulting spectrum to produce PLP parameter (Figure 2.7). The resulting parameters can then be transformed into their Cepstral coefficients (in the same way as in LPC) by the recursive relationship between the prediction coefficients and the Cepstral coefficients. These coefficients are known as Perceptual Linear Prediction Coefficients (PLPC) (Pillay, 2010). Unlike LPCC and MFCC, which are popularly used in speaker recognition (Pillay, 2010, Nengheng, 2005), PLPC is widely used with MFCC in speech recognition (Davis and Mermelstein, 1980, Picone, 1993).

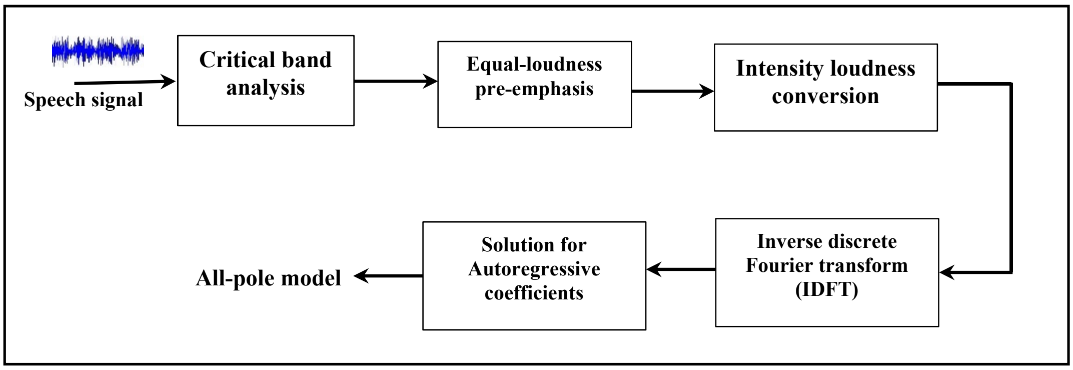

Speech signal
Critical band analysis
Equal-loudness pre-emphasis
Intensity loudness conversion
Inverse discrete Fourier transform (IDFT)
Solution for Autoregressive coefficients
All-pole model

Figure 2.7: Block diagram of PLP speech analysis based on (Hermansky, 1990)

Revathi and Venkataramani (2009) investigated the effectiveness of combining each of PLP and MF-PLP (Mel-frequency perceptual linear predictive cepstrum which uses the same process of extracting PLP but with a Mel-filter bank instead of a Bark filter bank) with a pitch for a robust text independent speaker identification system. In this system, training models are developed by a Vector Quantization (VQ) codebook of size M=L/10 , which is designed to represent L vectors of training data and the system evaluated by measuring minimum distance between test features and clusters. The database used in these experiments contains 8 speakers selected randomly from 8 different dialect regions and 8 speakers from the same dialect region in the TIMIT speech database. The results of these experiments show that both perceptual features (PLP, MFPLP) combined with pitch based systems have a higher accuracy than using perceptual features only without combination for both datasets.

2.3.D Cochlear Filter Cepstral Coefficients (CFCC)

An auditory-based feature extracted algorithm is presented by Li and Huang (2010), and Li and Yan (2011) based on an auditory transform in addition to a set of modules to emulate the signal processing functions inside the cochlea (Li, 2009). The auditory transform in this algorithm is a wavelet transform which comprises a pair of forward and inverse transforms. By using a bank of cochlea filters, the speech signal can be decomposed into a number of frequency bands in a forward transform. The frequency distribution of these cochlear filters is identical to the cochlea of the human ear and the impulse response of the filters is identical to that travelling wave (Li and Huang, 2010). In the inverse transform, the speech signal can be reconstructed from decomposed band-pass signal. Figure 2.8 illustrates the block diagram of CFCC extraction. The speech signal passes through the cochlear filter bank, then hair cell functions with variable windowing are applied, after that cubic-root nonlinearity, and finally a discrete cosine transform (DCT) has been used to obtain CFCC features.

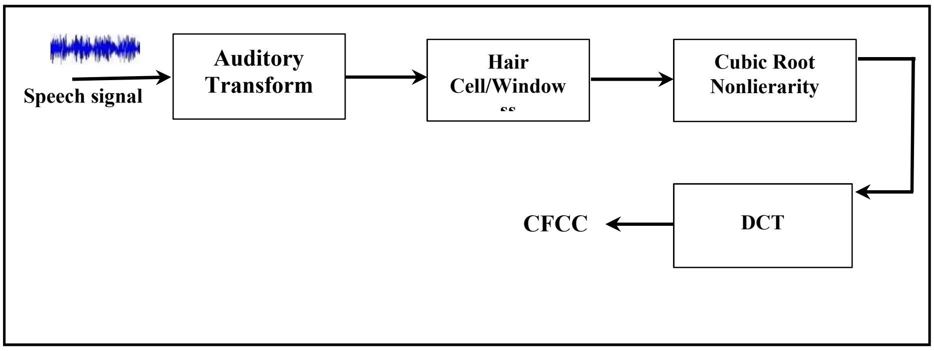


Figure 2.8: Block Diagram of CFCC Feature Extraction

The CFCC shows significant performance in speaker recognition under noisy conditions particularly in low signal to noise ratio (SNR). Li and Huang (2010) and Qi and Yan (2011) made a comparison between CFCC based speaker identification system with different types of features (MFCC, PLP, and RASTA-PLP) in clean and different noisy environments. Three types of noisy speech were used in these experiments: white noise, car, and babble. The results show that CFCC and MFCC features have the same accuracy with 96% in clean environments, while CFCC appears more robust in noisy environments with 88.3% against 41.2% accuracy for MFCC. On the other hand, the results show that CFCC features outperform PLP for all types of noise and outperform RASTA-PLP under white noise, while both CFCC and RASTA-PLP showed the same performance under car and babble noise.

2.3.E Gammatone Frequency Cepstral Coefficients

Further auditory features were proposed, firstly by Shao et al. (2007), and Shao and Wang (2008), to improve the robustness of speaker recognition for noisy speech. This kind of feature simulates 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 the Gammatone filterbank (Shao and Wang, 2008, Beigi, 2012). This filter bank consists of 128 filters (or sometimes 64-filters) centred the frequencies that are quasi-logarithmically spaced from 50 Hz to 8 KHz and equally distributed on an *equivalent rectangular bandwidth* (ERB) (Patterson et al., 1992)which models the human cochlea (Shao et al., 2007) (Figure 2.9). The gammatone filter bank is responsible for decomposing the input signal into the time-frequency (T-F) domains. This T-F demonstration is a variant of cochleagram (Wang and Brown, 2006). The Cochleagram has the ability to retain the higher frequency resolution at the low-frequency range for the same number of frequency components. This ability makes the Cochleagram different from the spectrogram which has linear frequency resolution. The time frame of the 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 (GFCC). Despite being known as cepstral coefficients, GFCCs are not classified as one of them, since cepstral analysis requires a log operation between first and second frequency analysis for deconvolution purposes. However, it is considered as one of them because the procedure used to derive them is similar to the one used with MFCC.

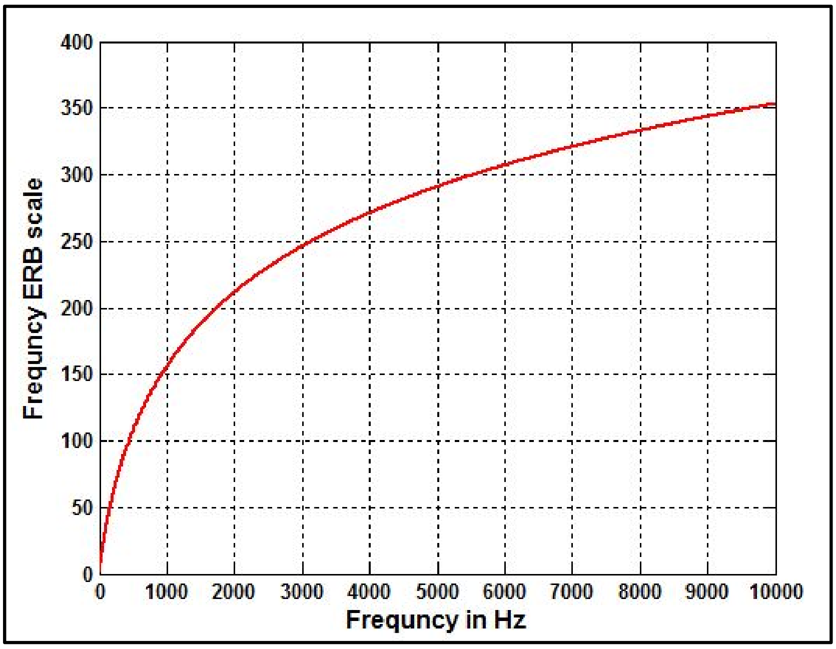


Figure 2.9: Equivalent rectangular bandwidth (ERB) scale

The GFCCs play a major role in speaker recognition under mismatch, since they are more robust than other features in noisy environments. A variety of research studies deal with using GFCC as features in speaker recognition. Shao et al. (2007) and Shao and Wang (2008) proposed this kind of feature for speaker identification (SI) under additive noise conditions. The evaluation of speaker identification based on GFCC is compared with the performance with both MFCC baseline and ESTI Advanced front-end feature extraction (ETSI-AFE features) based system(ETSI, 2007) under clean and a wide range of signal to noise (SNR) conditions. The systematic evaluations illustrate speaker identification based on GFCC features outperform systems based on both MFCC and MFCC ETSI-AFE features over a wide range of SNR and different types of noise.

A recent in-depth study (Zhao and Wang, 2013) confirmed the intrinsic noise robustness of GFCC relative to MFCC. In addition, this study includes further details about strong points that make GFCC more robust to additive noise compared with MFCC by carefully examining all differences between the two features using the speaker identification system. The authors also indicated that use of a cubic root rectification in the GFCCs instead of a logarithmic process in the MFCCs might contribute to their robustness because this cubic root operation makes features scale variant (energy level independent), which helps to maintain this information, while the log operations in MFCC features do not encode this information. This study also suggests some modification on MFCC extraction to improve their robustness to noisy environments. A similar study was documented in Moinuddin and Kanthi (2014) to compare a GFCC-GMM speaker identification based model with a baseline model using TIMIT and NTIMIT databases. This comparison also included using logarithmic and cubic root operations in extraction for both kinds of features (MFCC, GFCC) and the impact of this on the robustness of speaker identification in clean and noisy environments. However, this work shows that replacing the logarithm with the cubic root in both types of features extraction decreases the accuracy of identification in both clean and noisy speech. Das and Bhattacharjee (2014) investigate using GFCC with Joint Factor Analysis (JFA) based modelling to handle environmental noise in a text independent speaker verification (SV) system. This system is compared with a MFCC-JFA system and their performance is evaluated depending on error equal rate (EER). The results of these investigations show GFCC-JFA based systems outperform a MFCC-JFA based system especially in lower SNR (0dB and 5dB), while a MFCC-JLA system works better in higher SNR.

Based on aforementioned literature. GFCC feature spaces will be adopted in this research since they represent one of noise robust features in the field of speaker recognition. More details on GFCC will be presented in Chapter 6.

2.3.F Some of the Alternative Features

As has been seen in the last section, most of the features used in speaker recognition are based on MFCC and LPCC. In this section, some of the other features that have been used in this field have been described. Some of these features are generated by wavelet filter banks, instantaneous frequencies, and the others are variations of Cepstral coefficients with different frequency scaling (Beigi, 2012, Beigi, 2011).

Wavelet Octave Coefficients of Residues (WOCOR), are also known as vocal source features. These features are extracted by making a pitch-synchronous wavelet transform of the linear predictive (LP) residual signal (Chan et al., 2007). These kinds of features are used as complementary features for MFCC to produce a better result in speaker recognition (Nengheng, 2005, Wang et al., 2011) since the former are derived in trying to model vocal source characteristics, and the latter is related to the shape of the vocal tract (Beigi, 2011). Linear frequency cepstral coefficients (LFCC) and exponential frequency Cepstral coefficients (EFCC) were used by Xing and Hansen (2011) who used an unvoiced consonant detector to split the frames which contain such phones and used LFCC and EFCC for these frames. These features were then used to train up a GMM-based speaker identification system. Mel-frequency Discrete Wavelet Coefficients (MFDWCs) were used by Tufekci and Gurbuz (2005) in a speaker verification system, and they reported some improvement in the verification system relative to an MFCC baseline system in aggressively noisy environments. MFDWCs are computed by the same procedure as used for MFCC but using Discrete Wavelet Transform (DWT) instead of discrete cosine transform (DCT). Another noise robust feature set for speaker verification systems was proposed by Meriem et al. (2017). The proposed feature set, which is called Multitaper Gammatone Cepstral Coefficients (MGCC), is based on combining the advantage of the low variance multitaper short term spectral estimators with the noise robustness of the auditory Gammatone filterbanks. The experimental result using MGCC features in speaker verification outperformed both MFCC and GFCC under different noisy environments.

2.4 Speaker Modelling and Classification Approaches

As mentioned before, after a sequence of feature vectors are extracted from the speech signal of an anonymous speaker, the function of a speaker recognition system is to check whether that feature vector belongs to one of the registered speakers. To achieve that, speaker modelling is responsible for generating reference model(s) for each registered speaker during the enrolment phase depending on the features extracted from the speech signal. During the classification phase (recognition phase) the test utterance (recognition utterance) from an unknown speaker is compared with a reference model of the claimed speaker (or with all speaker models in case of speaker identification) to get the matching score, which indicates the degree of matching.

In general, these approaches can be classified into **generative** and **discriminative** models. Generative models try to capture the whole underlying distribution, i.e., the class mean (centroid) and variation around that mean, of training data. In addition, this model is trained to best represent the entire distribution space of the training data generated from a particular class. The trained class model considers only matching data, discarding the distribution of the other classes (Nengheng, 2005). The generative models can also be classified into the template models (e.g. vector quantization (VQ) codebooks) and stochastic models, (e.g. Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM)).

Discriminative models, otherwise, do not need to model the entire distribution, but only the most discriminative areas of distributions. The aim of training a discriminative model is to minimise the classification errors for a set of training samples. Consequently, not only samples from the matching class but also those from all the rival classes are taken into account when training the discriminative model for each class. These models include Support Vector Machine (SVM) and Artificial Neural Network (ANN).

The models generated in a discriminative model require the training sample from both the target speaker and all the imposter speakers, and modelling only the boundary may cause discarding of some information from the client which may contain the boundary information between the target speaker and other imposters. As a result, the discriminative model may work poorly for these imposters. However, in Generative models, all the target information is retained, which makes the generative models more robust against these imposters. Furthermore, in discriminative models, if the reference models in the speaker recognition system are updated with some new speaker, all the reference models have to be retrained (re-enrolled). The generative models, on the other hand, do not require re-training because each target model is trained independently (Nengheng, 2005). This makes the generative model more appropriate for speaker recognition compared with discriminative models; Table 2-1 illustrates the comparison between the two types. In the next section, we present a general review of the most commonly used modelling techniques in speaker recognition. Cerva et al. (2009) made a comparison among three different speaker recognition approaches deployed in broadcast using a novel modelling system. In this work, the authors focus mainly on the effect of generative and discriminative approaches on speaker recognition frameworks. For a generative approach, they used a Gaussian Mixture Model-Universal Background Model (GMM-UBM) SR based system; while discriminative approaches were represented by Gaussian Mixture Model-Support Vector Machine (GMM-SVM) and Maximum Likelihood Linear Regression- Support Vector Machine based system (MLLR-SVM). The experiments were performed on a very limited amount of data in both enrolment and recognition phases. The results illustrate that the system based on GMM-UBM outperforms both discriminative based systems but a higher computational cost is required. A comparative study to evaluate the performance of text-independent speaker verification based on different classifier approaches in noisy environments had been made by Sukhwal and Kumar (2015). These classifier approaches included Euclidian distance, Back Propagation Neural Network (BPNN), and Self Organization Map (SOM). Furthermore, this study involved the usage modified MFCC, which included Blackman windowing instead of hamming window. When the performance of these classifiers is compared, the results show SOM Speaker verification system based on Blackman window has given the best results in testing phase.

Table 2-1: Comparison of Generative and Discriminative modelling approaches

Generative Models	Discriminative Models
Attempt to include the whole underlying distribution	Attempt to include only the most discriminative areas of distributions by modelling the boundary between classes.
Generate a full model of enrolled speaker speech independent of the availability of imposter utterances. Therefore, the generating models become more robust to imposter attack.	Modelling the boundary may ignore some information from the target which may contain the necessary information between the target speaker and other imposters. Thus, the created model is vulnerable to imposters not present during the training phase.
Since the building model is based on clustering feature vectors, the variation in length of training data did not affect the process.	Cannot handle the variation length of data in both enrolment and recognition phase
Since the generative models deal with the whole underlying distribution, a full model requires high storage capacity.	Since discriminative models deal with the boundary of class only, they do not need high storage capacity.
Updating the speaker recognition with new speakers requires only new models to be trained independently from reference speaker models.	Updating the speaker recognition with new speakers requires all reference models to be re-trained.

2.4.A Vector Quantization (VQ)

Vector quantization represents one of the simplest and earliest template classification models for speaker recognition (Soong et al., 1985, Lin and Shuxun, 2006). This technique, which is also known as a centroid model, includes separating the features vectors into a set of non-overlapping clusters which individually represent different acoustic classes. Each of these clusters is represented by a code vector which is the centroid of that cluster (mean vector). Therefore, in this approach, a collection of centroid vectors represents a speaker reference model which is referred to as a codebook (Pillay, 2010). These codebooks are generated based on one of the standard clustering algorithms such as the Lindw-Buzo-Gray (LBG) algorithm (Soong et al., 1985) or the k-means algorithm (MacQueen, 1967). VQ approaches offer an efficient way of decreasing data storage requirements while preserving the fundamental aspect of the original distribution (Soong et al., 1985). During the classification phase, the distance of each of the extracted features vectors of the recognition utterance to its closest codebook vector is accumulated to produce an utterance score (Pillay, 2010).

$$\boldsymbol{score} = \sum_{j=1}^{M} \min_{x_i' \in C} \boldsymbol{dis}(\boldsymbol{x_j}, \boldsymbol{x_i'}) \qquad \textbf{(2.1)}$$

where M is length of input vector, C= $\{x_i' | i=1, 2,..,K\}$ the VQ codebook for each enrolled speaker. The main drawback of VQ classification is ignoring the probability that a specific enrolment vector may also belong to another cluster (Jayanna and Prasanna SR, 2009).

In earlier research Soong et al. (1985) applied an algorithm named LBG for generating speaker-based vector quantization (VQ) codebooks for speaker recognition. They demonstrated that a larger codebook and larger recognition data can give improved recognition performance. Moreover, their study suggests that Codebook of Vector Quantization can be updated from various recording conditions and intra-speaker variations. (Chatzis et al., 1999, Lin and Shuxun, 2006) proposed using fuzzy vector quantization as an alternative to VQ as a classifier in speaker recognition. FVQ gives better performance than VQ, since each feature vector in FVQ has an association with all clusters instead of being related to only one, as in VQ. Chatzis et al. (1999) and Lin and Shuxun (2006) proposed using fuzzy vector quantization as an alternative to VQ as a classifier in speaker recognition.

2.4.B Support Vector Machine (SVM)

One of the discriminative binary classifiers adopted in speaker verification is the Support Vector Machine (SVM), which includes modelling the linear boundary between two classes as a separating hyperplane (Burges, 1998, Beigi, 2011). In speaker verification, the first class consists of target speaker enrollment vectors (labelled +1) and the second class consists of training vectors from a huge number of background features. (labelled -1) (Kinnunen and Li, 2010). In addition, SVM can be used to learn non-linear boundary regions between samples by mapping the input samples into higher dimensional space. This can be done by using a kernel function (Pillay, 2010). Depending on these labelled training vectors, SVM is responsible for finding a splitting hyperplane that maximises the margin of separation between two classes. In the recognition phase, a classification score is then obtained by evaluating the distance of the recognition sample in relation to the hyperplane.

Campbell et al. (2006a) proposed using SVM for speaker and language recognition. The key part of their technique is the use of a novel derived sequence kernel (named the Generalised Linear Discrepant Sequence GLDS) that compares sequences of feature vectors and generates the measure of similarity. The GLDS kernel was shown to be very efficient and easy to use with SVM. This study shows that the result produced by using SVM is similar to those obtained from the GMM baseline system in error equal rate (EER) and give slightly reduced performance in detection error trade-off (DET). Furthermore, the authors suggest that the combination of SVM and GMM will achieve good performance since the SVM was provided complementary scoring information. Other researchers also investigated the effects of the combination of SVM and GMM to improve the performance of speaker recognition, and they reported that this combination yielded better recognition performance than individual systems (Campbell et al., 2006b, You et al., 2009).

2.4.C Artificial Neural Network

An Artificial Neural Network (ANN) is a discriminative model that is widely used in speaker recognition (Wahab et al., 2005, zohra Chelali et al., 2011). One of the main advantages of ANN is that the feature extraction and speaker modelling can be applied to a single network, enabling joint optimisation of the (speaker-dependent) feature extractor and the speaker model (Kinnunen and Li, 2010, Heck et al., 2000). This approach uses enormous parallel networks of many densely interconnected computational units known as neurones which are analogous to the neurones that exist in the human central nervous system (Beigi, 2011). Each neurone is responsible for sums a number of weighted inputs and passes the output through a non-linear activation function. Although there are many different kinds of ANN, the Multi-layer Perceptron (MLP) has been a commonly used architecture for speaker recognition. An MLP is made up of a network (multi-layers) of simple neurones which are known as perceptrons (Figure 2.10). The main idea of the MLP is based on a two stage process: first, compute a linear weighted sum of its input connections. And second, a non-linear activation function is applied in order to calculate the output of the neurone. An MLP with a non-linear activation function can estimate any non-linear mapping between input and output if given a sufficiently large number of neurones in the hidden layer. In speaker verification, an MLP has only one output neurone, since the goal, in this case, is to obtain a score over all the frames of the given recognition utterance (Pillay, 2010). Wahab et al. (2005) used an MLP neural network and Generic Self-organizing Fuzzy Neural Network (GenSoFNN) with extracted hidden features as an input for this network for a speaker verification system. The experiment was conducted on 10 speakers consisting of 6 males and 4 females recorded in a quiet room in a Digital system lab by using a digital tape recorder. Their experimental results showed the ability of both systems to verify speakers with high accuracy. Furthermore, the authors mentioned that the MLP can achieve high accuracy of verification with shorter training and testing time if it is applied to online speaker verification purposes.

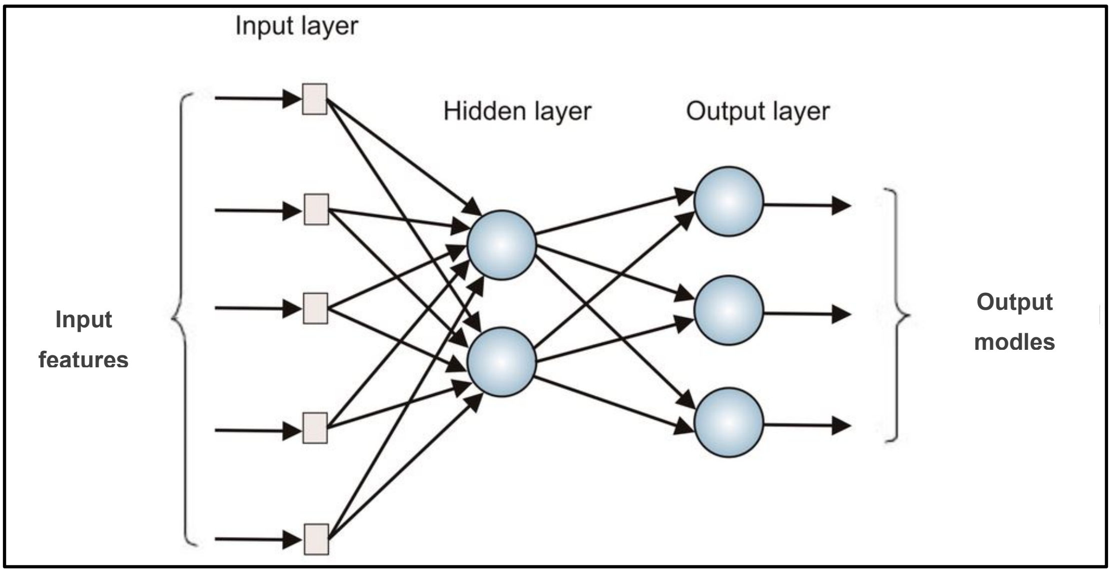


Figure 2.10: Multilayer perceptron (MLP) structure(de Araújo et al., 2013)

2.4.D Hidden Markov model (HMM)

Hidden Markov Model (HMM) is also commonly used in speaker recognition, especially for text dependent and text-prompted speaker verification, where a whole phrase is matched. HMM has the capability to model the temporal variations between the various acoustic classes (Pillay, 2010). HMM models are first-order discrete time series with some unobserved (hidden) information known as states. (Beigi, 2011). In the field of speaker recognition, each state may be referred to phones or larger units of speech. Through discrete time, the state of the HMM system is changed according to a set of probabilities related to it. The output from each current state is emitted after each transition. Although these outputs can be observed, the connected states are hidden and can only be deduced from outputs. This information between acoustic classes is encoded by moving from state to state along the allowed transition (Pillay, 2010). The amount of time consumed in the state represents the variability in speaking which depends on the training data. Matsui and Furui (1994) made a comparison between a vector quantization based texts–independent speaker identification system and a discrete/continuous ergodic HMMs based one. The experiments show that the continuous ergodic HMMs based system has the same robustness as the system that used VQ for variation of utterances. The authors also mentioned that the robustness of continuous ergodic HMMs based systems are restricted with the availability of sufficient data while VQ based systems show greater robustness when the amount of data is limited. On the other hand, discrete ergodic based systems show less robustness than the two other systems for variant utterances.

Hidden Markov Models are also used for storing the voiceprint of the speaker uniquely with a method called Hidden Markov model toolkit (MTK). This method helps to keep voice prints from being stolen by any other person since the raw voice is not stored. That means the imposter cannot access the system by providing a recorded copy of the speaker’s voice (Jayamaha et al., 2008).

2.4.E State of the Art Speaker Recognition Modelling

In the previous section, we mentioned some classical modelling approaches that are used in speaker recognition. In addition, some benefits and drawbacks of these approaches according to their classification (Generative and Discriminative models) were illustrated. This section will focus on the most common modern modelling that is used in speaker recognition systems. Some of these models represent a hybrid between different models such as the GMM-SVM combination(Campbell et al., 2006b) and some of them represent developments of other modelling techniques, such as GMM-UBM (Reynolds et al., 2000) and i-vector modelling(Senoussaoui et al., 2010).

A. Gaussian Mixture Model (GMM) and Universal Background Model (GMM-UBM)

The Gaussian Mixture model represents an extension of VQ techniques and has recently become the most frequent reference model approach in speaker recognition (Kinnunen and Li, 2010). Unlike VQ, the clusters in GMM are permitted to overlap with each other. GMM is a stochastic approach that expresses the probability density function of a random variable in terms of a weighted calculation of the sum of Gaussian components (mixture). These components are the mean, covariance statics of mixture, and weight associated with each of them which together represent each model of speaker in GMM (Beigi, 2011, Sinith et al., 2010). These models' parameters are generally computed by applying the iterative Expectation Maximisation algorithm EM (Dempster et al., 1977). Reynolds (1995) and Reynolds and Rose (1995) proposed using this model in speaker recognition, and they demonstrated the strength of using this model in text-independent speaker identification compared with other models like VQ, uni-modal Gaussian, and radial basis function. One disadvantage of GMM is the requirement for sufficient data to make a reference model for the speaker. To tackle this problem a developed model has been proposed by Reynolds et al. (2000), namely the Universal Background Model (UBM). The primary aim of using the UBM is to collect utterances from a large population of speakers, which are pooled to train the UBM using EM algorithms as speaker-independent models. After that, speaker dependent model is adapted from UBM by using Maximum a Posteriori (MAP) (Kinnunen and Li, 2010, Beigi, 2011). In the recognition phase (or classification phase), the recognition data are matched with the claimed speaker model or register speaker models using maximum likelihood rule. Over the last decades, GMM-UBM has become the most popular modelling approach in text-independent speaker recognition (Nengheng, 2005, Pillay, 2010, Pinheiro et al., 2013).

The significance of GMM-UBM in the speaker recognition field comes from the relationship between the reference speaker models and the UBM, since this provides a very efficient method for measuring log-likelihood ratio (LLR) to obtain a recognition score. Moreover, GMM-UBM represents one of the better modelling approaches when the speech data used in enrolled and recognition phases was limited (Bao and Juan, 2012). GMM/GMM-UBM have two important features: First, GMM have shown an efficient work with limited or small dataset than other modelling approaches Second, the speech signal is based on stochastic process, and to create an efficient model for the speakers the modelling are used should contain statistical analysis. That means the Gaussian built in by GMM become more efficient for the speech signal. For these reasons, GMM are adopted in this work. Additional details about GMM and GMM-UBM modelling are available in Chapter 4 and Chapter 5.

B. Supervector approaches

Recent approaches, first proposed by Campbell et al. (2006b), use a GMM-SVM combination technique to tackle utterances that have different durations and different numbers of feature vectors (Pillay, 2010, Kinnunen and Li, 2010). The term supervector refers to merging smaller dimensional vectors in a large fixed dimensional vector and then using this vector in SVM training, since SVM is proposed for speaker recognition to deal with each utterance using fixed dimension vector (Wan and Renals, 2005). In GMM-UBM modelling, the idea of a supervector is based on using a combination of the means from adapted enrolled speakers GMM and a large set of imposter GMMs to train the SVM speaker model. These means of GMM are obtained from an adaptation of UBM while the weight and covariance can be shared between speakers. Thus, it can also represent this adapted speaker model in GMM-UBM as a supervector. During the recognition phase, a supervector of means is extracted from the recognition model and then matched with the target's SVM model to obtain a classification score. Different kinds of kernel are used in the literature based on GMM and SVM supervectors; Campbell et al. (2006a) proposed one of the simplest SVM supervectors: a Generalized Linear Discriminant Sequence (GLDS) kernel, which generates supervector explicit mapping into kernel feature space by means of polynomial expansion (Campbell et al., 2002). McLaren et al. (2007) proposed Background Scaling Linear (BSL) as a GMM supervector approach where input supervectors are normalised, such that, they have unit variance in each dimension depending on the huge number of static background supervectors. A Gaussian supervector SVM (GSV) kernel, on the other hand (Campbell et al., 2006b, Campbell et al., 2006c), is extracted from a Gaussian supervector by bounding the kullback-leiber (KL) measure between two GMMs.

C. The i-vector model (total variability space)

The derived supervectors for specific speakers from different training utterances may not be the same since these training utterances may be collected from different handsets (different channels). In this case, Channel compensation becomes very important, to make sure that recognition data obtained from different handsets (than those used for training data) can be correctly scored against the speaker models (Kinnunen and Li, 2010). Therefore for channel compensation Kenny (2005) proposed that channel variability can be modelled explicitly using an approach called Joint Factor Analysis (JFA). This approach assumes that the generated JFA model represents the variability of a Gaussian supervector as a linear combination of speaker and channel components. Then for a given enrollment sample, the supervector is decomposed into two statistically independent components, one for a speaker-dependent supervector and second for a channel-dependent supervector. Dehak (2009) discovered that the channel components of JFA may also contain embedded information about the speaker. So Dehak et al. (2010, 2011) suggest that is not necessary to make a distinction between the channel component and speaker component in GMM supervector space. They also proposed a new technique in which the new space can contain the speaker and channel components simultaneously. This new space is defined by an Eigenvoice matrix and they called it *total variability space* (it was known later as **i-vectors** model, shortened to Identity model). In i-vectors the dimensionality is normally decreased from a GMM supervector before modelling by using a generative factor analysis technique called linear discriminant analysis (LDA) or recently Probabilistic LDA (PLDA) in order to suppress the direction of channel factor, thus increasing the discrimination between speaker subspaces. These vectors are then mean normalised using Within-class covariance normalisation (WCCN) (Sadjadi et al., 2013, Beigi, 2012). These i-vectors are also known as intermediate vectors since they are much smaller than GMM supervectors but at the same time, they are larger than the underlying feature vectors. Furthermore, unlike GMM-UBM, both enrolment and recognition models should be represented as i- vectors when this type of approach is applied. Many recent research studies that have used i-vector approach for speaker recognition (Glembek et al., 2011) propose simplifications to the i-vector formulation to increase the speed of extracting vectors and minimise their memory usage. The authors also investigate using principal components analysis (PCA) and Hetero Linear Discriminant Analysis (HLDA) to achieve of the components of the Gaussian mixture orthogonally.

D. Deep Neural Network (DNNs)

Deep neural networks represent one the newest approaches in speaker recognition fields (Hinton et al., 2012), particularly after the impressive results obtained from using DNNs for automatic speech recognition. DNNs is essentially a multi-layer perceptron (MLP) with more than two hidden layers that typically uses random initialization and stochastic gradient decent to initialize and optimize the weights (Hinton et al., 2012). In speaker recognition (and other speech applications), the input to DNN is a stacked set of cepstral features (e.g, MFCC, LPCC) extracted from frames of speech, while the output is a prediction of the posterior probability of the target class for the current input frame (Richardson et al., 2015). Two different methods are used to apply DNNs in Speaker Recognition. The first one is called the “Direct” method, in which a DNN is employed as classifier for the intended recognition task to discriminate directly between the speakers for SR. The second one is called the “Indirect” method, which employs a DNN, possibly trained for different purposes, to extract data that is then used to train another classifier algorithm for the intended recognition task (Richardson et al., 2015). Lei et al. (2014) proposed a new framework for speaker recognition in which extraction of sufficient statistics for i-vector model is derived by DNN instead of standard GMM-UBM. The proposed framework shows that the DNN approach significantly improved the i-vector speaker recognition system when compared with GMM-UBM.

2.5 Score Classification and System Performance Evaluation

For closed set Speaker Identification, to measure the similarity between recognition speech and reference model, the reference model (and sometimes models) that have the best match to the input feature vector of the unknown speaker are returned. However, in open-set speaker identification, the input feature vector is classified as an imposter speaker if the highest matching score is lower than a preset threshold (Nengheng, 2005).

In speaker verification, on the other hand, the decision is whether the input features vector belongs to the claimed speaker or it belongs to an imposter based on matching score. This matching score results from matching this input vector with a reference model for the claimed speaker. So, the speaker verification system should make a decision between two hypotheses: the input signal comes from the claimed speaker, H_1, or from an imposter H_0. Based on the stochastic model (GMM, HMM), the choice can be made depending on the expected log-likelihood ratio (Beigi, 2011).

$$\boldsymbol{D = log\frac{P(x_i|H_1)}{P(x_i|H_0)} = logP(x_i|H_1) - logP(x_i|H_0) >^? \beta} \qquad \textbf{(2.2)}$$

where $P(x_i|H_1)$ is the probability of observation x_i being produced by the target speaker, and $P(x_i|H_0)$ is the probability of the observation being produced by an imposter speaker. β is the threshold that can be determined by the difference between false accept rate (FAR)(also known as False Negative Rate FNR or Miss) and false rejection rate (FRR)(also known as False Positive Rate or False Alarm) (Nengheng, 2005, Beigi, 2011).

Theoretically, the performance of the perfect speaker verification system is capable of verifying all speech signals that belong to registered speakers and declining those that are generated from imposters. However, in real life, this does not always happen, since there are many environmental and speaker factors that adversely affect the performance of the system (as seen in the previous chapter). As a result, there are four possible decisions that are usually made, as shown in Figure 2.11:

- Accept input signal that comes from registered speaker (Correct).
- Accept input signal that comes from imposter speaker (Error).
- Reject input signal that comes from imposter speaker (Correct).
- Reject input signal that comes from registered speaker (Error).

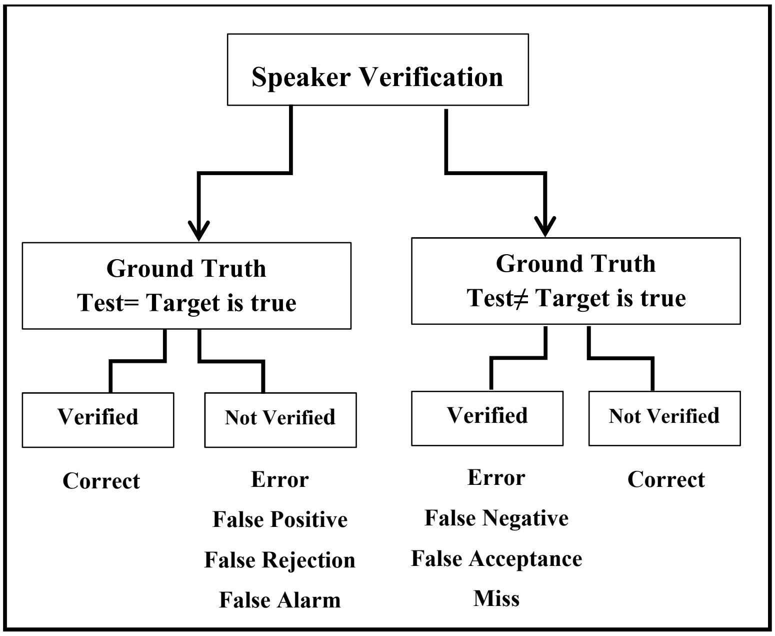


Figure 2.11: Speaker Verification Decisions

So, according to Figure 2.11 and depending on statistical hypothesis testing, there are two types of errors (Beigi, 2011, Rao and Sarkar, 2014):

1. False Acceptance Rate (FAR) also known as false negative or miss probability , is defined as the number of verified identities for which the recognition speaker was different from the registered speaker normalised against total number of acceptances, such that:

$$\boldsymbol{FAR} = \frac{\boldsymbol{Number\ of\ FAs}}{\boldsymbol{Number\ of\ Imposter\ Trial}} \qquad \textbf{(2.3)}$$

2. False Rejection Rate (FRR), also known as false positive or false alarm ,is defined as the number of identities which were not verified for which the recognition speaker was the same as the registered speaker normalised against the total number of rejections, such that:

$$\boldsymbol{FRR} = \frac{\boldsymbol{Number\ of\ FRs}}{\boldsymbol{Number\ of\ Target\ \ Trial}} \qquad \textbf{(2.4)}$$

In order to evaluate the system performance into a single metric, the National Institute of Standard and Technology, NIST (1998) proposed Error Equal Rate (EER) to measure the accuracy of a speaker recognition system. ERR represents one of the most commonly used methods to measure the performance not only in speaker recognition fields but also in most biometric systems, and refers to the point where the chances of a false rejection rate (FRR) and false acceptance rate (FAR) are equal (Beigi, 2011, Rao and Sarkar, 2014). Consequently, the EER offers the means for producing a convenient percentage error.

2.5.A Detection Cost Function (DCF)

The Detection Cost Function (DCF), which has been proposed by Martin and Przybocki (2000) is measured as the weighted sum of FRR and FAR as follows:

$$\boldsymbol{DCF} \triangleq (\boldsymbol{C_{Miss}}\, \boldsymbol{p(Miss|Target)P(Target)}) + \\ (\boldsymbol{C_{False\ Alarm}}\, \boldsymbol{p(FalseAlarm|NonTarget)P(NonTarget)}) \qquad \textbf{(2.5)}$$

where C_{Miss} is a cost of Miss (FAR) =10,

$C_{falseAlarm}$ is a cost of False Alarm (FRR)=0.1,

P (target) is a probability of Target =0.01

P(NonTarget) is a probability if non –target =1-P(Target)=0.99

The minimum values of DCF with EER are used as metrics for accuracy evaluation of speaker recognition (Rao and Sarkar, 2014, Przybocki and Martin, 2004).

2.5.B Receiver Operating Characteristics (ROC) Curve

A Receiver operating characteristics (ROC) curve is a way of graphically representing the trade-off between FAR and FRR (Martin et al., 1997). ROC curves (as mentioned in Figure 2.12) normally plot the FAR as X-axis and FRR as Y-axis. The space under the curve is a measure of the accuracy of a speaker recognition system. ROC curves can usually be plotted in a linear scale; but sometimes to avoid the resolution problems, they may be plotted in log scales.

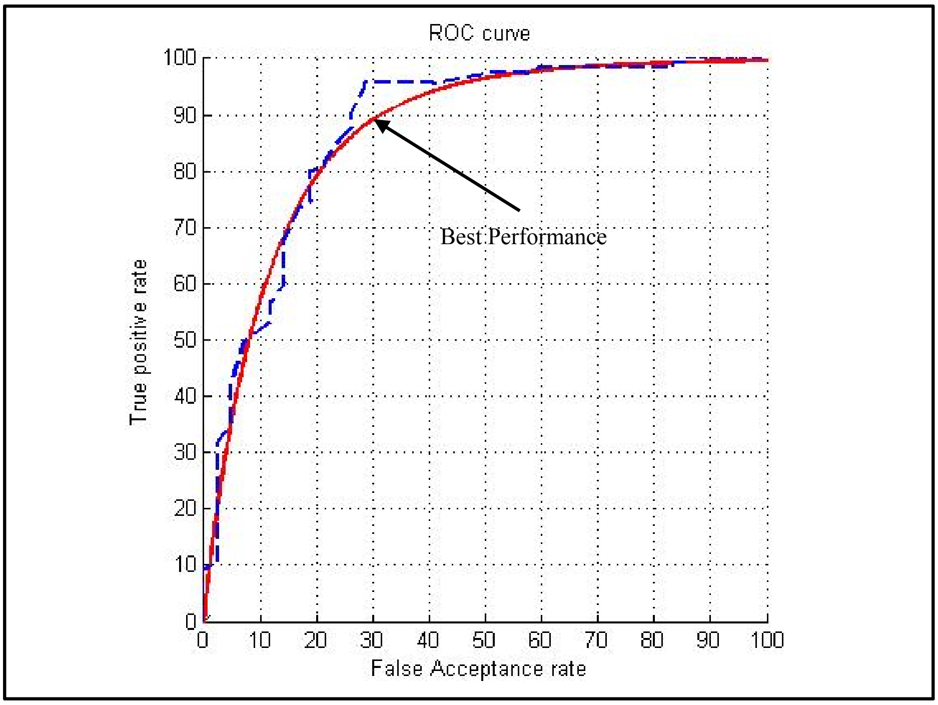


Figure 2.12: Illustration of Receiver Operating Characteristics (ROC)

2.5.C Detection Error Trade-Off (DET) Curve

Another graphical method was proposed by Martin et al. (1996) to evaluate the performance of speaker verification. In contrast to ROC, DET plots the Miss Probability (FAR) as Y-axis and False Alarm (FRR) as X-axis, and both are represented in percentage form. Moreover, DET curves are based mainly on log scales, which makes these curves more linear compared with

ROC (which is more concave) (Przybocki and Martin, 2004, Martin et al., 1997).The linear property makes the DET curve more flexible to visualise relative differences between various classifiers. Figure 2.13 shows the DET example for different verifications. Since NIST proposed it in 1996, DET has become most widely used to evaluate speaker recognition systems. Martin et al. (1997) made a deep comparison between ROC and DET curves. The authors mentioned many advantages that make DET better than ROC curves. Since 1996, another study has been made by NIST on the evaluation metrics of speaker recognition systems (Martin and Greenberg, 2010). In this study, EER with a DET graph is adopted to evaluate the speaker recognition system. The main reasons for adopting these two evaluation metrics are that ERR are broadly used in the literature to evaluate the performance of Speaker recognition system. While the DET graph represents the best graphical demonstration of both FPR and FNR in recognition system. Furthermore, the DET graph represents the graphical representation for Detection Cost Function (DCF).

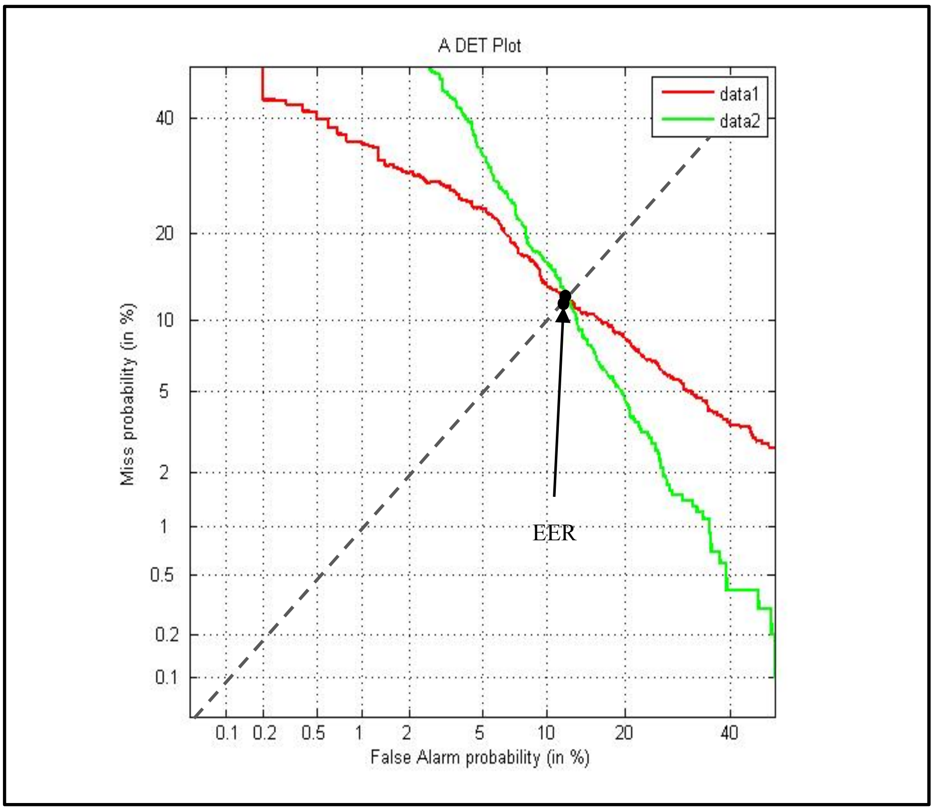


Figure 2.13: Illustration of Detection Error Trade-off (DET)

2.5.D Performance Evaluation of Open-Set Speaker Identification

As mentioned before, Open-set identification involves two stages: close-set identification and verification. To evaluate the performance of this system, the verification stage is evaluated by using the same EER discussed before for speaker verification. In this case, the verification performance is expressed in term of Open-Set Identification Error Equal Rate (OSI-EER), while the identification stage is measured in terms of Identification Error Rate (IER), which is evaluated as follows (Pillay, 2010):

$$\boldsymbol{IER = \frac{No.of\ Incorrectly\ idetification\ clint\ speakers}{No.of\ clint\ Trails} \times 100\%} \qquad \textbf{(2.6)}$$

2.6 Noise Robustness approaches in Speaker Recognition

The main of aim of this research study is to handle the channel mismatch that occurs as a result of environmental noises. In this section, we focus mainly on the approaches that deal with the impact of noise on the performance of speaker recognition and some of the main studies that deal with this problem. First, we talk about environmental noise and the effect of noise on the speech signal. Then we move to the main techniques and studies that tackle this challenge.

2.6.A Knowing the Environmental Noise enemy

In order to understand the nature of the impact of the various noises on the accuracy of speaker recognition, we should first understand the characteristics of background noises. Figure 2.14 illustrates the long-term average spectrum of the five types of noises (Cafeteria speech Babble, inside the car, inside the train, street noise, and white noise) that have been used in this study.

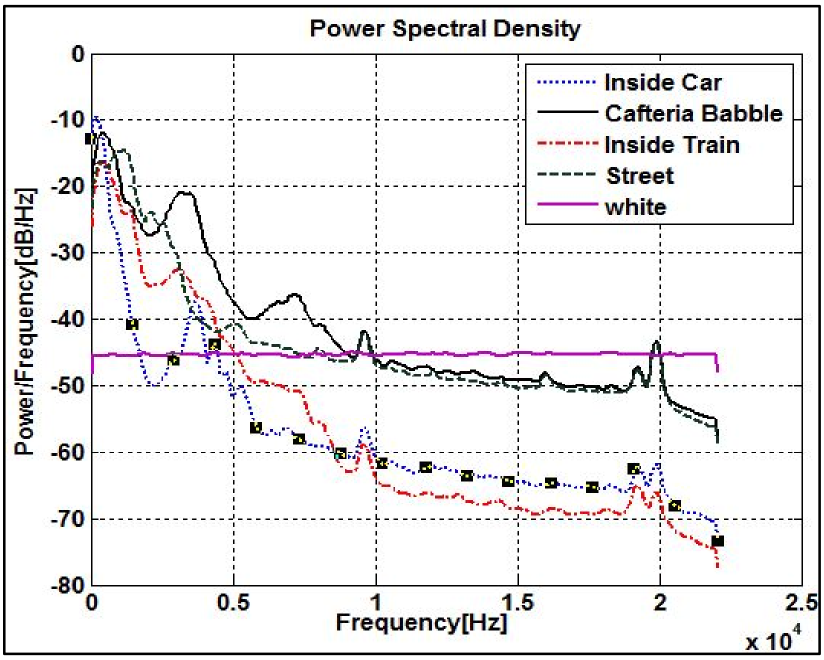


Figure 2.14: Power Spectral Density of Different types of noise

The first observation refers to the absence of uniformity in the spectrum, which gives a different identity for each kind of noise. In addition, noises can be generally categorised into ***stationary or Wide Sense Stationary (WSS).*** The noise signal $d(n)$ is known as wide sense stationary if the following three conditions are provided (Gruber, 1997) :

1. The mean of the signal is constant, $m_d(n) = m_d$.
2. The autocorrelation $r_d(k,l)$ depends only the difference,$k-l$, where $d(k)$ and $d(l)$ are random variables, $k-l$ is separating the two-random variable in time (also known as lag).
3. The variance of the process if finite, $c_d(0) < \infty$.

On the other hand, ***non-stationary*** in which the spectral characteristics are continuously changing, such as in cafeteria babble noise in Figure 2.14. This makes suppression of this kind of noise more difficult than suppressing stationary noise. Furthermore, the spectral distribution of these kinds of noise gives them a different impact on the accuracy of speaker recognition. For instance, Cafeteria Babble noise may represent one of the hardest types of noise to handle within speaker recognition, since it occupies a wider frequency range. Furthermore, it has very similar spectral characteristics to a speech from the target speaker (Loizou, 2013, Mandasari et al., 2012, Beritelli, 2008).

The power level of difference between the signal and additive noise is known as Signal to noise ratio (SNR). SNR is usually measured in Decibels (dB) such that if the ratio of the speech signal is equal to the ratio of additive noise then SNR=0dB. So to measure SNR for any noisy speech signal using logarithmic Decibel scale:

$$\boldsymbol{SNR_{db} = 10\log_{10}(\frac{P_{signal}}{P_{noise}})} \qquad \textbf{(2.7)}$$

where P_{signal}, P_{noise} represent the average power of speech signal and average power of additive noise respectively.

2.6.B Noise suppression techniques for Speaker Recognition

Different kinds of techniques have been proposed in the literature to deal with the effect of Background noise on speaker recognition. These techniques can be categorised depending on the stage at which they are applied. In this section, we will focus on some these techniques based on each stage and some of the literature in each stage:

1.Pre-processing stage techniques

These types of techniques include those methods that are applied at speech signal or acoustical level. The main function of these techniques is to clean the speech signal from additive noise by improving the SNR of these signal. These techniques usually produce an estimation of the enhanced short-time speech spectra by cleaning speech from the noise components (Pillay, 2010). These techniques can also be divided into three main types: spectral subtraction (Drygajlo and El-Maliki, 1998, Gustafsson et al., 2001), statistically based model (Cohen et al., 2006), and Subspace techniques (Jabloun and Champagne, 2003, Hu and Loizou, 2003).These kinds of techniques are more effective with stationary noise. Furthermore, these techniques can enhance speech quality but not speech intelligibility. More details on these techniques are given in Chapter 7.

One of the earlier research studies was done by Vizinho et al. (1999) using techniques to estimate the noise and signal to noise ratio SNR, which enabled them to remove the noise portion from the spectrogram in order to obtain the reliable portion of the spectrogram of speech to perform speech recognition. The performance of this technique achieved up to 10 dB SNR on factory noise and 0 dB of car noise. This method is suitable when the noise is stationary or slowly varying. Drygajlo and El-Maliki (1998) investigated the effect of a combination of missing features theory and spectral subtraction on improving the performance of text-independence speaker verification in environmental noise. The spectral subtraction algorithm is used in this study as a simple missing features detector instead of an enhancement system. This combination significantly improved EER, compared with using traditional spectral subtraction as pre-processing speech enhancement. This study was only based on using speech data contaminated with white noise with limited SNR (between 9dB-15dB).

Different enhancement techniques had been applied to speaker identification systems by Ortega-Garcia and Gonzalez-Rodriguez (1996) to improve robustness identification accuracy against two kinds of stationary noises (white and Fan noises) with different SNRs. They made a broad analysis on speech enhancement techniques of a single channel (based on classical spectral subtraction), dual channel (based on adaptive noise cancelling), and multi-channel (using microphone arrays). The results of these experiments showed good recognition results especially for multi-channel speech enhancement in SNR greater than 5 dB. These experiments were applied only on stationary noises.

Ming et al. (2006) made a comparison between different speech enhancement techniques on the robustness of speaker verification depending on noisy samples and using the same features and model framework for characterising the speaker. These techniques included Wiener filtering, noise compensation, missing-feature technique, and universal compensation. Furthermore, they investigated the effect of combining these techniques to improve noise robustness. Limited enrollment data were used in the handheld-device database (26 male, 22 female) with involved realistic noise and transducer mismatch between enrollment and testing. The results of these experiments indicate that traditional noise filters and noise compensation provide very limited robustness to noise corruption. However, the combination method that was used in this experiment showed a good improvement in noise robustness.

2.Features Stage techniques

This type of techniques applies additional noise suppression algorithms to the signal feature domains (feature extraction stage approaches). These are mainly based on the assumption that in practice, the features representing the speech signal can be separated into two different parts. The first part is represented as the noisy speech part, which refers to the unreliable or missing features part, while the second part represents the reliable or present features. These techniques include missing feature theory (MFT) (Beigi, 2012, Lim et al., 2011, May et al., 2012), Relative Spectral method (RASTA) (Hermansky and Morgan, 1994), and Cepstral Mean Subtraction (CMS) (González-Rodríguez et al., 1996). The objectives of these kinds of the technique are to estimate or discover those missing unreliable features in order to reconstruct or ignore them during the recognition process (Rao and Sarkar, 2014). Different techniques have been proposed for this aim. Some of these techniques are based on prior estimation for noise regions in order to compensate them (Drygajlo and El-Maliki, 1998). Other types of these methods have proposed to completely discard the severely corrupted speech data segment when there is insufficient information about the noise, or it cannot be easily estimated. The recognition process then depends only on the portion of the speech signal that is not contaminated or at least contains a low ratio of noise (Seltzer et al., 2004, Besacier and Bonastre, 1998). Other features stage techniques deal with retaining only speech features vectors, which produce reliable scores for measuring the overall likelihood score to increase speaker recognition robustness under noisy environments. However, these kinds of techniques are suitable only when the speech signal is partially contaminated with noise. Furthermore, these methods should lead to removing some important speaker discriminative information. In addition, identifying new features that are more robust to noise (for example GFCC, and CFCC) are also classified as feature type techniques.

Seltzer et al. (2004) presented a new estimation mask using a Bayesian classifier to extract information from a noisy signal to determine the reliability of this information, instead of trying to estimate the noise portion. This estimation mask based Bayesian classifier shows a real improvement in recognition performance over a conventional noise estimation based mask, especially when the environment is unknown to the mask estimator. However, this kind of estimation masks compared with other traditional mask is more complicated and requires many more parameters to optimise. Pullella et al. (2008) proposed a combination of spectrogram mask estimation and dynamic feature selection to increase the accuracy of speaker identification in the presence of noise. The technique depends on the uneven distribution of speaker-specific information in the frequency domain, allowing the erasure of non-discriminative frequency sub-bands and the formation of a reduced feature set. Furthermore, multi-condition enrolments were used in the study to dynamically select the most discriminative features for a group of speakers given an estimate of the global (SNR) of the evaluation environments. Experimental results of that combination show a good performance in the system under stationary white noise. For non-stationary factory noise, spectral subtraction is unable to provide accurate mask estimation. Shin-Cheol et al. (2011) proposed a modification in Advanced Missing features theory mask (AMFT) to reduce the number of reliable spectral components of the noise signal. This modified mask was based on an estimated weighted function by computing the number of reliable spectral components. In the proposed mask, namely Hard mask MFT-8 (HMFT-8), they chose only eight elements out of ten spectral components in the features vector. The results of the new mask show a decreasing identification error rate compared with a conventional full band system and AMFT respectively. In addition, the proposed mask shows a reduction in computational complexity from 307 arithmetic and conditional operations in AMFT to just 41 in this mask. May et al. (2012) investigated the combination of the missing features recognition with the adaptation of a speaker model depending on a Universal Background Model (UBM). For missing features recognition, the missing data mask based on local signal to noise ratio (SNR) has been applied to identify whether the feature component is reliable or unreliable. Comparing this system to a GMM-Based recognizer, the addition of UBM was shown to be very useful in representing the spectral feature with the presence of highly non-stationary background noise. A robust speaker identification using a computational auditory scene (CASA) time-frequency mask was proposed by Zhao et al. (2012) to isolate speech from noise in the signal. The corrupted components that are obtained from the CASA mask are then either reconstructed or marginalised. Furthermore, they combined the two methods (reconstructed or marginalisation) in one system depending on the complementary advantage. They also propose using Gamma Tone Frequency Cepstral Coefficients (GFCC) for this system (Zhao and Wang, 2013) which achieve better performance in Speaker identification compared with a MFCC-based system. The proposed system achieved significant performance over conventional systems under a wide range of Signal to Noise Ratio (SNR). In another recent study, Zhao et al. (2014) investigated the combined effects of background noise and reverberation on the accuracy of Speaker identification system. They first suppressed the background noise through a CASA mask using a neural network. Then they dealt with reverberation depending on training models in a noise-free reverberant environment while they assumed a little information about the amount of reverberation in recognition data. The CASA mask was utilised for automatic speaker identification in two strategies: bounded marginalisation and direct masking. The outputs of the two strategies were then combined to make a Speaker Identification Final Decision. GFCC features were also used in this system.

3.Model stage techniques

These techniques are usually applied at the model stage to deal with the impact of noise by decreasing mismatch between the reference model for registered speaker and recognition materials such that they have the same noise characteristics (Pillay, 2010). These kinds of approach are either based on the estimation of the noise characteristics through the enrolled and/or recognition phases in order to decrease the mismatch or depend on generating multiple reference statistical models for the same registered speaker. These multiple models represent different noisy conditions for speech signals belonging to each registered speaker. Then, during the recognition phase, the reference model that gives the closest matches to the characteristics of the input speech signal (based on higher likelihood score) is selected for recognition. These techniques include parallel model combination (PMC) (Wong and Russell, 2001, Tufekci and Gurbuz, 2005) and multi-SNR method (Pullella et al., 2008, Yoshida et al., 2004, Matsui et al., 1996). Matsui et al. (1996) and Lit Ping and Russell (2001) investigated the utility of parallel model combination (PMC) for a noise robust HMM-based speaker recognition model. In these experiments, they combined a speaker HMM and noise source HMM into a noisy speaker HMM with a particular SNR. The approach used in this study generates several noise-added speakers HMM with different SNRs for each speaker, and the HMM that has the highest likelihood score for input speech signal is chosen. The results of these studies show significant improvement in both identification and verification accuracy. In a similar study, Tufekci and Gurbuz (2005) investigated using the PMC along with Mel-Frequency Discrete Wavelet Coefficients (MFDWCs) to exploit the advantages of both for improving the robustness of speaker verification in noisy environments. They evaluated the performance of this system over a MFCC based system also using PMC. The proposed system gives significant performance improvement over an MFCCs based system for -6dB and 0dB SNR values respectively. Ji et al. (2007) used the combination techniques in speaker recognition with the presence of noisy conditions. They used a method that combines multi-condition training models with missing features theory (MFT) to model the noise speech. Multi-conditional training was applied using simulated noisy data of simple noise characteristics and missing feature theory was used to refine the combination by discarding noise variation outside the given enrollment conditions. This technique was tested by using two kinds of database: a Noisy TIMIT database, collected by recording the data of different control noise conditions, and a handheld device database obtained in realistic noise. The experiments on both databases showed a significant improvement in the recognition system when compared with a baseline system using the same amount of information. Kim and Kim (2010) studied the effect of combining speech enhancement and multi-condition training in a speaker recognition system in a noisy environment where different groups of features are extracted in the speech enhancement domain and then the model of each speaker is trained depending on the mixture of speech enhancement-domain. They measured the mismatch between the training and testing data depending on a modified version of Kullback-Leiber distance (KLD), namely intra-KLD. In different background noise environments, the proposed system produced 20.29% error-rate compared with a speaker recognition system based on speech enhancement 43.51% and speaker recognition based on multi-conditional training 25.0%. Kheder et al. (2014) proposed an i-vector "de-noising" approach, called i-MAP, to deal with additive noise. i-MAP includes a full covariance Gaussian modelling of the clean i-vector and noise distribution in the i-vectors space, then introduces a technique to estimate a clean i-vector given the noisy version and the noise density function using a MAP approach. In an early study, Kheder et al. (2017) proposed an extension of the previous approach. This extension is based on building a noise distribution database in i-vector space in an off-line step to decrease the computational time imposed by the i-MAP de-noising technique. The results from the proposed extension approach show it is possible to achieve comparable results with sufficiently large noise distribution.

One of the major limitations with this type of system is that they rely on prior information of the noise sources. Therefore, it is difficult for this type of system to handle arbitrary environmental noise. In this thesis, therefore, we propose a new approach to SR systems used in real-world noisy environments. This technique re-trains the enrollment models of the speakers based on estimated the signal to noise ratio (SNR) and noise profile estimated from signals submitted to the system for recognition. This technique is called "**Noisy training on the fly for speaker recognition**".

4. Score stage techniques

The main aim of applying this kind of method is to alleviate the impact of differences in characteristics of the speech signal as a result of background noise by decreasing the overlap in the score distribution for the target speaker and an imposter (Pillay, 2010), since decisions depend mainly on these recognition scores and this score is channel dependent. Different normalisation techniques have been adopted for this purpose (Beigi, 2011). Some of these techniques are based on using the Bayesian equation for likelihood estimation. The matching score gained from a registered speaker model is normalised with the score obtained from a UBM (Ariyaeeinia and Sivakumaran, 1997) or a cohort of background speaker models (Zajic et al., 2007). Another group of techniques, which is mainly used with speaker verification, focuses on normalisation of the scores obtained from a recognition engine. These types of technique are known as score normalisation techniques. The most popular approaches in this type are Zero-normalization (Z-Norm) (Auckenthaler et al., 2000) and Test-normalization (T-Norm)(Reynolds, 1997).

2.7 Chapter Summary

This chapter has focused mainly on giving the background to speaker recognition systems and the major approaches that are employed in this field. This chapter has also concentrated on the majority of the literature that deals with the robustness of speaker recognition in noisy environments. First, we reviewed Human speech production. Then we covered the type of features and the most commonly used low-level features extraction algorithms in these fields, and different techniques used for generating reference speakers' models and classifying the recognition utterances. After that, this chapter focused on various approaches used to evaluate the robustness of speaker recognition systems. Finally, we covered some of the literature that deals with the degradation that occurs in the accuracy of speaker recognition systems due to additive environmental noise.

From this chapter, the following point can be summarised as follows:

- The Features of the speech signal can be classified for different type but low-level features and specifically the Cepstrum features are the most efficient in the field of speaker recognition, since this type of features is easy to extract, compared with other types of features. Furthermore, the major of study shows this give the best result for speaker recognition from other types of features.

- Although there are many types of Cepstrum features are proposed in the literature, the MFCC is still the most common features space in the field of speaker recognition and still used as baseline features in this filed. The main reason that the most experiments of this features in the literature still show an efficient performance for speaker recognition when evaluating in normal conditions.
- For the noisy environment, the auditory features (such as GFCC, CFCC) shows a significant improvement in accuracy of Speaker Recognition.
- Various types of modelling approaches are used in speaker recognition. However, the generative models (such as GMM and its extension GMM-UBM and i-vector) represent the modelling approach for speaker recognition. On the other hand, some research suggests usage of the combination between generative and discriminative modelling approaches (such as supervector) which outperform the performance of Speaker recognition.
- To evaluate the performance of speaker recognition, NIST proposed different types of metric based on two types of error false positive and false negative. The majority of literature uses Error Equal Rate (EER) as well as Detection Error Trade-off (DET) to evaluate speaker recognition application, since the first are used broadly in biometric fields, while the second demonstrates good graphical representation for speaker recognition performance.
- Various approaches are suggested to deal with noise effect on speaker recognition. These approaches can be classified according to the stage of speaker recognition are applied for, pre-processing stage, Feature stage, modelling stage, and score stage.

CHAPTER 3
SPEECH AND NOISE DATABASES

Chapter Overview

The previous chapter presented the background on speaker recognition and the most common approaches used in it. This chapter is intended to cover some details on the speech and noise databases that have been used in this research and will focus on the new database provided for speaker recognition, namely SALU-AC, followed by the methods used to generate the noisy speech data sets with various signal to noise ratios (SNRs) and the type of noises chosen for this research. Section 3.1 describes in detail the two speech databases that are adopted in this thesis and the types of noise that are used in this research in section 3.2. Finally, an explanation for mixing procedure to provide noisy speech with different Signal to Noise Ratios (SNRs) is given in 3.3.

3.1 Speech Databases

In this study, two types of speech data set have been adopted. These are TIMIT database (Garofolo et al., 1993) and SALU-AC which has been recorded by the researcher in at the University of Salford. In the next section an overview of TIMIT is given and then the details about SALU-AC database are described. The main aim of using two different databases in this research is to validate the results using a different type of speech sample recording in different environments. TMIT represent one of the most common database that is used broadly to evaluate speaker recognition. SALU-AC represents good examples to dealing with environmental noise separately from other adverse conditions (e.g. echo and reverberation) when mixing with different types of noise.

3.1.A TIMIT database

The TIMIT database represents one of the most commonly used speaker recognition corpora (Garofolo et al., 1993). It was developed for DARPA and distributed by the Linguistic Data Consortium (LDC) (Beigi, 2011).This speech database was collected at Texas Instruments (TI), transcribed at the Massachusetts Institute of Technology (MIT) and confirmed and prepared by The National Institute of Standards and Technology (NIST) (Pillay, 2010). The main purpose of designing this database was to evaluate and develop speech recognition systems and it became widely used to evaluate speaker recognition systems. The TIMIT database can be summarised as follows:

1. Speech samples in this database were obtained from 630 speakers (438 Male and 192 female speakers) of 8 various dialects of American English. In addition, all participant speakers are native speakers of American English.
2. Each speaker provided ten speech utterances for 10 specific sentences. Some of these sentences are fixed for all participants (e.g. “She had your dark suit in greasy wash water all year” and “Don't ask me to carry an oily rag like that”) while the other sentences vary from one participant to another (e.g. “Bagpipes and bongos are musical instruments” and “It's hard to tell an original from a forgery”). In total, TIMIT contains 6300 utterances.
3. The speech samples were recorded in a noise-free environment at a sample rate of 16 kHz with the time duration between 1 to 3 seconds based on the length of sentence and the rate of speech. So the corpus comprises 5 hours of speech.

4. The database contains several files associated with each speech sample, including speech waveform files (.wav) and text files (.txt), which contain each sentence pronounced by each participant.

Table 3-1 illustrates the dialect distribution of speaker participants for TIMIT

Table 3-1: Dialect Distribution of speakers(Garofolo et al., 1993)

Dialect Region	Male Speaker	Female Speaker	Total Speaker
1.New England	31 (63%)	18 (27%)	49 (8%)
2.Northern	71 (70%)	31 (27%)	102 (16%)
3. North Midland	79 (67%)	23 (23%)	102 (16%)
4. South Midland	69 (69%)	31 (31%)	100 (16%)
5. Southern	62 (63%)	36 (37%)	98 (16%)
6. New York City	30 (65%)	16 (35%)	46 (7%)
7. Western	74 (74%)	26 (26%)	100 (16%)
8.Army brat	22 (67%)	11(33%)	33 (5 %)
Total Speaker	438(70%)	192 (30%)	630 (100%)

3.1.B University of Salford Anechoic Chamber Database (SALU-AC)

In this section, a new database called SALU-AC dedicated to speaker recognition is presented. The main characteristics of this database are that it contains English spoken by native and non-native speakers, provides speech samples in different languages (in addition to English speech samples), and the environment where they were recorded. The main objective of creating this database for speaker recognition is to provide very clean speech in English and other languages, since data were collected in the Anechoic Chamber at the University of Salford that make it more efficient when want to deal with one adverse condition (for example noise) separately from other adverse conditions. Furthermore, this database can also be used in other speech processing fields like speech and language recognition. This database has been evaluated using MSR toolbox (Sadjadi et al., 2013), and was shown to provide a good speaker recognition rate. The description of SALU-AC is as follows:

1. Recording Environment

In this database, the audio speech samples were recorded in an Anechoic chamber in the University of Salford. An anechoic chamber is an acoustic room designed to be isolated from any sound reflection from wall, floor, and ceiling. It represents an ideal environment for any acoustic experiments since the chamber doesn't affect the measurements. The dimensions of this chamber are$(5.4 \times 4.1 \times 3.3\mathrm{m})$; the wall, floor, and ceiling are made of heavy Accrington bricks and concrete to prevent sound getting into the room with two heavy acoustic doors with rubber seals used to minimise airborne sound. In addition, every surface is covered with absorbent materials to absorb any sound, and the floor is covered with a wire trampoline stretched between the walls with an acoustically transparent catch net below. The design of this chamber made it one of the quietest rooms in the whole world with background noise level approximately -12.4 dBA and cut off frequency 100 Hz (Figure 3.1). Accordingly, the speech samples recorded in this room are a very clean sample from any environmental noise and any other adverse conditions.

2. Recording Equipment

The speech samples were recorded using a Zoom H6 Handheld recorder manufactured by Zoom Electronics. On-board X/Y stereo condenser microphones were arranged with the right and left microphones on the same axis. This design ensured that the microphones were always an equal distance from the sound source for perfect localisation without phase shifting. Throughout recording, only one side of the microphone was used (one channel), the distance between speaker and microphones was approximately 15-30cm, and samples were recorded with a sampling rate of 16 kHz with a bit rate of 16. The recording file can be recorded in different formats (such as .wav, and mp3). In the database, the speech samples were recorded into the most commonly used file type:- .wav.

Figure 3.1: Anechoic Chamber at University of Salford

3. Recording procedure and speaker set details

In the SALU-AC database, the audio speech samples were recorded from 110 volunteer speakers (55 male, and 55 female). Each volunteer was instructed to speak 5 speech samples with different duration times (approximately 60-70 seconds for the first sample and 30-40 seconds for the other samples) by using general texts from different readable resources (newspapers, books, leaflets, articles, etc.) without any constraint for the volunteers on what text to read (i.e. the volunteers were completely free to select any text they wanted to read or speak). The main aim of using these readable resources was to help the speaker to speak fluently without any long pause, in order to get very clean audio speech samples in addition to providing text-independent samples. Furthermore, each volunteer speaker was also instructed to provide two types of samples. The first type of samples was recorded using English. The second type was recorded based on the native language of the volunteer in order to provide language and phoneme independent samples (if the speakers only spoke English, they were requested to provide all the audio samples in this language).The main goal of providing non-English speech samples was to investigate the effect of language mismatch between enrolment and recognition phases. Furthermore, these samples could be used for language recognition. The audio speech samples for each speaker were divided into short 5 seconds sample utterances in order to obtain between 25 to 30 utterances for each speaker divided into two groups: one for English utterances and one for native language speaker utterances (where the speaker was not an English native speaker) Table 3-3 demonstrates the language distribution of speaker participants for SALU-AC. The database thus contained between 2500 and 3000 speech utterances. Nine utterances from each speaker were used to generate reference models in the enrolment phase (45 seconds training) and the remaining utterances were used as recognition samples.

The main advantages of the SALU-AC databases over TIMIT were that it was not limited to one language or particular sentences, which enabled the evaluation of speaker recognition based on this database to be more accurate and more generalised than TIMIT. Table 3-2 illustrates a comparison between the two data sets.

Table 3-2 Comparison between SALU-AC and TIMIT

SALU-AC	TIMIT
• Speech samples are recorded in an anechoic chamber which makes them very clean.	• Recorded in noise-free environments
• English speech samples were obtained from 110 Speakers (55 male, 55 female) including English native and non-native speakers.	• English speech samples were obtained from 630 speakers (438 Male and 192 female speakers) of 8 various dialects of American English
• For non-native speakers, SALU-AC provides two sets one for English utterances and the second for native language of the speaker.	• Provides speech samples in English only.
• The time duration for each speech sample is 5 seconds with sample rate 16 kHz.	• The time duration for each speech sample between 1-3 seconds with sample rate 16 kHz.
• Speakers are not constrained by specific sentences.	• Each speaker provided 10 utterances for specific sentences

Table 3-3: Language Distribution of SALU-AC

Speaker Native Language	Male Speaker	Female speaker	Total Speaker
English (Native language)	21	15	36
Arabic	8	14	22
Spanish	0	4	4
Hindi	4	3	7
Polish	1	3	4
Italian	1	3	4
French	8	2	10
Chinese	2	2	4
Farsi	1	2	3
Kurdish	0	1	1
Romanian	0	1	1
Portuguese	1	1	2
Shona	0	1	1
Swahili	0	1	1
Bengali	2	1	3
Greek	2	1	3
Urdu	1	0	1
German	1	0	1
Nigerian	1	0	1
Indonesian	1	0	1
Total Speakers	55(50%)	55(50%)	110 (100%)

3.2 Noise Dataset

Six types of noises were used in this research, four of which were environmental noises, and they were recorded by Loizou (2013) using a laptop and a MC391 (Shure) microphone. Each one was recorded with a 44.1 kHz sample rate. In addition to artificial white noise and Pink noise, the types of noises were as follows:

1. **Cafeteria Speech Babble Noise**: This type of noise refers to multi-talker babble in the University's cafeteria where the interference is speech from other speakers in the vicinity. This noise is uniquely challenging because of its highly time evolving structure and similarity to the target speaker.
2. **Interior of moving Car:** the noise heard inside a moving car with fixed speed 60m/second. This type of noise has a similar frequency content to brown noise and it is periodically modulated due to the uneven road surface.
3. **Interior of underground train compartment:** The sounds heard inside a moving train compartment, containing static noise, pulses of sound from motion along tracks, and sounds changing frequency over time due to change of train speed.
4. **Street Noise**: Sounds heard while walking in the streets of a city. This type contains noise from traffic, sirens, wind noise, and miscellaneous sound from the street.
5. **White Noise**: representing a random signal of equal intensity at different frequencies, giving it a fixed power spectrum density. In discrete time, white noise is described as a discrete signal whose samples are regarded as a sequence of serially uncorrelated random variables with zero mean and finite variance.
6. **Pink Noise:** refers to a signal with a frequency spectrum such that power spectral density is inversely proportional to the frequency of the signal. In pink noise, the total sound power in each octave in the same as the total sound power in the octave immediately above or below it. This type of noise is used only in noise training investigation described in **Chapter 5**.

3.3 Mixing Procedure and noisy data

As mentioned before, the evaluation speech datasets were typically recorded in clean environments (Anechoic Chamber) with no environmental noise affecting the recorded speech signal. The noisy speech samples were therefore obtained by mixing speech samples with the various types of noise mentioned in the last section in different signal to noise ratios (SNR) (20, 15,10,5, and 0 dB). Figure 3.2 illustrates the mixing process of speech and noise signals.

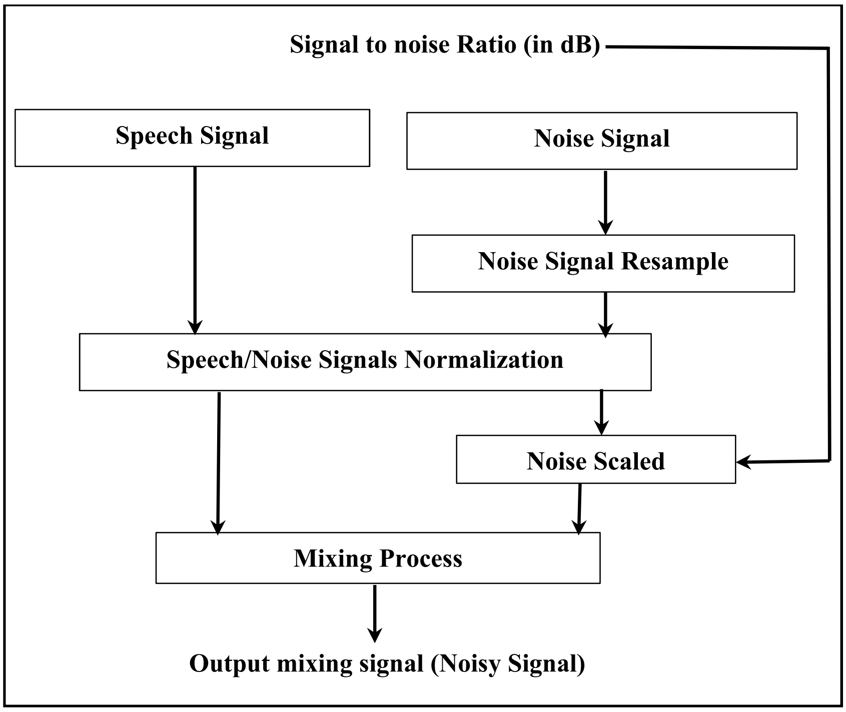

Signal to noise Ratio (in dB)
Speech Signal
Noise Signal
Noise Signal Resample
Speech/Noise Signals Normalization
Noise Scaled
Mixing Process
Output mixing signal (Noisy Signal)

Figure 3.2: Mixing process block Diagram of Speech and noise Signal

The procedure of mixing was as follows:

1. The Noise signals were truncated to become the same length as target speech utterances. The aim of this step was to make sure that mixing of noise would occur on the entire speech signal. This signal was then resampled to be equal to the speech signal sample rate.
2. The signal to noise ratio SNR (in dB) was specified to determine the ratio at which the speech signal and noise were mixed. For research purposes, a range of mixing between 20dB to 0dB was selected, since the SNR above 20dB is almost clean (i.e. the ratio of speech is high compared with the noise, which is too low) and SNR lower than 0 dB is barely recognised by the human ear.

3. The speech and the noise signals were normalised (root mean square was used for signal normalisation).
4. Finally, the noise signal was scaled to reach the desired SNR before it was mixed with the speech signal. Figure 3.3 illustrates the same signal contaminated with cafeteria babble noise with different SNRs.

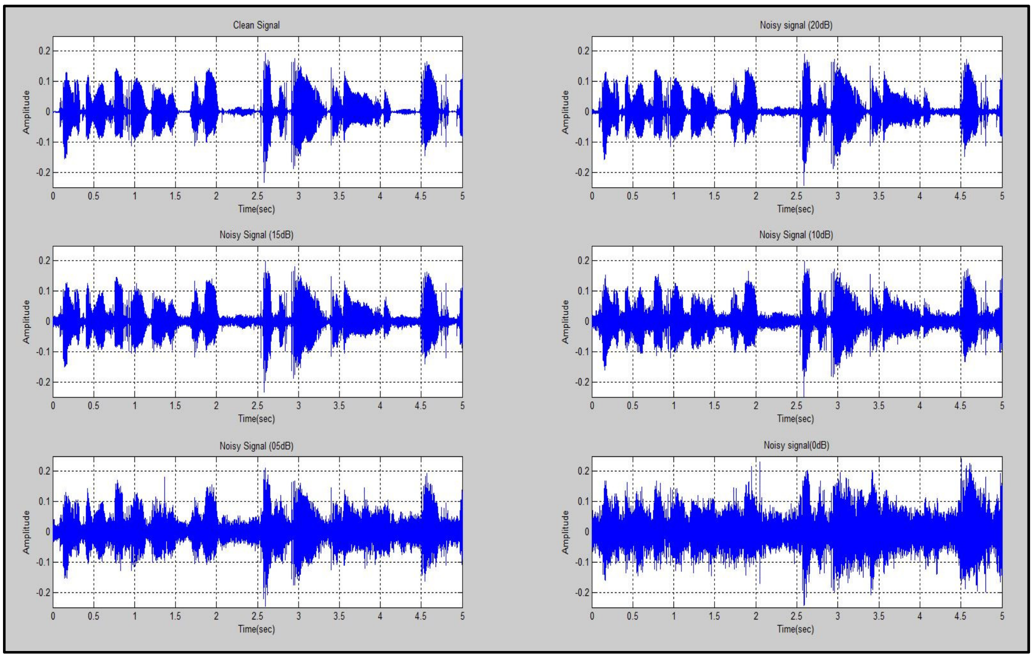


Figure 3.3: Speech Sample contaminated with different SNRs

3.4 Chapter Summary

This chapter has given a description of the speech databases used in this research and the types of noise adopted. A new speech database for speaker recognition is presented in this chapter (SALU-AC), with the main characteristics of this database followed by an explanation of the mixture process used to obtained noisy speech samples contaminated with different SNRs. Since it is recorded in a clean environment, SALU-AC represents a good model to deal with environmental noise separately from other types of adverse condition when mixing with different types of noise. Then, it is easy to deal with the noise issue far from any other effect of adverse conditions. TIMIT, on the other hand, represents the most commonly used database to evaluate the performance of speaker recognition application.

CHAPTER 4

GMM-UBM BASED SPEAKER RECOGNITION

Chapter Overview

This Chapter focuses mainly on explaining the structure of the Speaker Recognition based on GMM-UBM which is adopted in this research and how this system works. This is followed by a description of experiments that investigate the variant effects of different kinds of noise with different SNRs on the robustness of Speaker Recognition. A review, and how GMM-UBM based SR works, are presented in section 4.1, and the experiments that show the impact of different environmental noises on speaker recognition are presented in section 4.2.

4.1 GMM-UBM based Speaker Recognition

In order to implement Speaker Recognition based on GMM-UBM, the Microsoft speaker recognition (MSR) identity toolbox (Sadjadi et al., 2013) was adopted for this purpose. The next sections give a review of the GMM-UBM, and then the process of this system is explained.

4.1.A Review of GMM-UBM based Speaker Recognition

GMM-UBM Speaker Recognition (like other Speaker Recognition system models) includes two primary components: a front-end and a back end. The front-end of this system is responsible for transforming the speech signals to acoustic features in the features extraction process. The Cepstral features are most commonly used with this SR model (as seen in chapter 2). The back–end, on the other hand, includes the training (enrolment) and testing (recognition) phases. The training (enrolment) phase is responsible for estimating a model for each registered speaker to generate a reference model. This can be done by using one of the modelling approaches explained in Chapter 2. In the test phase (recognition), on the other hand, 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 the claimed speaker to make a decision on whether the speaker is the target speaker or an imposter.

4.1.B How GMM-UBM based Speaker recognition works

As mentioned earlier, the GMM-UBM model has been adopted for this research. The operation of this SR model, like any other recognition system, can be summarised in the following processes (Figure 4.1):

1. Pre-processing: This step includes the following processes: read signal, use Speaker Activity Detection (SAD) to remove the silence from the speech signal, ignore the SNR frames, and remove the linear channel effects.
2. Features parameters extraction: Extract Cepstral features for each audio speech utterance in the enrolment and recognition phases. This process is responsible for converting the audio speech signal for each speaker to Cepstral features (this study focuses mainly on MFCC and GFCC features as will described later in Chapter 6). The features of each speaker are then stored in a unique file. The created files are classified into training files (Train*i*-*j* , *i*=1,..,number of users *j*=1,..,number of utterances for *i*th speaker)which

represent the features space for each user (*i*) that has been used to create the reference model, and, testing file (Test*i*, *i*=1,…number of users) which represents the feature space for each user(*i*) in recognition phase. More details about features extaction algorithms are given in Chapter 6.

3. Training a UBM from background data: This process is responsible for creating background models from a huge number of speakers by fitting GMM to acoustic features using binary splitting and Expectation Maximisation (EM).
4. Maximum a Posteriori (MAP): This process is responsible for adapting speaker specific GMM from UBM. The output of this process represents the reference model of each speaker. More details about generating UBM models and adaptation process are given in Chapter 5.
5. Scoring verification: Computes the verification score between the reference model of the claimed speaker and recognition features of the input speech signal (Test *i*). The scores are measured as a log-likelihood ratio between the two models.
6. System Evaluation: Making an evaluation for the performance of the system depends on the output of verification (in this work error equal rate (EER) and Detection Error Trade-Off (DET) Curve were adopted).

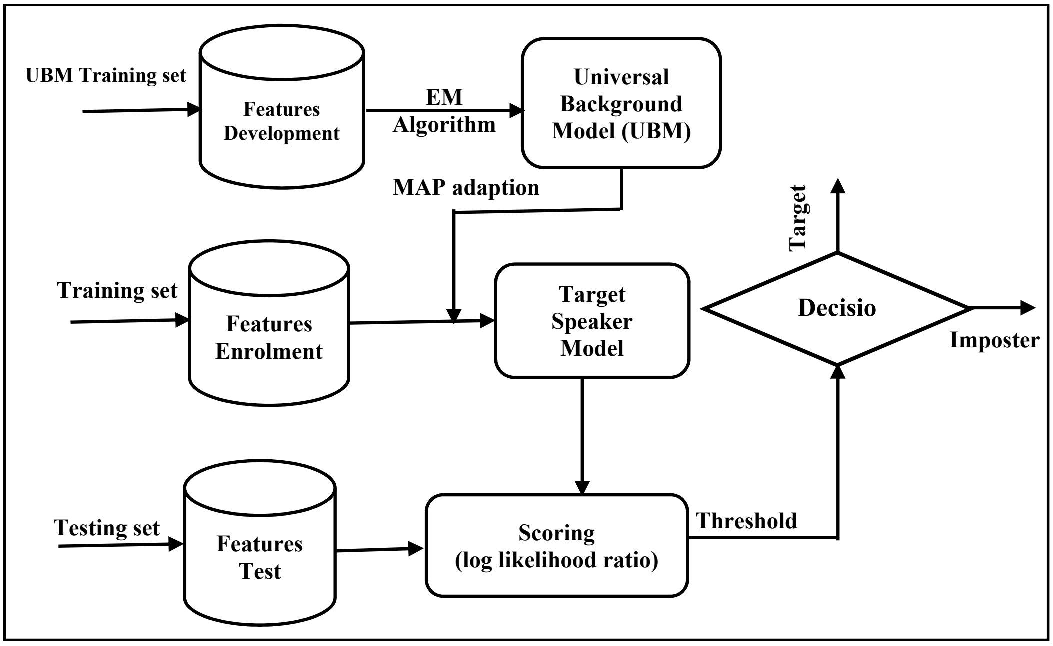


Figure 4.1: Gaussian Mixture Model -Universal Background Model Speaker Recogniton (Sadjadi et al., 2013)

In most of the experiments in this research, a GMM baseline speaker recognition was used instead of a GMM-UBM one. The main reasons behind this are the limitation of data in SALU-AC compared with the data required to train UBM. Furthermore, in a training on the fly approach, the adapted model for claimed person requires to be built from scratch instead of adapted from UBM. To be more precise, in these experiments UBM is still used, but the speech data used in UBM and in training are the same. Thus each speaker will adapt his/her own model from UBM instead of adapting an independent model using MAP. This is also useful because the adapted model from UBM will be treated as a supervector in order to make the recognition more robust. For simplicity, we call this the GMM based system. More details about the difference between GMM and GMM-UBM baseline speaker recognition will be given in the next chapter.

4.2 The Impact of environmental noises on the performance of SR

In the last section, a detailed description GMM-UBM based system was given. In this section, the impact of various noises with different SNRs on the accuracy of verification of this system will be investigated. Statistical relationships between recognition accuracy and SNR have therefore been established. Results show various noises with different SNRs degrade the performance of verification to different extents.

4.2.A Experimental Setup

The aim of this experiment is to characterise the effectiveness of noisy speech sets contaminated with different SNRs and different types of noise on the accuracy of the GMM-UBM speaker recognition. The details of these experiments are as follows:

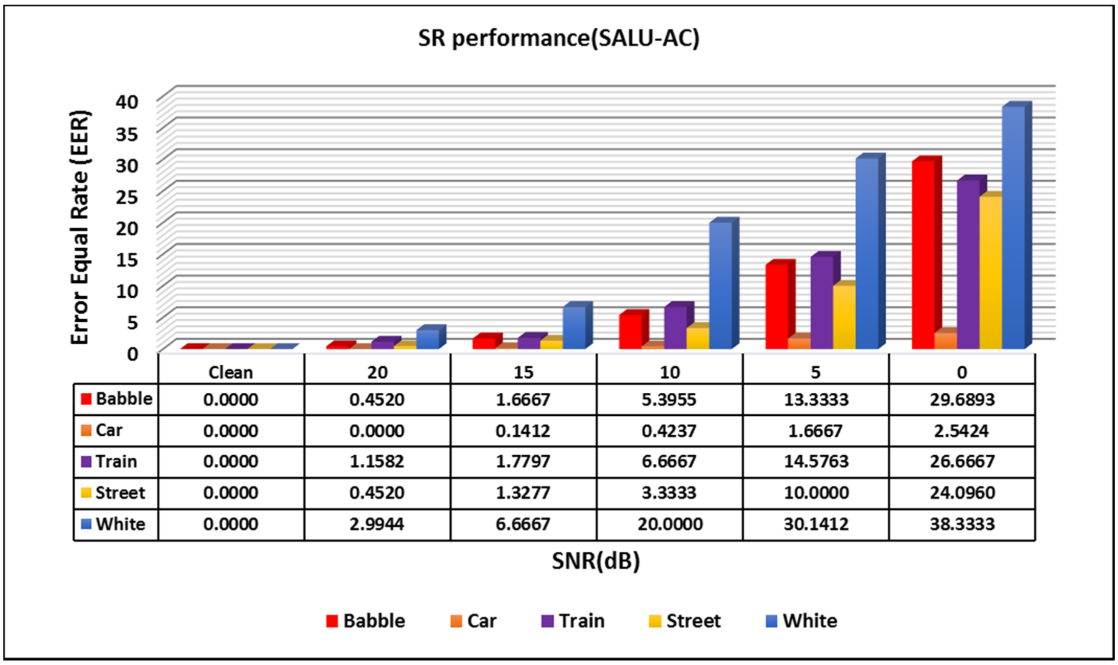
- Mel Frequency Cepstral coefficients MFCC including delta and delta-delta coefficients were used in these experiments with dimension 39*514 where the row represents the number of Mel-filters and the column represents the number of frames (more details about MFCC extraction are given in Chapter 6).
- Gaussian mixture model GMM was applied for creating reference models and classification with 256 Mixture. In this experiment, only GMM was applied.
- The experiments were conducted on two speech samples from 60 speakers (30 Male and 30 Female) in both SALU-AC and TIMIT databases. Five groups of noisy speech samples

were used in these experiments. Each group represents speech samples contaminated with a specific type of noise. Each group also includes five sub-groups which represent the same data contaminated with different signal to noise ratios, in addition to one group that includes a clean speech sample.

- The evaluation of the GMM-UBM based speaker recognition, in these experiments, is based mainly on Error Equal rate and Detection Error Trade-off (DET) graphs. These evaluations depend on 60 cases of target speakers (authorised speakers) against 3540 cases of imposters (unauthorised speakers) based on log-likelihood scores.

4.2.B Experimental Results and Discussion

As mentioned before, the aim of these experiments is to investigate the impact of differed types of noise with a various speech to noise ratios (SNRs). Figure 4.2 illustrates the degradation accuracy based on EER when the SNR decreases for each type of noise using the SALU-AC database. The X axis represents the SNR (in dB) between clean signals (which is usually greater than 25 dB) and 0 dB (where the level of speech and noise are equal) and each bar in the figure represents a different type of noisy speech (Cafeteria Babble, Interior Car, Interior Train, Street, and White noise).



	Clean	20	15	10	5	0
■ Babble	0.0000	0.4520	1.6667	5.3955	13.3333	29.6893
■ Car	0.0000	0.0000	0.1412	0.4237	1.6667	2.5424
■ Train	0.0000	1.1582	1.7797	6.6667	14.5763	26.6667
■ Street	0.0000	0.4520	1.3277	3.3333	10.0000	24.0960
■ White	0.0000	2.9944	6.6667	20.0000	30.1412	38.3333

Figure 4.2: Speaker Recognition Performance for different noises for SALU-AC

It is very clear that the various noises have different impacts on the accuracy of speaker recognition. Interior Moving Car noise, for instant, has only a slight effect on accuracy of SR (0% EER and 0.14% for 20, and 15 dB respectively) if compared with other types of noise like white and train, which have a higher impact on SR accuracy (2.9% and 6.6% EER for 20, and 15 dB in white noise, and 1.15% and 1.7% EER for Train for the same SNRs). On the other hand, the Cafeteria Speech Babble noise has more effect on accuracy at 0dB with 29.6% EER than interior moving train noise; while for other SNRs, it is clear the effect of interior moving train noise is higher. Figure 4.3 shows the DET graph for different types of noise in 10dB SNR. As can be seen, the effect of car noise on FPR (FRR) and FNR (FAR) is lower, while the white noise has a higher effect on both compared with other types of noise. Furthermore, the interior moving train, street, and cafeteria babble also have a high effect on accuracy, especially on false positive rate.

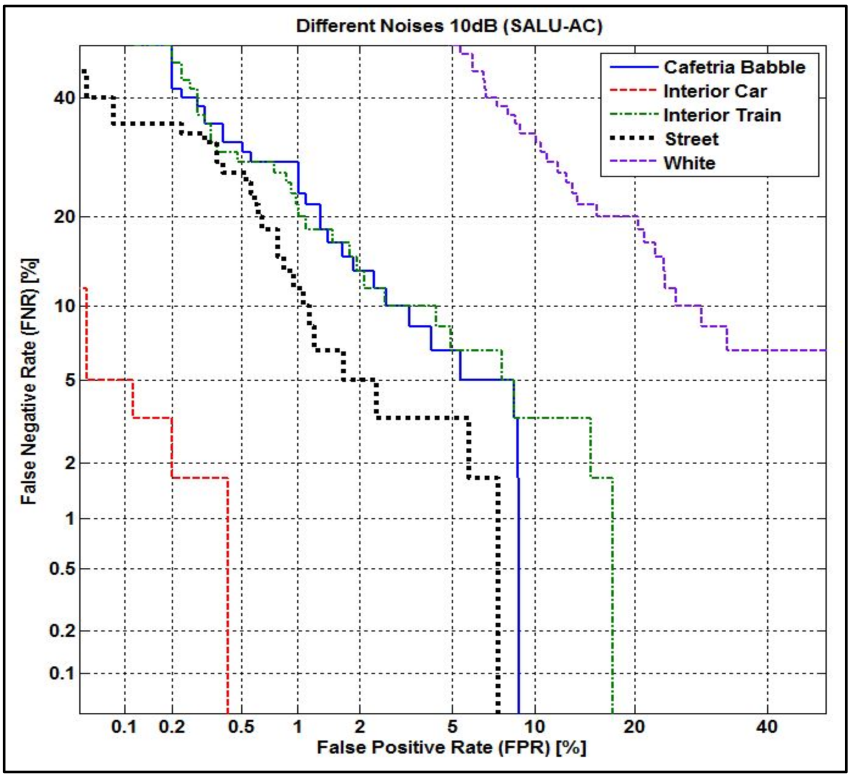
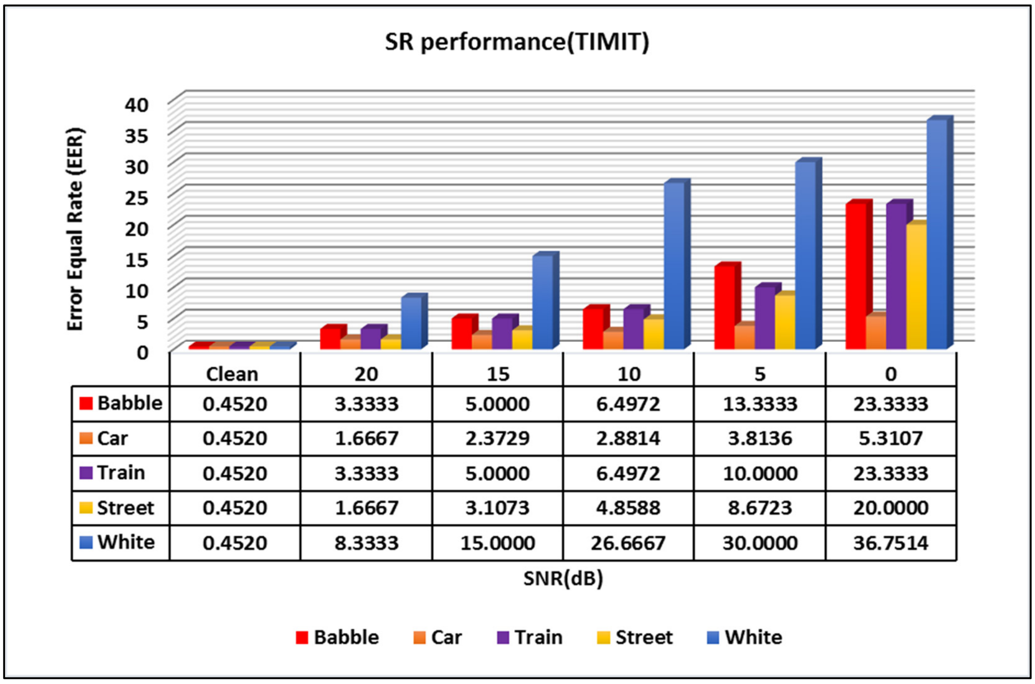


Figure 4.3: DET Graph for 10 dB SNR for Different types of noise in SALU-AC



	Clean	20	15	10	5	0
Babble	0.4520	3.3333	5.0000	6.4972	13.3333	23.3333
Car	0.4520	1.6667	2.3729	2.8814	3.8136	5.3107
Train	0.4520	3.3333	5.0000	6.4972	10.0000	23.3333
Street	0.4520	1.6667	3.1073	4.8588	8.6723	20.0000
White	0.4520	8.3333	15.0000	26.6667	30.0000	36.7514

Figure 4.4: Speaker Recognition Performance for different noises for TIMIT

When TIMIT was used, it is still clear that the white noise has a higher impact on accuracy than the other kinds of noise, while the car noise still has a limited effect compared with other types of noise (Figure 4.4). Cafeteria Babble and interior moving train noise have approximately the same effect, through whole SNRs except at 5dB where the effect of Cafeteria Babble is higher. Also, it can be noticed that the interior moving car and street noise have the same EER in higher SNR (20dB and 15dB with 0.45% and 1.6% EER respectively), while the effect of street noise becomes higher at 10, 5, and 0 dB (i.e. when SNR are decreased). Figure 4.5 shows a DET graph for these types of noise at 15 dB.

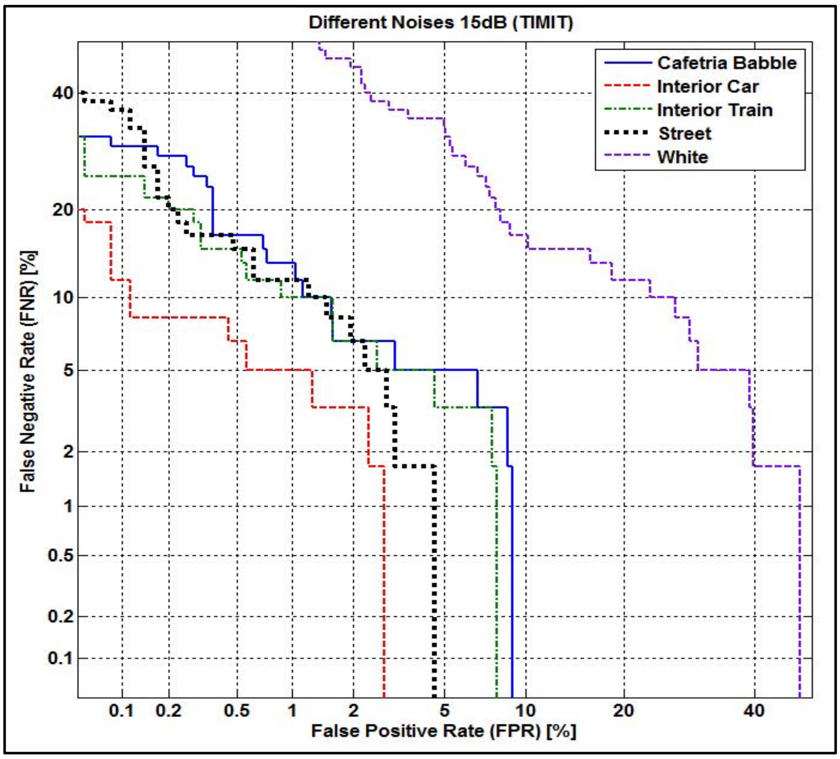


Figure 4.5: DET Graph for 15 dB SNR for Different types of noise in TIMIT

Again, the white noise showed the most effect on both FPR and FNR; Babble and Train have approximately the same effect on FPR, and the car noise has the least impact on SR for both FPR and FNR. Overall, through these experiments, it is clear that the impacts of different types of noise on speaker recognition accuracy are varied from one type of noise to another. Furthermore, the ratio of noise in signal (i.e. SNR) also plays a major role in the SR performance. In the next chapter, an investigation of using different noisy speech data with different SNRs in the enrollment phase has been investigated, as well as how much the robustness of speaker recognition can be improved.

4.3 Chapter Summary

In this chapter, the GMM-UBM based speaker recognition adopted in the study has been reviewed, with an explanation of how this type works. Finally, an implementation of this type of system with the impact of different types of noisy speech with different types of SNR has been investigated. The results show the variation of the effects of different types of noise on the performance of speaker recognition. Clearly, signal to noise ratio has the main role in the robustness of speaker recognition. Furthermore, the value of SNR can play a major role to find the threshold at which the speaker recognition performance is affected with additive noise. In the next chapters, different approaches are employed to improve the robustness of speaker recognition in environmental noise.

CHAPTER 5
IMPROVING THE ROBUSTNESS OF SPEAKER RECOGNITION IN NOISY CONDITIONS VIA TRAINING

Chapter Overview

As mentioned before, environmental noises are known to be one of the greatest challenges in speaker recognition since they significantly compromise the system reliability due to Channel Mismatch. The previous chapter investigated the effect of different types of noise and different SNRs on the robustness of speaker recognition. This chapter describes attempts to improve robustness by including possible channel mismatching, i.e. using noisy speech to create training models. Validation testing was carried out in emulated noisy conditions with a controlled signal to noise ratio. Below (Section 5.1) is a description of how to employ Gaussian mixture model (GMM) and Gaussian mixture model - Universal Background Model (GMM-UBM) for modelling and classification, followed by experiments to investigate the robustness of using GMM and GMM-UBM with limited data in section 5.2. After that, the experiments using noisy data in the enrolment phase and their results are described in 5.3.

5.1 Gaussian Mixture Model (GMM) and Universal Background Model (UBM)

The Gaussian Mixture model is one of most commonly used models in the field of speaker recognition. This model represents the probability density functions (PDFs) of a random variable by a weighted sum of **α** components (or Mixtures), where α is an integer. This weighted summation is given by Reynolds (1995) and Reynolds and Rose (1995):

$$\boldsymbol{p(\sigma|\beta)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{\alpha}} \boldsymbol{w_i \mathcal{N}(\sigma|\mu_i, \theta_i)} \qquad \textbf{(5.1)}$$

where σ is a H-dimensional feature vector, w_i are the mixture weights of each of the components . $\mathcal{N}(\sigma|\mu_i, \theta_i)$ is a Gaussian density function parameterized by μ_i which represents the mean of the σ vector , θ_i is a covariance matrix, and i=1,…,α,. $\mathcal{N}(\sigma|\mu_i, \theta_i)$ are constrained by $\theta_i w_i = 1$.

The Gaussian density function is given by:

$$\boldsymbol{\mathcal{N}(\sigma|\mu_i, \theta_i)} = \frac{\boldsymbol{1}}{\boldsymbol{(2\pi)^{\frac{d}{2}}|\theta_i|^{\frac{1}{2}}}} \times \boldsymbol{exp\{-\frac{1}{2}(\sigma-\mu_i)'(\theta_i)^{-1}(\sigma-\mu_i)\}} \qquad \textbf{(5.2)}$$

for $i = \{1,2,\ldots,\alpha\}$. Operations $|.|$ and $(.)'$ indicate the determinant and transpose respectively. The collection of the mixture weights w_i, means μ_i , and covariance θ_i can be represented in notation *β*, such that $\beta = \{w_i, \mu_i, \theta_i\}$. The form of the models generated by GMM supports full variance matrices. Therefore, the GMM implemented based on diagonal covariance matrices are applied. The main reasons are in three parts. First, as a result of employing cepstral feature parameters, they are to be highly uncorrelated. The covariance of these parameters is therefore trivially small compared with the covariance of the matrix which is diagonally dominant. Second, using one covariance matrix for each mixture offers better modelling capabilities (Reynolds and Rose, 1995). In addition, employing diagonal matrices requires small storage, and improves computational efficiency and simplicity (Pillay, 2010).

To estimate the parameters of GMM the iterative Expectation-maximization algorithm (EM) is commonly used (Dempster et al., 1977). EM is an unsupervised approach, which depends mainly on the maximum likelihood (ML) principle. Reynolds et al. (2000) proposed a novel approach for creating speaker dependent GMM. This approach depends on using Maximum a Posteriori (MAP) adaptation of a speaker independent model, which is based on a Bayesian framework. Conversely to maximum likelihood enrolling, this approach is based on the assumption of the prior distribution of a model, which is derived from speaker independent distributions. This is generated by applying the EM algorithm depending on using a large set of speakers. This set of models from independent speakers is commonly known as the Universal Background Model (UBM) or world model, while the whole approach is referred to as GMM-UBM or adapted-GMM (Reynolds et al., 2000).

The approach of GMM-UBM can be described as follows:

- Given β_{UBM} which represents UBM extracted by EM, and a set T of enrolling vectors, $\sigma = \{\sigma_1, \sigma_2, \dots\dots, \sigma_T\}$ (extracted from speech segment), the probabilistic alignment of enrolling features vectors related to the α mixture of the UBM is determined. This can be done by measuring the *a posteriori* probability for acoustic mixture *i*, given the observation σ_t

$$\boldsymbol{p(i|\sigma_t, \beta) = \frac{w_i p_i(\sigma_t)}{p(\sigma_t|\beta)} = \frac{w_i p_i(\sigma_t)}{\sum_{k=1}^{\alpha} w_k p_k(\sigma_t)}} \qquad \textbf{(5.3)}$$

- Next, calculate the appropriate statistics for weights w_i, means μ_i , and variance θ_i of each mixture *i* as follows (Reynolds et al., 2000):

$$\boldsymbol{m_i = \sum_{t=1}^{T} p(i|\sigma_t, \beta_{UBM})} \qquad \textbf{(5.4)}$$

$$\boldsymbol{E_i(\sigma) = \frac{1}{m_i} \sum_{t=1}^{T} p(i|\sigma_t, \beta_{UBM}) \sigma_t} \qquad \textbf{(5.5)}$$

$$\boldsymbol{E_i(\sigma^2) = \frac{1}{m_i} \sum_{t=1}^{T} p(i|\sigma_t, \beta_{UBM}) {\sigma_t}^2} \qquad \textbf{(5.6)}$$

where m_i, $E_i(\sigma)$ and $E_i(\sigma^2)$ are the count and the first and second moment of the enrolment features respectively.

- Based on statistics from the previous step, the new estimates for each mixture of the adapted model are then obtained by merging them with existing UBM parameters as follows (Reynolds et al., 2000):

$$\widehat{\boldsymbol{w}_i} = \left[\boldsymbol{\delta}_i^{w} \boldsymbol{m}_i \frac{\mathbf{1}}{\boldsymbol{T}} + (\mathbf{1} - \boldsymbol{\delta}_i^{w})\boldsymbol{w}_i\right]\boldsymbol{\gamma} \qquad \textbf{(5.7)}$$

$$\widehat{\boldsymbol{\mu}_i} = \boldsymbol{\delta}_i^{\mu} \boldsymbol{E}_i(\boldsymbol{\sigma}) + (\mathbf{1} - \boldsymbol{\delta}_i^{\mu})\boldsymbol{\mu}_i \qquad \textbf{(5.8)}$$

$$\widehat{\boldsymbol{\theta}}_i^2 = \boldsymbol{\delta}_i^{\theta} \boldsymbol{E}_i(\boldsymbol{\sigma}^2) + \left(\mathbf{1} - \boldsymbol{\delta}_i^{\theta}\right)\left(\boldsymbol{\theta}_i^2 + \boldsymbol{\mu}_i^2\right) - \widehat{\boldsymbol{\mu}}_i^2 \qquad \textbf{(5.9)}$$

where γ is a scaling factor that ensures all adapted mixture weights sum to unity. The adaptation coefficients $\{\delta_i^w, \delta_i^\mu, \delta_i^\theta\}$ are coefficients for mixture i^{th}, which are responsible for controlling the balance between old and new estimates for weight w_i, mean μ_i, and variance θ_i. These compute as:

$$\boldsymbol{\delta}_i^{g} = \frac{\boldsymbol{m}_i}{\boldsymbol{m}_i + \boldsymbol{R}^{g}} \qquad \textbf{(5.10)}$$

where $g \in \{w, \mu, \theta\}$, $R^g$is the fixed relevance factor for parameter g . Generally, single adaptation coefficients are used, such as $\delta_i = \delta_i^w = \delta_i^\mu = \delta_i^\theta$. Moreover, the adaptation of only mean statistics gives better performance of speaker recognition compared with full adaptation of weight, means, and covariance for GMM as reported by (Reynolds et al., 2000).

As discussed earlier, the GMM-UBM approach includes the adaptation of the background model in UBM using register enrolling materials to obtain particular speaker GMMs. This UBM, as mentioned before, is enrolled by applying the EM techniques on a large amount of developed data from different speakers. Consequently, the parameters of each speaker model, through adaptation procedure, are derived by updating the well-enrolled parameters in the universal model according to available enrolling materials. Hence, the adaptation procedure results in tighter coupling between the particular speaker model and UBM (Pillay, 2010). The reason for this is that the parameters of the mixture, which are not supervised in the enrolling speech of a particular speaker are simply copied from the UBM. Furthermore, this tighter coupling obtained by the GMM-UBM approach allows a fast scoring technique to be implemented through the recognition phase without any degradation in accuracy (Reynolds et al., 2000).This approach depends on two observations: Firstly, for each feature vector of a recognition utterance, it is noticed that only a few of the mixture contributes significantly to the overall likelihood value. Second, it is also seen that feature vectors, which are near to a specific mixture in the general model (UBM) tend to be near the corresponding mixture in the speaker model (Pillay, 2010).

As a result, the fast-scoring approach merges these two observation as follows (Pillay, 2010): the Largest D scoring mixtures in UBM are determined for each feature vector, and the UBM likelihood is calculated depending only on those mixtures. Then, to evaluate the speaker's likelihood, the feature vector is scored with the corresponding D mixtures of the speaker model.

For instance, if we consider a UBM has D mixture, the fast scoring would include α+D computation for each feature vector instead of 2α log-likelihood in normal procedure. This is particularly useful especially when the amount of mixtures is large.

In the classification stage, as soon as the speaker GMMs are produced, this model has been used for recognising speakers based on their recognition utterances. In the case of speaker verification, the function is evaluating the likelihood of hypothesis (claimed) speakers model λ for given observation σ which is written by using Baye's theorem in form $p(\lambda|\sigma)$ such that:

$$\boldsymbol{p(\lambda|\sigma) = \frac{p(\sigma|\lambda)p(\lambda)}{p(\sigma)}} \qquad \textbf{(5.11)}$$

where $p(\lambda)$ is a priori likelihood of target speaker model which is considered similar for all models and can be ignored. $p(\sigma)$ is the unconditional likelihood of the observation σ being generated by any speaker. $p(\sigma)$ can also be represented as a constant value.

Next, in order to obtain a log-likelihood function, equation 5.11 is transformed into log-domain.

$$\boldsymbol{L(\sigma) = \log(p(\sigma|\lambda))} \qquad \textbf{(5.12)}$$

The decision of verification is then made depending on whether the result of log-likelihood $L(\sigma)$ is greater or less than a specified threshold φ ; for example, if $L(\sigma) \geq$ φ the speaker is classified as Target, otherwise classified as Imposter:

$$\mathbf{Dec}_v = \begin{cases} \boldsymbol{Imposter}, & \boldsymbol{L(\sigma) < \varphi} \\ \quad \textbf{Target}, & \boldsymbol{L(\sigma) \geq \varphi} \end{cases} \qquad \textbf{(5.13)}$$

where $\mathbf{Dec}_v$ represents the verification decision.

In the case of Speaker Identification (SI), a similar principle is applied, but in this case, the enrolled speaker model should be found which produced the highest log likelihood against the given recognition segment. This is given as:

$$\boldsymbol{S} = \mathbf{arg}\max_{\mathbf{1 \leq n \leq N}} \mathbf{log}(\boldsymbol{p}(\boldsymbol{\sigma} \,|\, \boldsymbol{\lambda_n})) \qquad \textbf{(5.14)}$$

5.2 GMM vs. GMM-UBM Experiments

In the previous chapter, the experiments of speaker recognition in noisy environments were presented based on a GMM approach instead of GMM-UBM. The reason for the use of Speaker Recognition SR[1] based GMM will be discussed in the report on the next experiments. These experiments studied the difference between using both systems (GMM, GMM-UBM) with limited data.

The experiments in this section were intended to evaluate the performance of speaker recognition in a noisy environment with limited data from a SALU-AC database using two different modelling based systems: GMM and GMM-UBM. Furthermore, the effect of language mismatch on both systems was investigated in the same environments. For the GMM-UBM based system, 80 speakers were used to train the UBM and 30 speakers as register speakers in SR (the enrolment speaker). For the GMM based system, only 30 were used as register speakers. To be more particular, the GMM based system also used UBM, but in this case, the data used to train UBM was the same data used in the training section. As a result, each enrolled speaker adapted his/her model from UBM instead of adapting an independent speaker model. For the recognition (test) phase, two sets of data were applied to both systems. The first set included

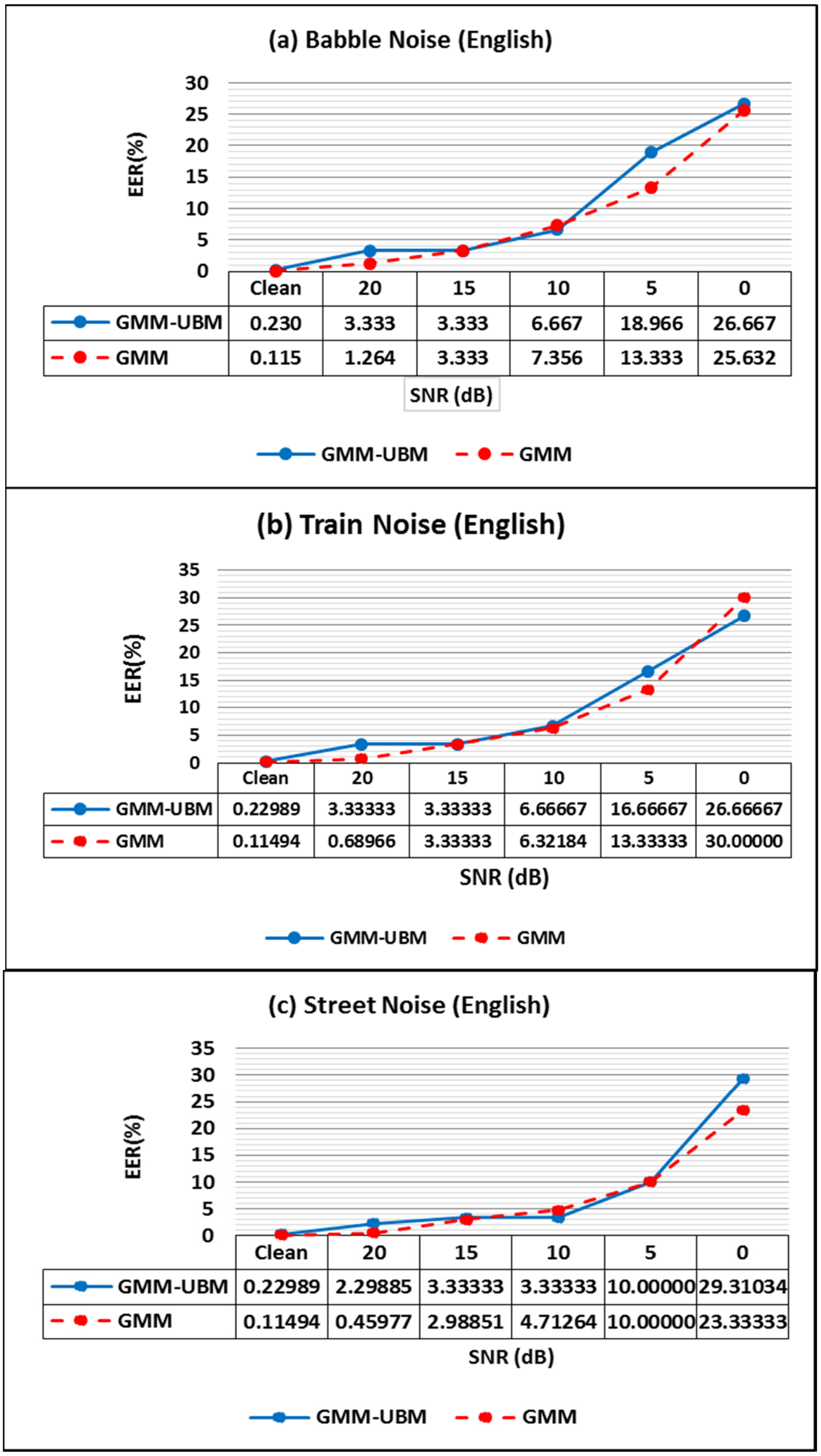
[1] the terminology of Speaker Recognition (SR) in this thesis is referred to Speaker Verification (SV)

speech samples using English only and the second recognition set included speech samples for the same speakers with different languages: each one with his/her native language. The details of these experiments are as follows:

- MFCC including delta and delta-delta coefficients were adopted in these experiments with dimension 39*514 where the row represents the number of Mel-filters and column represents the number of frames.
- GMM and GMM-UBM were used for generating reference models and classification with 256 components (mixture).
- The experiments were conducted on speech samples from 30 speakers (15 male, and 15 female) and 80 speakers for training UBM in the case of GMM-UBM model.
- For the recognition phase in both systems, two sets of recognition speech samples were used: the first set included the English speech utterances from 30 speakers, and the second set included different language utterances (except English) from the same speakers. In addition, each set included a number of subsets. Each subset represented the same speech samples contaminated with different SNRs (20, 15, 10, 5, and 0) for a specific type of noise. In these experiments, three types of noise were used (Babble, Train, and Street).
- In these experiments, the evaluation of both verification systems is based on using EER and detection error trade-off DET graph.

5.2.A Experimental Results

As discussed before, the main purpose of these experiments was to investigate the performance of SR using two modelling approaches, including GMM, GMM-UBM with limited enrolling data sets. The main motivation for these experiments was to check the performance of the universal model for SR using SAL-AC, which is, unlike TIMIT, based on speech samples from a limited number of speaker participants. In particular, as discussed before, UBM requires a huge number of speaker utterances to achieve robust recognition performance. Figure 5.1 a, b, and c shows the accuracy of speaker recognition (based on EER) using the first recognition set (the set that includes English utterances) based on GMM and GMM-UBM modelling approaches. The abscissa represents the signal to noise ratio SNR in dB, while the ordinate represents the recognition accuracy of the system based on EER. Each line graph represents the degradation of the impact of specific noise on the accuracy of speaker recognition. It is clear that both systems have very close EER in clean and high SNR, and possibly, that SR based GMM has slightly better EER than SR based GMM-UBM. For instance, in the clean and 20dB SNR for the three types of noisy speech (Babble, Train, and street) the SR based GMM has slightly less ERR than SR based GMM-UBM. However, both have the same accuracy in 15dB for speech contaminated with babble and train noise (with 3.33% ERR), while for noisy street speech the better performance is from the GMM based system with very small differences (2.98% EER). However, for the noisy speech with 5 and 0 dB, the accuracy of both systems varies from one type of noise to another.



(a) Babble Noise (English)

	Clean	20	15	10	5	0
GMM-UBM	0.230	3.333	3.333	6.667	18.966	26.667
GMM	0.115	1.264	3.333	7.356	13.333	25.632

(b) Train Noise (English)

	Clean	20	15	10	5	0
GMM-UBM	0.22989	3.33333	3.33333	6.66667	16.66667	26.66667
GMM	0.11494	0.68966	3.33333	6.32184	13.33333	30.00000

(c) Street Noise (English)

	Clean	20	15	10	5	0
GMM-UBM	0.22989	2.29885	3.33333	3.33333	10.00000	29.31034
GMM	0.11494	0.45977	2.98851	4.71264	10.00000	23.33333

Figure 5.1: Speaker Recognition accuracy (EER) for GMM/GMM-UBM based systems (first set)

Figure 5.2 shows a DET graph for both systems for 15 dB noisy speech for three types of noise. It is clear that the GMM based system (the red dashed curve) gives the best performance over GMM-UBM (the blue bold curve) for the three types of noise in both FPR and FNR, but in some cases (e.g. babble noise) the accuracy of FPR is slightly different.

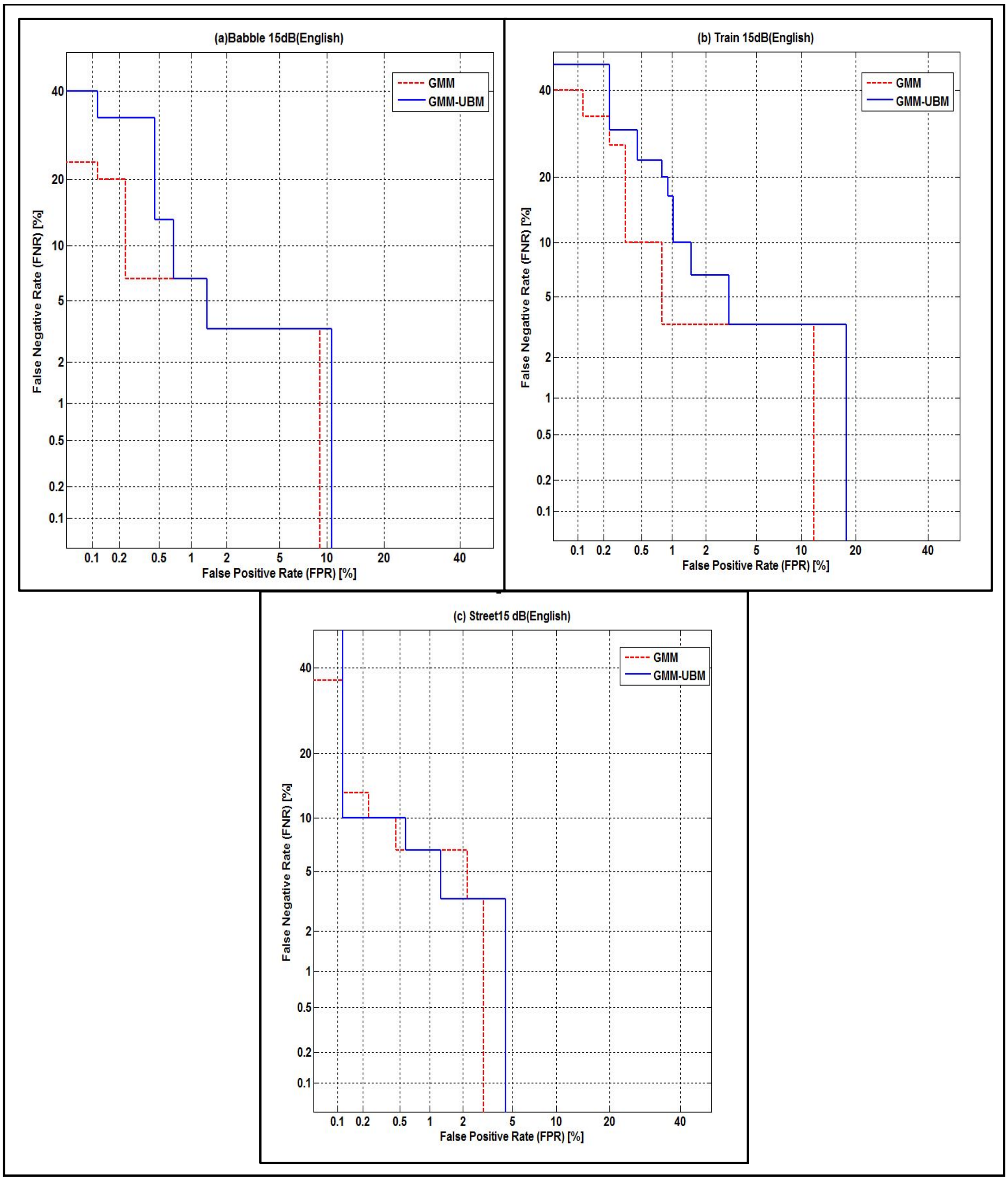


Figure 5.2: DET graph for both GMM/GMM-UBM based systems (first set)

Conversely, when the second set of speech samples are used (the set that includes utterances from different languages) in the recognition phase (Figure 5.3), the performance of SR based GMM-UBM shows a significant decrease in accuracy in the clean and high SNRs (20dB and 15dB) for the three types of noise, while SR based GMM shows more robustness to language mismatch in the same SNRs. However, both based SRs show almost the same EER between 10 and 0 dB for speech utterances contaminated with train and street noise (Figure 5.3 b, c). Babble noise, on the other hand, shows higher performance from the GMM-UBM based system for the same SNRs (Figure 5.3 a).

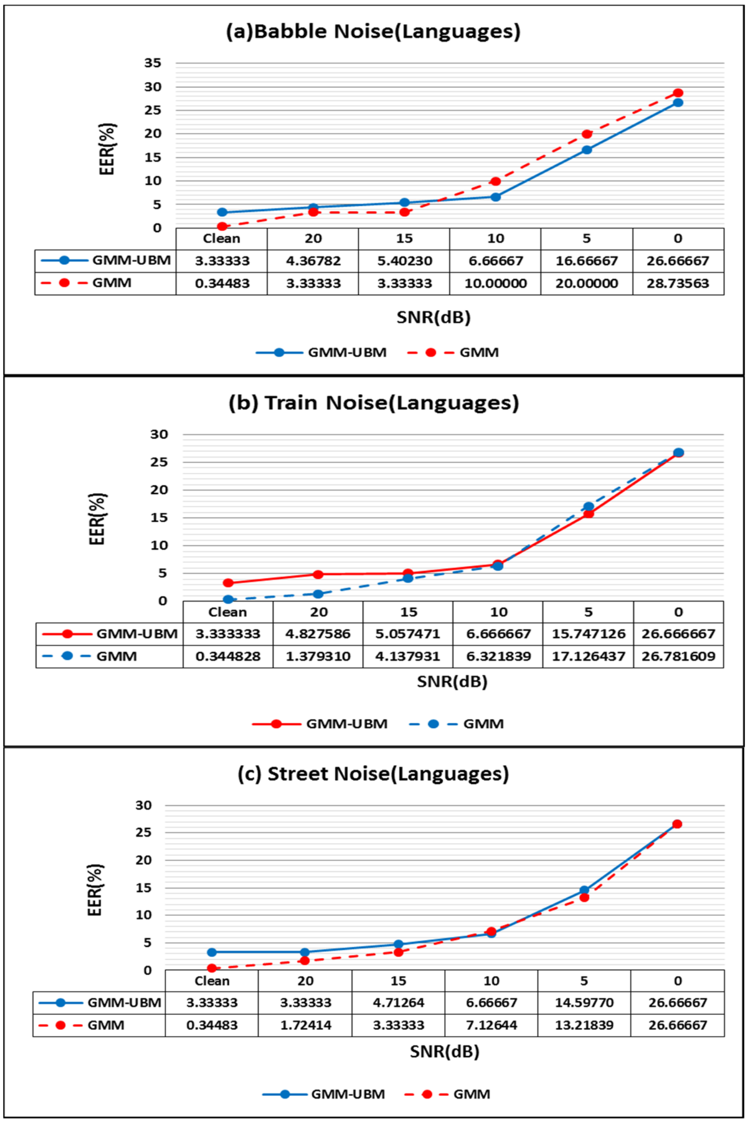


Figure 5.3: Speaker Recognition accuracy (EER) for GMM/GMM-UBM based systems (Second set)

In the DET graph for the second set (Figure 5.4), which represents the performance of both GMM (the red dashed curve), and GMM-UBM based SRs in 20dB SNR for the three type of noise, it is clear that there is a significant difference in performance, with GMM base line demonstrating improvement in both FPR and FNR over GMM-UMB baseline.

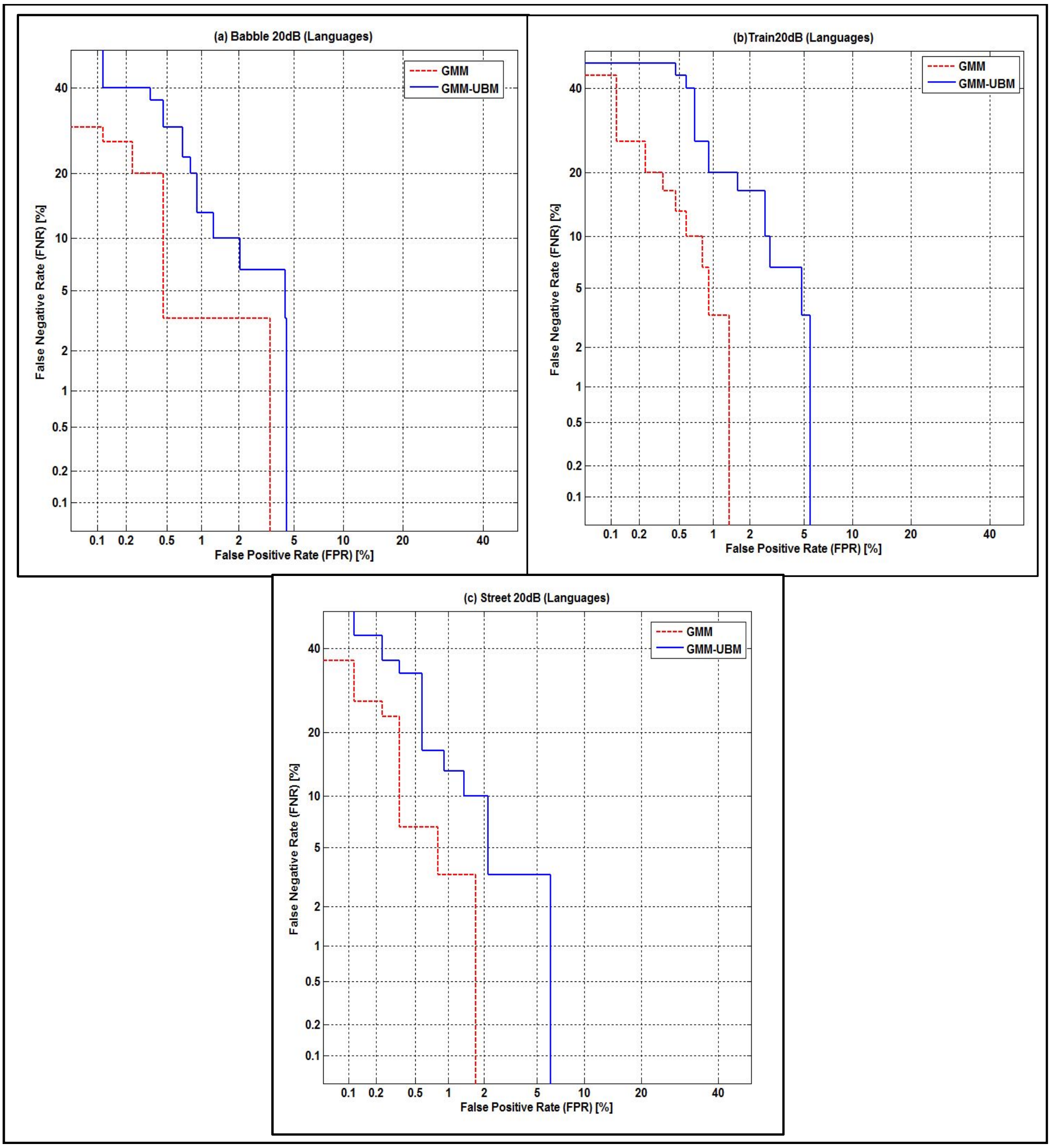


Figure 5.4: DET graph for both GMM/GMM-UBM based systems (Second set)

From the last experiments, the main findings can be summarised as follows:

- Both baseline(GMM and GMM-UBM) give different performance in different types of noise and different SNRs. However, in general, the GMM baseline is more robust to noise in clean and high SNRs while in low SNRs both systems have almost the same performance in most cases.
- The GMM-UBM baseline is affected by language mismatches more than GMM-baseline, especially when the signal is clean or has high SNR, which makes GMM baseline more robust, especially when it is known that each speaker will adapt its model instead of using an independent model.

As a result, the findings of this study suggest that it is better to use GMM baseline than GMM-UBM baseline with limited data, especially with a database like SALU-AC. Therefore, these experiments may lead to the following question: is the SR with GMM baseline better than the one with GMM-UBM baseline? The answer is no, and that was shown by Reynolds et al. (2000). However, the main important point in the GMM-UBM baseline is providing a large number of utterances for training the background model to obtain a good SR performance. In this research, it was possible to do this with the TIMIT database, since there was enough data for the background model to learn. However, in the case of the SAL-AC database, it is difficult to apply UBM, since it is based on a very limited number of speech samples.

One other important reason, as will be seen in Chapter 8, 'Training on the fly', where it is intended to build an adaptive model based on estimated SNR and type of noise, it is necessary to build one from nothing instead of adapting one from UBM. Otherwise, the UBM should re-train again to make it suitable for the on the fly model.

Therefore, based on previous experiments, the next experiments in this thesis are based mainly on using SR based GMM.

5.3 The robustness of SR using noisy speech Training

This section describes an attempt to improve speaker recognition robustness in a noisy environment by including noises with different controlled SNRs in the enrolment phase of typical Gaussian Mixture Model based speaker recognition systems. Validation testing was carried out in emulated noisy conditions with a controlled signal to noise ratios SNR. The aim of these experiments was to characterise the effectiveness of using noisy speech samples contaminated with different types of noise with different signal to noise ratios (SNRs) on the accuracy of speaker recognition, and what kind of improvement was obtained when using noisy speech samples in enrolment phase.

5.3.A Experimental setup

The experiments on noisy speech were based on using three different scenarios: the first one was based on using speech samples contaminated with five types of noise (cafeteria speech babble, interior moving car noise, interior moving train, street noise, and white noise) in enrolment phase (Training phase). These enrollment speech samples were contaminated with fixed signal to noise ratio (15 dB in this case) for each type of noise. Results were then compared with those obtained from clean speech samples training. The second scenario was based on using speech samples contaminated with different SNRs of Cafeteria speech, Babble noise, and interior moving train noise (in this case 4 types of SNR were used: 15, 10, 5 and 0 dB). The third scenario included using speech samples contaminated with a specific type of noise (in this experiment pink noise was used) in the enrollment phase (with 15dB SNR) and investigated the performance of SR with different types of noisy recognition data. The details of these experiments are as follows:

- Mel-frequency Cepstral coefficients MFCC with delta and delta-delta coefficients were used in these experiments with dimensions 39*514 where the row represents the number of Mel-filters and column represent the number of MFCC frames (more details about MFCC extraction in Chapter Five).
- Gaussian mixture model GMM was applied for creating reference models and classification with 256 Mixture.
- The experiments were conducted on speech samples from 60 speakers (30 male and 30 female) in the SALU-AC database.

- One set of speech samples was used. This set was divided into a number of subsets. Each subset represents the same speech samples contaminated with different SNRs (20, 15, 10, 5, and 0 dB) for a specific type of noise. In these experiments, five types of noise were used (Cafeteria speech babble, Interior moving car noise, interior moving train, Street noise, and white noise).
- The evaluation of the MSR speaker recognition toolbox, in these experiments, is based mainly on using Error Equal Rate (EER) and Detection Error Tradeoff (DET).

5.3.B Experimental Results and Discussion

In the first scenario, Figure 5.5 a, b, c, d and e illustrate the differences that occur in degradation in recognition accuracy (based on EER) for each sample of noisy and clean enrollment speech when SNR is decreased. As discussed before, the abscissa represents the signal to noise ratio SNR in dB, the ordinate represents the recognition accuracy for SR based on EER, and the table included with each line graph represents the values of ERR for each SNR.

Regarding additive babble noise (Figure 5.5 a), the accuracy of SR shows greater noise robustness when noisy speech samples are used in the enrolling phase (training phase) compared with using clean speech samples for enrolling. In particular, when the SNR becomes low (i.e., between 10 and 0 dB) the difference in accuracy between both cases (clean and noisy) is increased. For example, at 15, 10, and 5 dB, the EERs for Babble noisy enrolment are 0.113%, 0.59%, and 3.92% respectively, with only 1.66%, 5.39% and 13.33% respectively for clean speech enrolling. Likewise, the additive interior train and Street noise show the same improvement in noisy training. However, a significant improvement can be demonstrated between 10 and 5 dB, with EER 1.6% and 3.33% respectively for interior train noisy enrolment against 6.66% and 14.57% EER respectively for clean enrolment. Street noisy enrolment demonstrated EER 0.48% and 1.66% respectively compared to 3.33% and 10% EER respectively for clean training. On the other hand, the interior car noise (Figure 5.5 b) shows the best improvement at 5 and 0 dB while the white noise samples also showed very significant improvement between 15dB and 5dB (Figure 5.5 e). In addition, the experiments discovered, for some noise, that the best results were obtained when the enrolment and recognition samples have the same SNR (in this case 15 dB) as seen in cafeteria babble, interior car, and white noisy speech sample. However, this cannot be generalised for all types of noisy speech training.

Conversely, using noisy speech in the training phase with clean speech recognition samples in the recognition phase causes degradation in performance.

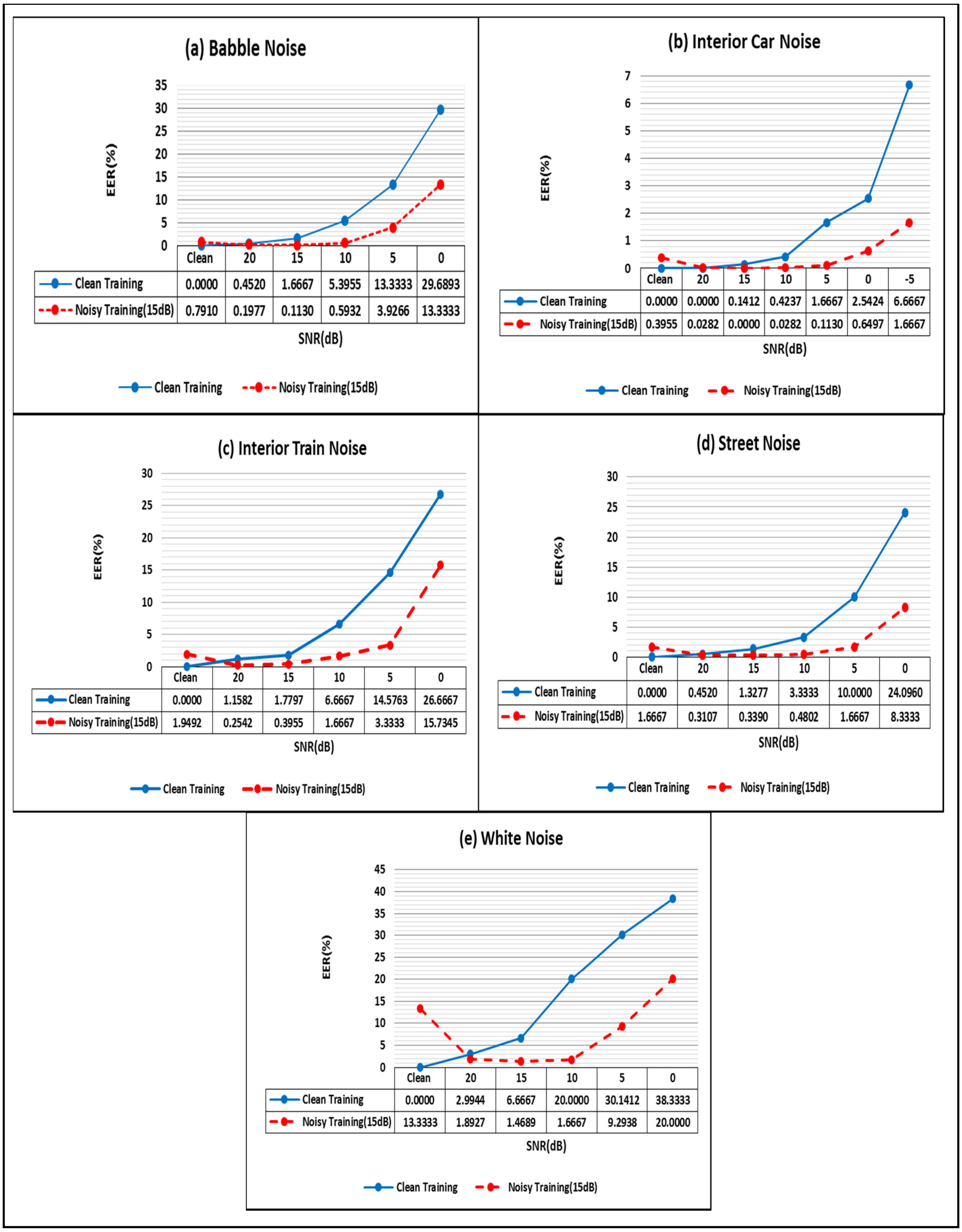


Figure 5.5: EER of SR using Noisy Speech Training

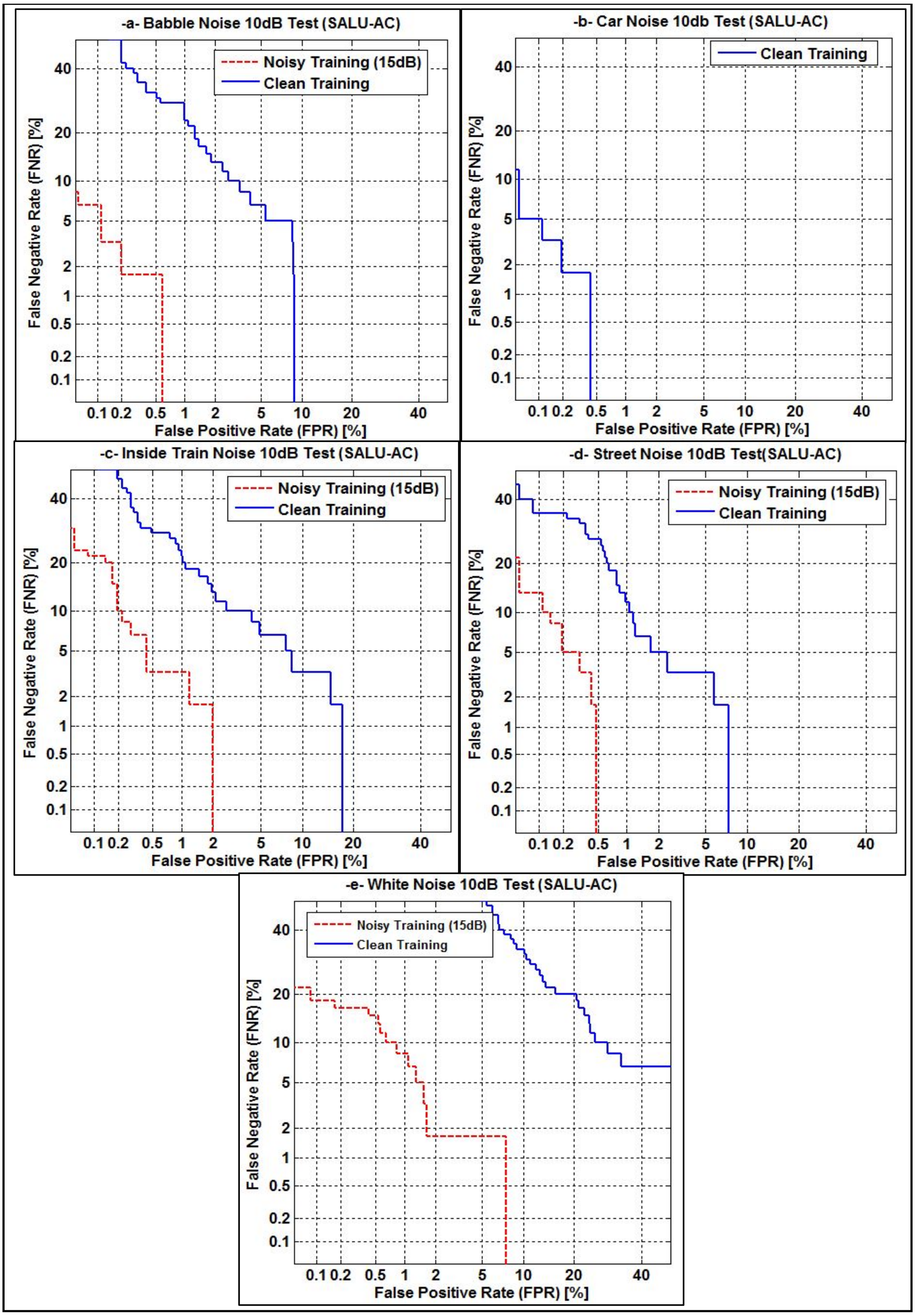
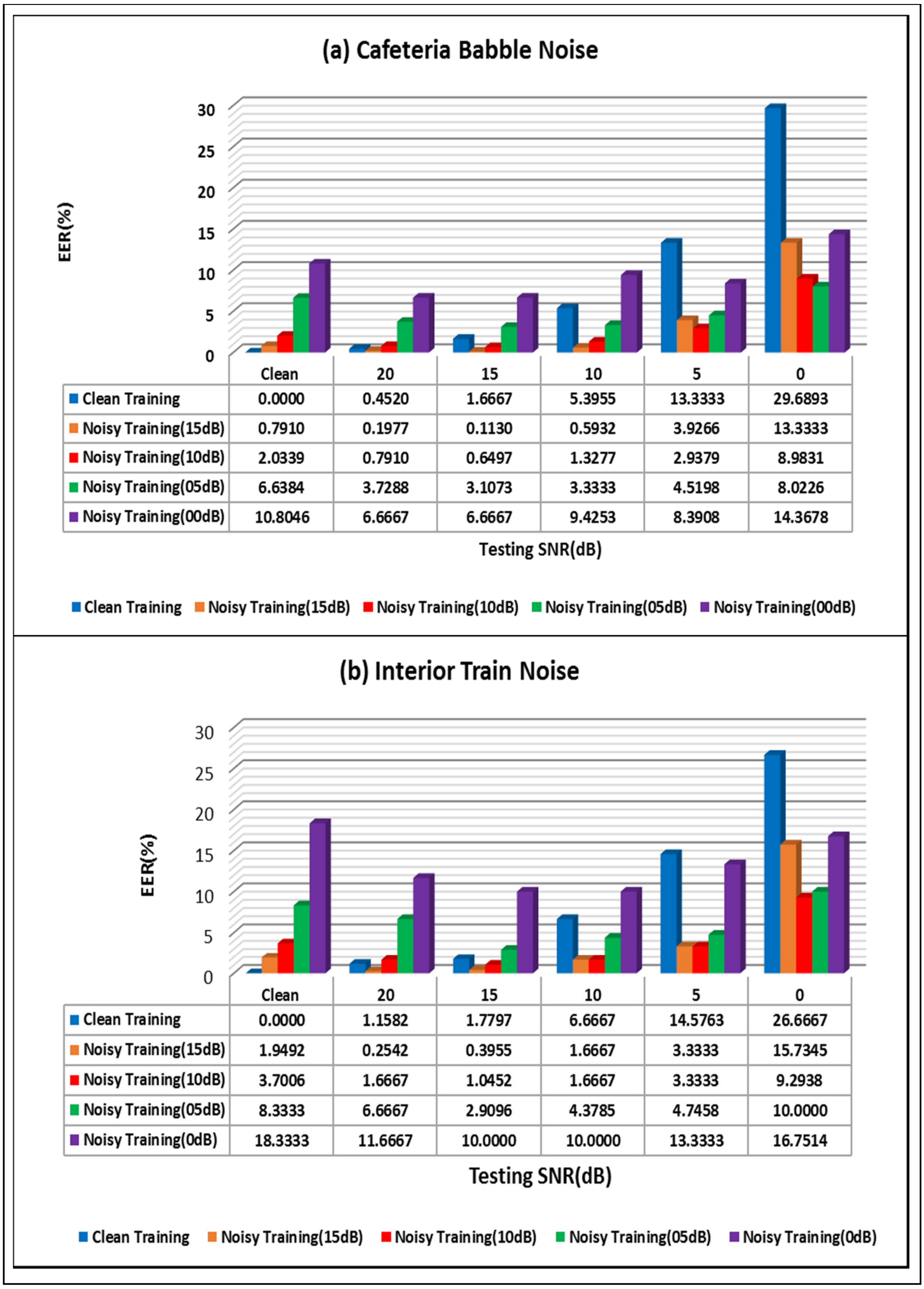


Figure 5.6: DET Graphs for Noisy and Clean Training at 10dB recognition data

Figure 5.6 a-e represent the detection error trade-off (DET) graph for clean and noisy training for noisy recognition speech samples contaminated with 10 dB SNR. The red dashed curve represents the noisy training accuracy of SR, while the solid blue one represents the accuracy of clean training. It is noticeable that the improvements obtained from noisy training in both false positive rate FPR and false negative rate FNR are high compared with clean training for all types of noisy speech training. Note that the curve of noisy training for interior car noise is not shown since the EER, in this case, is too small (0.02 %). Furthermore, it can be seen that the improvement of white noise speech training is the highest.

More details about DET graphs for the other SNRs are illustrated in Appendix -I-.

In the second scenario, as mentioned before, noisy speech samples contaminated with different SNRs (15, 10, 5, and 0) of specific types of noise (in this case cafeteria babble noise and interior train noise) were applied in the enrolment phase. Figure 5.7 a, b demonstrates the EER of each SNR training (orange, red, green, and purple bars for 15, 10, 5, and 0 dB respectively) in addition to clean training (which is represented as a blue bar). Compared with clean speech training, it is clear that babble noisy speech training at 15, 10, and 5 dB outperforms the clean speech training, especially when SNRs of recognition samples (test sample) are decreased. For example, in cafeteria babble noisy speech training, when the recognition speech samples (test speech samples) are contaminated with 10 dB SNR, the best performance was obtained in 15dB and 10 dB noisy training with 0.59% and 1.3 % EER respectively compared with 5.39% EER for clean training (Figure 5.7 (a)). The same result can be seen when the interior train noisy samples are applied in the enrolment phase.



(a) Cafeteria Babble Noise

	Clean	20	15	10	5	0
■ Clean Training	0.0000	0.4520	1.6667	5.3955	13.3333	29.6893
■ Noisy Training(15dB)	0.7910	0.1977	0.1130	0.5932	3.9266	13.3333
■ Noisy Training(10dB)	2.0339	0.7910	0.6497	1.3277	2.9379	8.9831
■ Noisy Training(05dB)	6.6384	3.7288	3.1073	3.3333	4.5198	8.0226
■ Noisy Training(00dB)	10.8046	6.6667	6.6667	9.4253	8.3908	14.3678

(b) Interior Train Noise

	Clean	20	15	10	5	0
■ Clean Training	0.0000	1.1582	1.7797	6.6667	14.5763	26.6667
■ Noisy Training(15dB)	1.9492	0.2542	0.3955	1.6667	3.3333	15.7345
■ Noisy Training(10dB)	3.7006	1.6667	1.0452	1.6667	3.3333	9.2938
■ Noisy Training(05dB)	8.3333	6.6667	2.9096	4.3785	4.7458	10.0000
■ Noisy Training(0dB)	18.3333	11.6667	10.0000	10.0000	13.3333	16.7514

Figure 5.7: EER of Speaker Recognition for clean and different Noisy training SNRs

The performance of SR using noisy training showed significant improvement over the clean training, especially when the recognition speech samples were at 10, and 5 dB SNR. For instance, when the recognition sample was contaminated with 5 dB SNR less EER was obtained than when SR was trained in 15 dB and 10 dB SNRs, with 3.33% for both cases, while in the case of clean training the EER was 14.5. However, the training in 0 dB SNR (where the level of noise and speech was the same), for both type of noise, did not give any improvement even when the recognition speech data had the same rate. Similar to the first scenario, the accuracy of using noisy speech training with a clean speech in the recognition phase shows degradation in the accuracy compared with using clean speech in enrolment phase.

Moreover, the experiments show that when recognition speech samples are contaminated with high SNRs (e.g. 15 dB) the best results can be obtained when enrolment noisy speech is trained with the same signal to noise ratio (i.e. when channels are matching between testing and enrollment phases). However, for recognition samples with low SNRs (e.g. 10, 5 and 0 dB) it is better to use a higher SNR than that in recognition samples, as seen in Figure 5.7 a, b.

Figure 5.8 shows the DET Graph for clean and noisy training with different SNRs for 05 and 10 dB speech samples contaminated with cafeteria babble noise (Figure 5.8 a, c) and interior Train noise (Figure 5.8 b, d). The bold black curve represents the clean training, while the remaining curves represent different SNRs training. Again it is clear that noisy training for 15, 10, and 5 dB outperformed the clean training in both FPR and FNR and with the two types of noise. Moreover, for the noisy recognition speech at 10dB, the best performance was obtained when training at 15dB for both types of noise (Figure 5.8 c, d); while for the 5dB noisy recognition speech set, the best performance is at 10dB training in case of cafeteria babble noise and 15db training in case of interior moving train. More details about DET graphs for the other SNRs are illustrated in Appendix -I-.

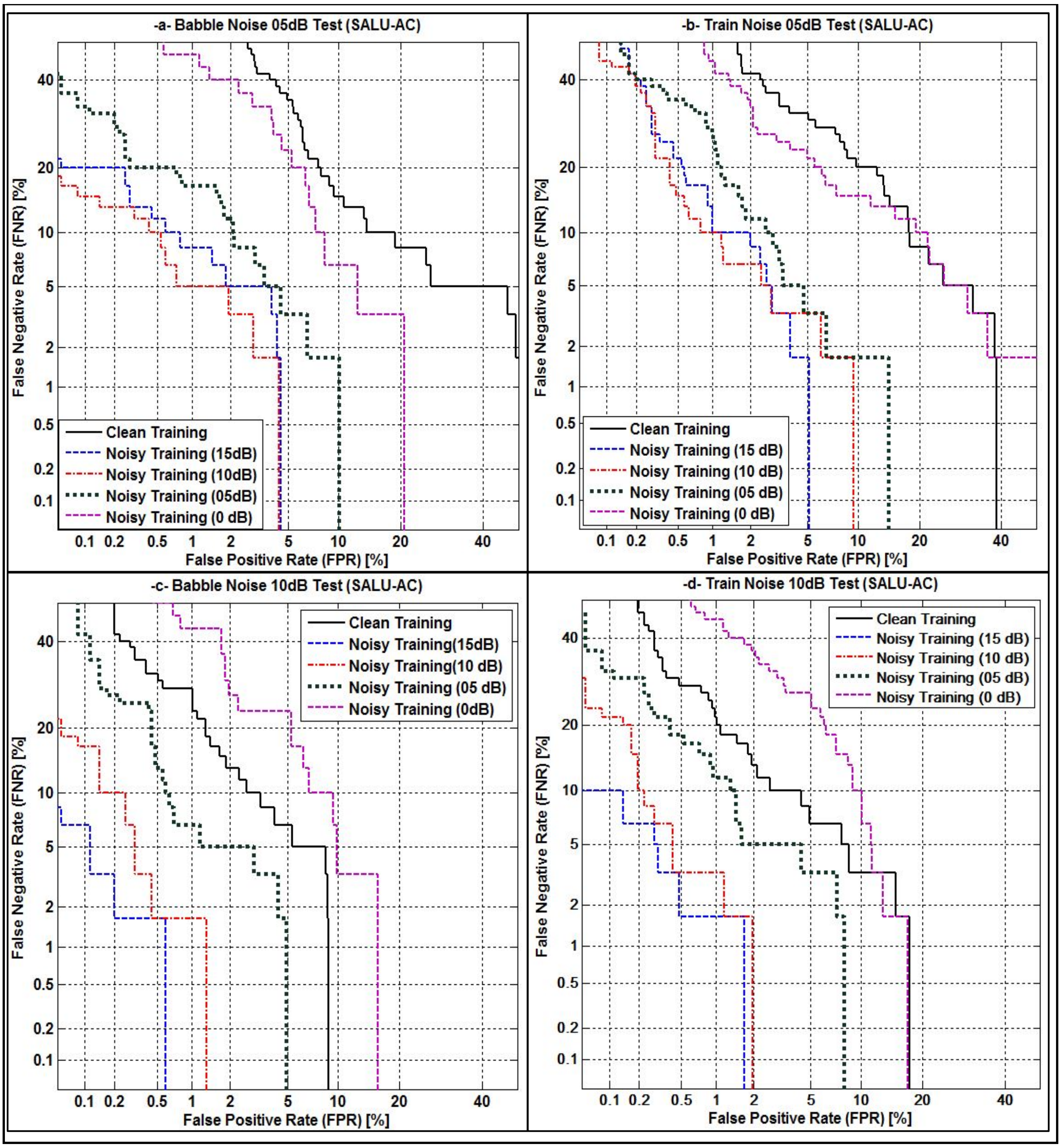
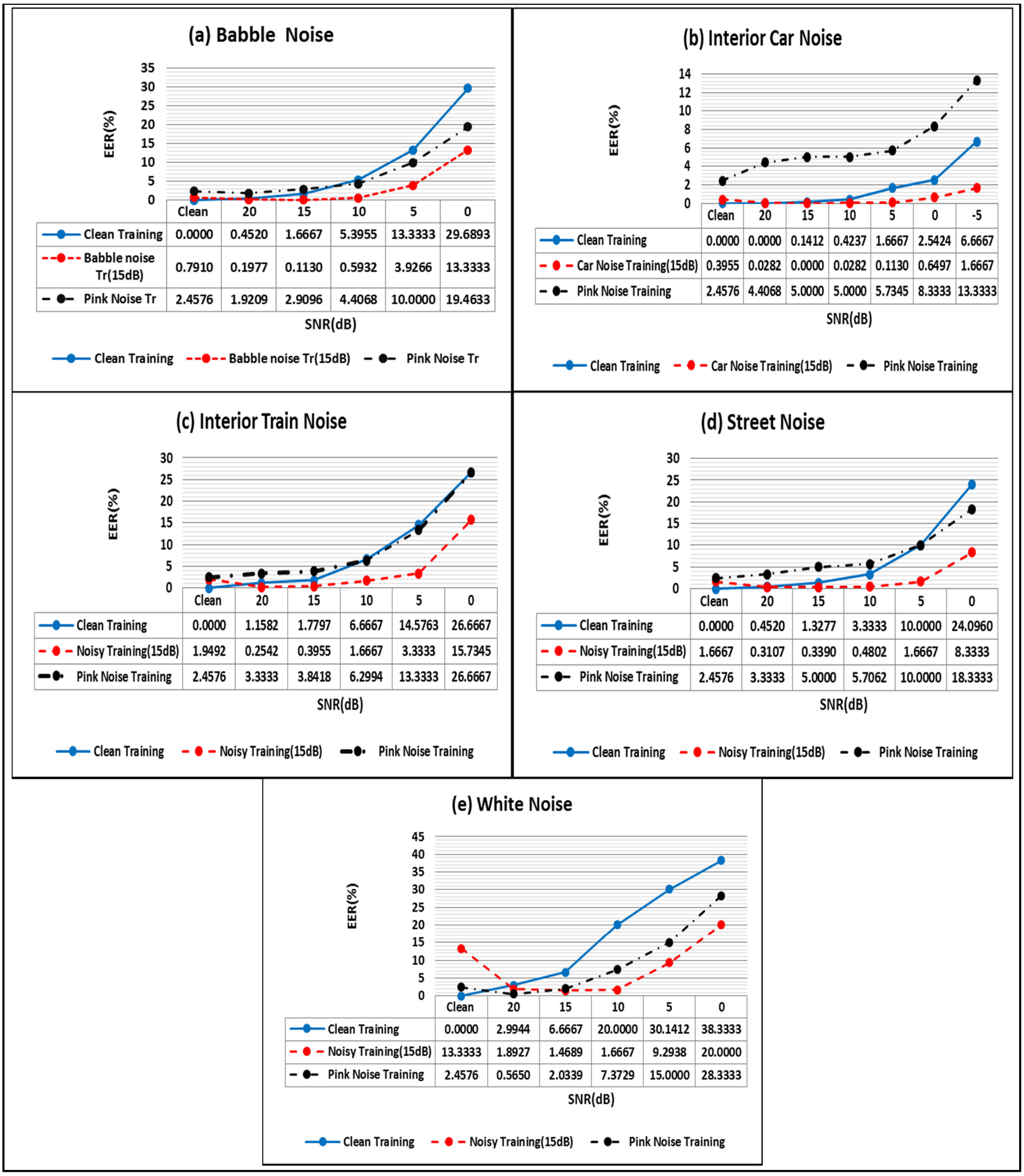


Figure 5.8: DET Graph for Clean and Different SNRs Training



(a) Babble Noise

	Clean	20	15	10	5	0
Clean Training	0.0000	0.4520	1.6667	5.3955	13.3333	29.6893
Babble noise Tr(15dB)	0.7910	0.1977	0.1130	0.5932	3.9266	13.3333
Pink Noise Tr	2.4576	1.9209	2.9096	4.4068	10.0000	19.4633

(b) Interior Car Noise

	Clean	20	15	10	5	0	-5
Clean Training	0.0000	0.0000	0.1412	0.4237	1.6667	2.5424	6.6667
Car Noise Training(15dB)	0.3955	0.0282	0.0000	0.0282	0.1130	0.6497	1.6667
Pink Noise Training	2.4576	4.4068	5.0000	5.0000	5.7345	8.3333	13.3333

(c) Interior Train Noise

	Clean	20	15	10	5	0
Clean Training	0.0000	1.1582	1.7797	6.6667	14.5763	26.6667
Noisy Training(15dB)	1.9492	0.2542	0.3955	1.6667	3.3333	15.7345
Pink Noise Training	2.4576	3.3333	3.8418	6.2994	13.3333	26.6667

(d) Street Noise

	Clean	20	15	10	5	0
Clean Training	0.0000	0.4520	1.3277	3.3333	10.0000	24.0960
Noisy Training(15dB)	1.6667	0.3107	0.3390	0.4802	1.6667	8.3333
Pink Noise Training	2.4576	3.3333	5.0000	5.7062	10.0000	18.3333

(e) White Noise

	Clean	20	15	10	5	0
Clean Training	0.0000	2.9944	6.6667	20.0000	30.1412	38.3333
Noisy Training(15dB)	13.3333	1.8927	1.4689	1.6667	9.2938	20.0000
Pink Noise Training	2.4576	0.5650	2.0339	7.3729	15.0000	28.3333

Figure 5.9: Pink Noisy Training compared with other types of Training

The third scenario includes applying a noisy enrollment speech contaminated with a specific type of noise (the pink noise was chosen) at 15dB and investigating the performance of SR with different sets of recognition noisy speech contaminated with different types of noise and different SNRs. The main goal of these experiments was to answer the following question: can an improvement in SR be obtained by using noisy speech data contaminated with a fixed type of noise in the enrollment phase for all types of recognition speech sample contaminated with different types of noise?

Figure 5.9 illustrates the degradation of EER for pink noisy speech samples training (the black dashed & dotted line) compared with clean samples training (the blue bold line) and speech samples training using the same noise already used in the test sample (the red dashed line).

From this figure, it is observed that using speech samples contaminated with the pink noise in training phase causes degradation in performance of SR for four types of testing noisy data (Cafeteria Babble, interior car, interior train, and street noise). In some cases, the EER of SR even becomes worse than using the clean sample for training (Figure 5.9 a, b, c, and d). On the other hand, when this training was used with test samples contaminated with white noise, a promising improvement in accuracy was found, compared with clean samples training (Figure 5.9 e). However, this improvement is still limited compared with white noise sample training which produces a significant improvement in accuracy when compared with other training types (pink and clean training). Figure 5.10 shows the DET graph for the test samples contaminated with 15 dB SNR for the five types of noise using the three types of training.

According to Figure 5.10, the training with pink noisy sample affects both FPR and FNR (the dash-dot curve) negatively for the four types of noisy test sample mentioned before, while there is an improvement with noisy test samples with white noise. However, the best results still come from using the same type of noise as used with the test sample.

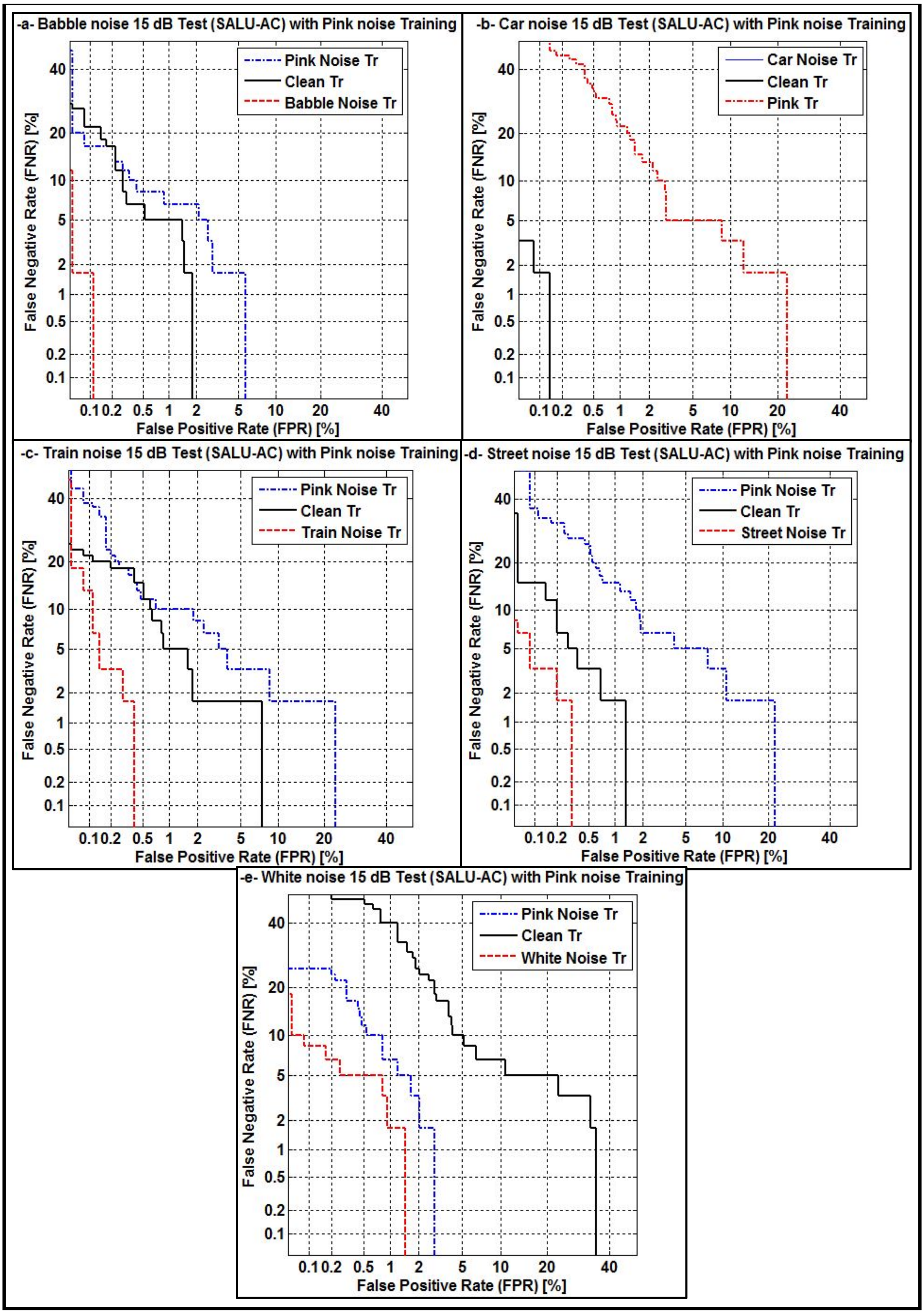


Figure 5.10: DET graph of pink noisy training for 15 dB test samples

From the results of this scenario, it can be concluded that each type of noise has a unique profile and unique characteristics that make it different from other kinds of noise. Therefore noisy reference models cannot be generalised based on one type of noise for other type speech sample data contaminated with a different type of noise.

From the last three scenarios, the findings can be summarised as follows:

- Using noisy speech samples in a training phase gives significant improvement in performance of a speaker recognition system compared with clean speech training, especially when the system is used in noisy environments.
- To provide suitable noisy models that closely resemble the noisy input signal, two factors should be taken into account: signal to noise ratio SNR that matches or is at least close to SNR of the input signal, and a noise profile which includes the same characteristics that are embedded in the input signal.
- It is very necessary to specify SNR for noisy speech training, since the investigations from the last scenarios showed that, to ensure a good SR performance, the noisy speech training must have the same SNR or at least close to that included in the recognition signal if SNR is high (e.g. 20, and 15 dB), while for low SNR it is better to provide noisy training samples with SNR greater than that included in the recognition signal. However, this still differs from one noise to another.
- Using noisy reference samples contaminated with one type of noise with recognition speech samples contaminated with a different type of noise in order to improve the robustness of speaker recognition is not a good solution (sometimes this works with some kinds of noise but not for all noises) ,since each type of noise has its own unique characteristics.

As a result, if suitable SNR and noise profile estimations can somehow be provided, it is possible to re-build a noisy adaptive reference which comes closer to matching the noisy input signal to improve the performance of speaker recognition in noisy environments. In chapter 8 a new approach has been proposed to make the speaker recognition more robust in noisy environments, namely “Training on the fly” based on creating a noisy adaptive model dependent on SNR and noise profile.

5.4 Chapter Summary

This chapter can be divided into two parts.

The first part included an investigation of robustness of using GMM and GMM-UBM based SR with limited speech data and performance of each based system in noisy environments and mismatched languages.

The second part included investigations made to deal with the degradation in speaker recognition caused by environmental noises. This investigation involved enrolment with various types of noisy speech with different SNRs. Results show using noisy speech samples in a training phase gives significant improvement in performance of a speaker recognition system compared with clean speech training.

These experiments suggest that inclusion of noisy environmental conditions in the enrolling can mitigate the performance degradation to some extent, but this works best when channels are perfectly matched or at least close. This further suggests that if acoustic conditions can somehow be estimated and suitably pre-trained models chosen, the robustness of the system can be improved.

CHAPTER 6
COMPARISON OF FEATURE SPACES FOR ROBUST SPEAKER RECOGNITION AND MISMATCHED LANGUAGES

Chapter Overview

Chapter Two presented a review of the most common features extraction algorithms, which focused mainly on Cepstral features, since they are the most commonly used in the speaker recognition field. This chapter is focused on making a systematic comparison between the robustness of two kinds of features extraction algorithms in the performance of speaker recognition. Conventional Mel-frequency Cepstral coefficients (MFCC) - including the Dynamic Mel cepstral - were adopted, together with Gamatone Cepstral Coefficients (GFCC). Furthermore, the current work extends the experiments to include investigations into language independency features in recognition phases. The experiments used an MSR toolbox with two types of database (TIMIT, and SALU-AC) and two types of recognition data sets (test data). The first type was based only on using a text-independent English speech dataset for recognition purposes in various kinds of noisy environments with different SNRs. The second type was based on applying a text-independent noisy speech data set from various languages for recognition purposes in the same noisy environments. The chapter starts with a description of Extraction algorithms for each MFCC and GFCC and the main differences between the two features. The experiment details are given in Section 6.2, and results and discussion in Sections 6.3 and 6.4 respectively. This work is published by the author at 140th International AES Convention, Paris, France (Al-Noori et al., 2016).

6.1 Features Extraction Process

As discussed in Chapter Two, Low-level features (specifically Cepstral features) are the most appropriate and widely adopted speech representation for applying in speaker recognition fields since they are easy to calculate, and have a good performance compared with other types of features. Generally, most applications of speaker recognition utilise a front-end processing section in order to characterise the speech signal in this way. This section focuses on processes of extracting features of both MFCC and GFCC.

6.1.A Mel-Frequency Cepstral Coefficients MFCC

The Mel-Frequency Cepstral Coefficients are one of the most commonly used features in both speech and speaker recognition (Davis and Mermelstein, 1980) since they are a close representation of the human speech signal. The Mel-Scale in MFCC has been used to approximate the human auditory system (Kim et al., 2006). Figure 6.1 illustrates the process of MFCC extraction. The main steps of MFCC extraction are as follows:

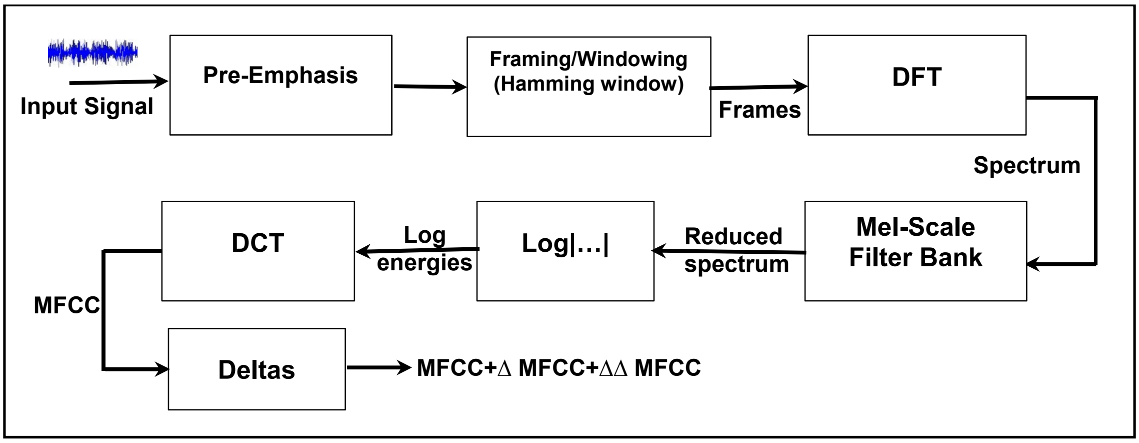


Figure 6.1: Extraction process of MFCC

1. Pre-emphasis: the input speech signal is pre-emphasized in order to flatten the speech spectral slope. This spectral slope represents the nature of the audio signal. Pre-emphasis represents a high pass filter responsible for amplifying the energy at high frequencies and decreasing the difference in the power signal's components (Combrinck and Botha, 1996). The transfer function of this filter is a first order function:

$$H_p(z) = 1 - \alpha z^{-1} \;, 0.9 \leq \alpha \leq 1.0 \qquad (6.1)$$

where the α is chosen to be 0.95.

2. Framing and Windowing: The aims of this process are slicing the speech signal into overlapping frames and reducing any signal discontinuities. The process of windowing is done by taking the samples of the signal and multiplying by a window function. *Hamming window* is most commonly (Figure 6.2) used in speaker recognition fields since it prevents any sharp edge like rectangular windows (Beigi, 2011). To trade-off the spectral and temporal resolution a 20-30ms window has typically been used. The equation of Hamming windows is:

$$w[n] = \begin{cases} 0.54 - 0.46 \cos\frac{2\pi n}{L} & 0 \leq n \leq L-1 \\ 0 \end{cases} \qquad (6.2)$$

3. The Discrete Fourier Transform (DFT): To convert the speech signal from time domain to frequency domain the DFT has been used. The DFT can be defined as:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i\frac{2\pi kn}{N}} \qquad (6.3)$$

where N is a number of samples, x_n is the signal at sample n , X_k is fourier tranform signal

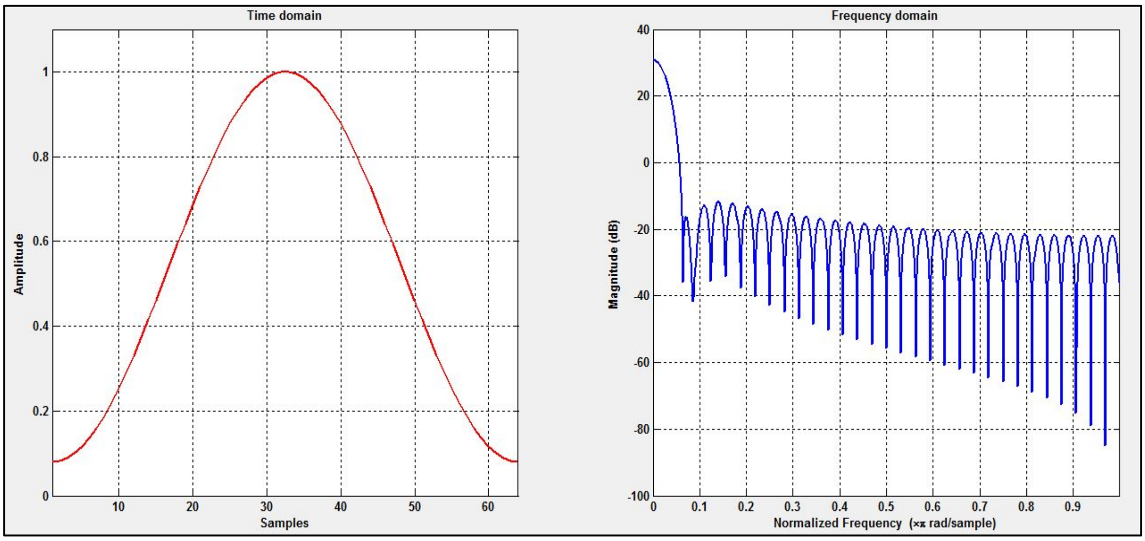


Figure 6.2: The Hamming window in time and Frequency Domain

4. Mel filter-bank: in order to wrap the magnitude spectrum into Mel-spectrum, a Mel –filter bank which effectively maps the frequency bin centre has been applied in this step. The main goal of this step is to approximate the frequencies to human auditory perception because the DFT computations only pertain to a linear frequency scale (Burgos, 2014). Therefore, the logarithmetic Mel Scale has been used to convert spectrum frequencies to a smaller number with a process called frequency wrapping (Equation 6.4). Figure 6.3 illustrates the shape of 24-filter Mel-filter bank sampled at 8000Hz

$$\boldsymbol{mel(f) = 1127l\,n(1+\frac{f}{700})} \qquad \textbf{(6.4)}$$

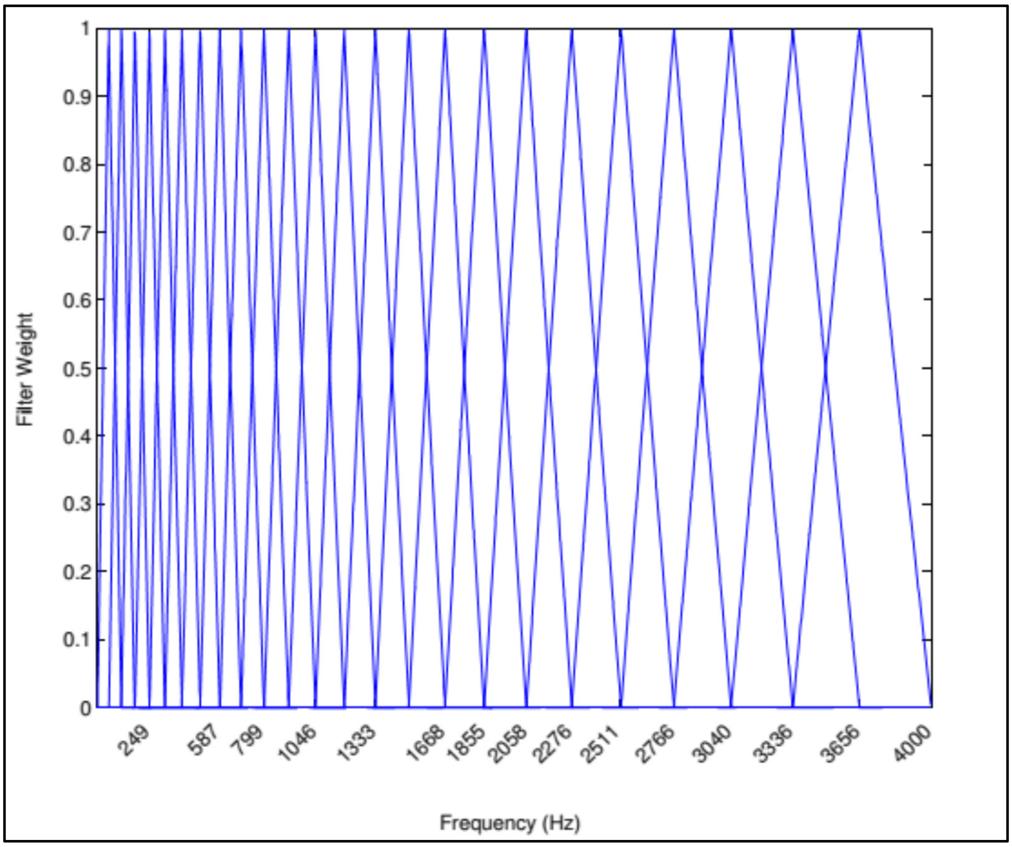


Figure 6.3: Mel filter bank for 24-filter with 8 kHz sample rate (Beigi, 2011)

5. Discrete Cosine Transform (DCT): this step is responsible for transforming the spectral information to Cepstral domain in order to produce MFCC, which contain the meaningful information and offer the particular characteristic of the speaker.
6. Dynamic MFCC (Delta and Delta Delta): in order to model the local dynamics of speech and make the extracted feature more robust to interfering noise, Wang et al. (2008 and 2009) proposed measuring the first and second order derivatives of the MFCC to capture the dynamic of Mel Cepstral Coefficients. These dynamic features, known as delta MFCC (ΔMFCC) and Delta Delta (ΔΔ MFCC), are somewhat independent of conventional MFCC.

As a result, the total number of final features for the speaker vector is 39 including the energy of MFCC and Dynamic MFCC.Table 6-1 illustrate the total extracting features of MFCC in standard speaker.

Table 6-1: Total Number of MFCC feature for Speaker Vectors

Feature Type	Number
Mel Cepstral Coefficients	12
Delta Mel Cepstral Coefficients	12
Delta Delta Mel Cepstral Coefficents	12
Energy Coefficients	1
Delta Energy Coefficients	1
Delta Delta Energy Coefficients	1

6.1.B Gammatone Frequency Cepstral coefficients (GFCC)

Gammatone Frequency Cepstral coefficients (GFCC) are auditory features based on the Gammatone Filter bank. This filter bank represents a standard model of the cochlear filtering of human ears, which is derived from psychological and physiological monitoring of auditory peripheral (Patterson et al., 1992). The output of a Gammatone Filter Bank can be represented as a time-frequency of the signal which is known as a *Cochleagram*. To extract GFCCs from the speech signal, the following steps should be followed: Figure 6.4 shows the block diagram of GFCC extraction.

1. Pre-Emphasis: as in MFCC, the input speech signal is pre-emphasised by a Pre-emphasis filter (high pass filter) in order to spectrally flatten the speech signal. According to Zhao and Wang (2013), this can be ignored.

2. Framing and windowing: dividing the signal into overlapping frames. The Hamming window is then applied to deal with the edges of the frame.

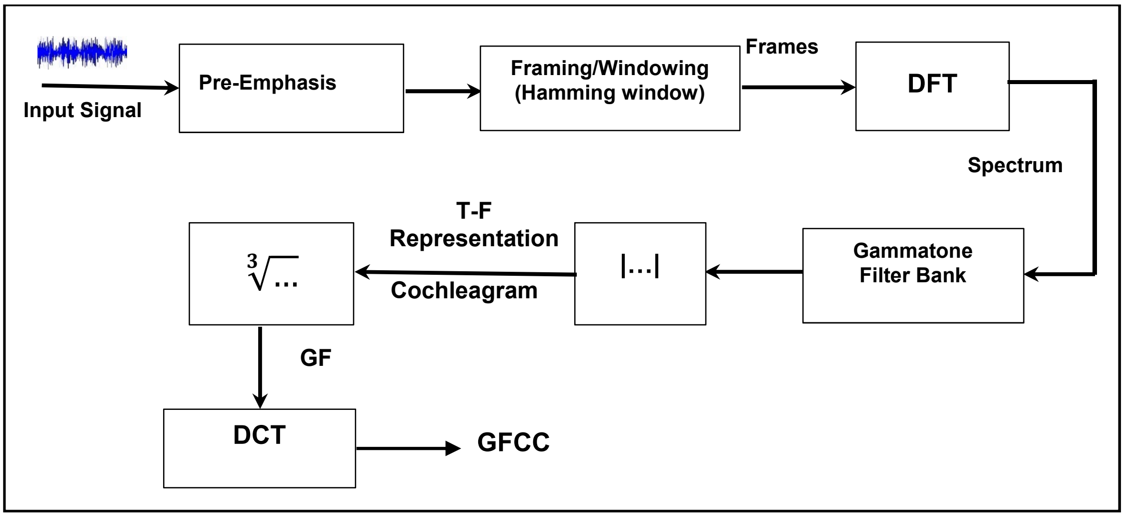


Figure 6.4: Block Diagram of GFCC features extraction

3. DFT: transforms the windowed frame of the speech signal from time domain to frequency domain.
4. Gammatone Filter Bank: the design of Gammatone Filter-Bank is based on the operation of the human auditory system. This filter bank consists of 64 filters, centred from 50 Hz to 8 kHz distributed equally on an *Equivalent Rectangular Bandwidth* (ERB) scale (Patterson et al., 1992) and the filters with higher centre frequencies having wider bandwidth(Figure 6.5). The impulse response of each filter centred at frequency f is represented by the following equation (Shao et al., 2007, Shao and Wang, 2008):

$$\boldsymbol{g(f,t)} = \begin{cases} \boldsymbol{t^{(a-1)} e^{-\pi bt} \cos(2\pi f t)} & \boldsymbol{, t \geq 0} \\ \boldsymbol{0} & \boldsymbol{, t < 0} \end{cases} \qquad \textbf{(6.5)}$$

where $\boldsymbol{t}$ refers to time ; $\boldsymbol{a} = 4$ is the order of filter; $\boldsymbol{b}$ is rectangular bandwidth which increase with centre frequencies $\boldsymbol{f}$

5. For each channel, the filter response is decimated to 100 Hz along the time dimension. The absolute value is taken afterwards.
6. The magnitudes of output sample are warped using a cubic root operation such as:

$$\boldsymbol{G_m[i]} = \sqrt[3]{|\boldsymbol{g_{decs}[i,m]}|} \quad , \qquad \textbf{(6.6)}$$

where $i = 0, .., N - 1$, N is number of filter channels, $m = 0, .., M - 1$. m is the frame index; M is number of frames obtained after decimate sampling, and *decs* is decimate sample.

The result of $G_m[i]$ is in the form of a matrix representing time-frequencies (T-F) of input. This (T-F) representation refers to the differences of cochleagram (Wang and Brown, 2006). The time frame $G[i]$ is known as Gammatone feature (GF). In some studies (Moinuddin and Kanthi, 2014) the cubic root operation is replaced by logarithm operation. In this work the cubic root operation has been used.

7. Discrete cosine transform is applied to GF in order to decrease the dimension of GF and de-correlate the component. The coefficients produced from this operation are known as Gammatone Cepstral Coefficients GFCC.

As a result, the total of number of GFCC features for each speaker vector is 23.

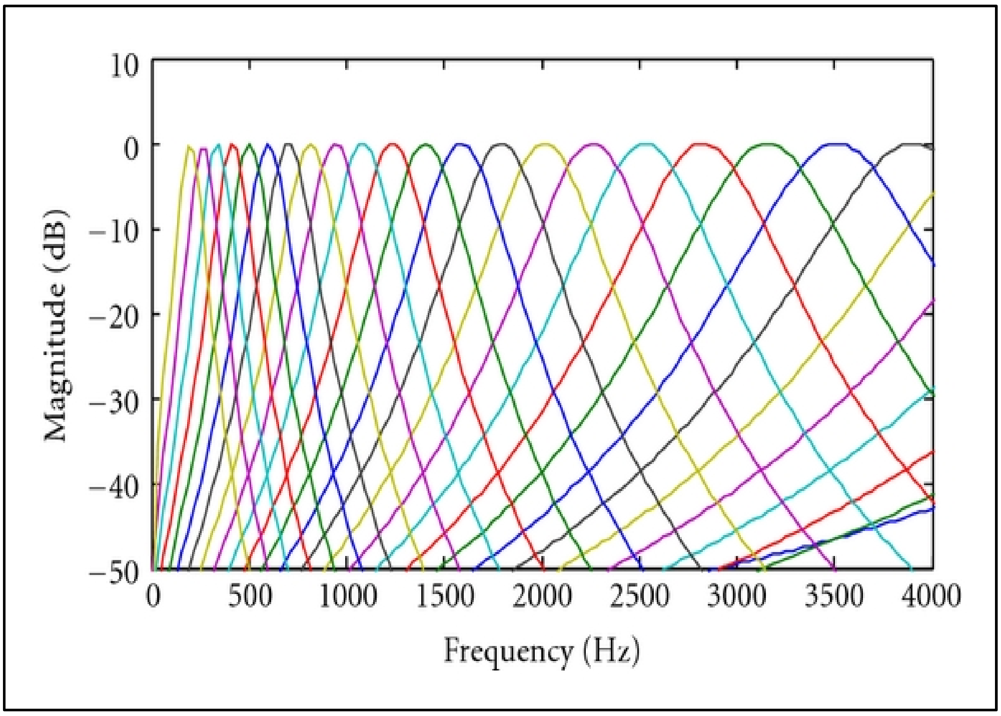


Figure 6.5: Frequency response of Gammatone filter bank

Table 6-2 shows the main differences between Mel Frequency Cepstral Coefficients MFCC and Gammatone Frequency Cepstral Coefficients GFCC.

Table 6-2: The Main Differences between MFCC and GFCCC

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

6.2 Robustness SR Based on MFCC and GFCC from noisy speech samples and Mismatched Languages

What will be the overall performance of these two features based speaker recognition systems? A comparative study is justified, hence these experiments. In addition to the comparisons of recognition accuracies in various signal to noise ratios and a variety of noises, system robustness against language mismatch is also investigated in this section.

6.2.A Experimental Setup

These experiments present a systematic comparison of robustness between the GFCC and conventional MFCC baseline speaker recognitions with various collected noisy speech datasets. The experiments were conducted using two data sets, namely the TIMIT and SAL-AC. For each speaker, nine utterances were used in the enrolment phase and one utterance for the recognition phase. Two data sets of recognition speech were used in SALU-AC. The first set included samples of utterances spoken in English. The second set included samples of utterances spoken in different languages (including English).

The aim of using two different sets of samples was to evaluate the performance of speaker recognition based on two kind of features (MFCC, GFCC) when the languages of the recognition dataset used in the recognition phase are different from those that have been used in the enrollment phase (note that only English speech data sets are used in the enrolment phase), and to evaluate the sensitivity of both MFCC and GFCC to changes of language. It was also intended to evaluate the robustness of speaker recognition systems based on those features in clean and noisy environments. The details of these experiments were as follows:

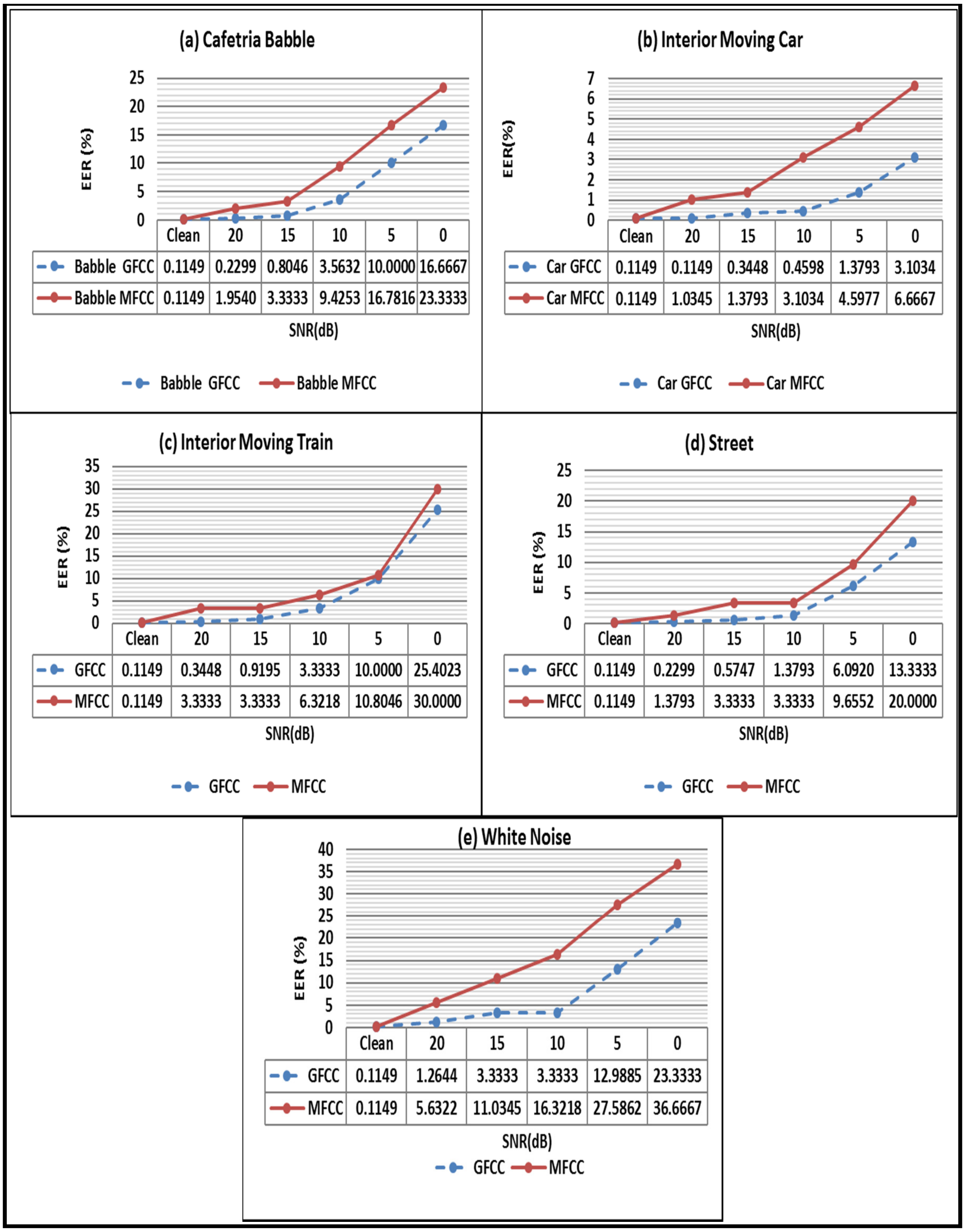
- The dimension of the GFCC features in these experiments was 23 for each speaker vector while the dimensional of the baseline MFCC was 39 for each features vector including the dynamic features as demonstrated in Table 6-1.
- Gaussian mixture model GMM was applied for creating reference models and classification with 256 Mixture.
- As mentioned earlier, the experiments were conducted on speech sample sets from two types of database TIMIT and SALU-AC. Each set included speech samples from 30 speakers (15 male and 15 Female). For each speaker, 9 utterances were used in the training phase, all in English (for both TIMIT and SALU-AC database). In addition, 1 utterance was used in the test phase for each speaker. For the SALU-AC database, two datasets of speech samples were used. The first set included utterances spoken in English, while the second set contained utterances spoken in various languages (Arabic, French, Indian, Italian, Persian etc.) in addition to English. Furthermore, each data set was divided into a number of subsets, each subset including the same utterances contaminated with different SNRs (20, 15, 10, 5, and 0dB). Table 6-3 illustrates the languages of utterances used in the recognition phase.
- The evaluation of the MSR speaker recognition toolbox was based mainly on using Error Equal Rate (EER) to evaluate the performance of the SR baselines, as well as Detection Error Tradeoff (DET). These evaluations are measured mainly based on the recognition scores (log-likelihood score) based on 30 authentic cases (verified cases) and 870 inauthentic cases (imposter cases).

Table 6-3: Recognition (test) speech sample using different languages

Language	Male Speakers	Female Speakers	Total speakers
English	7	1	9
Arabic	2	6	7
Itally	1	2	2
Spanish	0	2	2
Portuguese	0	1	1
Polish	0	1	1
Bangladeshi	2	0	2
Swahili	0	1	1
Hindi	1	1	2
France	2	0	2
Total	**15**	**15**	**30**

6.2.B Experimental Results

As described earlier, the evaluation of speaker recognition used in these experiments depended on using two metrics: EER, and Detection Error Trade-off. Figure 6.6 a to e show the impacts of various types of noise on the accuracy of speaker recognition based on GFCC and MFCC features using the TIMIT database. The x-axis represents signal to noise ratio SNR in dB (clean,20dB to 0dB) while the y-axis represents EER (in %). According to this figure, the performance of the GFCC-baseline shows greater robustness to the various noises with different SNRs compared with the performance of the MFCC baseline, especially when SNR becomes too low. For instance, in speech samples contaminated with Cafeteria babble noise (Figure 6.6 -a-), there is a significant improvement in EER for the GFCC baseline for all SNRs, especially at 10, 5, and 0 dB SNR with 3.56%, 10%, and 16.66 EER respectively, compared with only 9.42%, 16.78%, and 23.33% for the MFCC baseline for the same SNRs. Results for speech samples contaminated with Interior Moving Car noise, Street noise, and White noises show similar trends (Figure 6.6- b, c, d, and e) for all SNRs, specifically for 10, 5, and 0 dB SNR. For Interior Train noise speech samples (Figure 6.6 c), there is a slight improvement in EER for the GFCC baseline, and especially at 5dB with 10% EER, compared with 10.8% EER for the MFCC based system. This represents a clear improvement compared with the EER results of the MFCC baseline. In contrast, both systems show the same performance for a clean environment with 0.114% EER for both.



	Clean	20	15	10	5	0
Babble GFCC	0.1149	0.2299	0.8046	3.5632	10.0000	16.6667
Babble MFCC	0.1149	1.9540	3.3333	9.4253	16.7816	23.3333

	Clean	20	15	10	5	0
Car GFCC	0.1149	0.1149	0.3448	0.4598	1.3793	3.1034
Car MFCC	0.1149	1.0345	1.3793	3.1034	4.5977	6.6667

	Clean	20	15	10	5	0
GFCC	0.1149	0.3448	0.9195	3.3333	10.0000	25.4023
MFCC	0.1149	3.3333	3.3333	6.3218	10.8046	30.0000

	Clean	20	15	10	5	0
GFCC	0.1149	0.2299	0.5747	1.3793	6.0920	13.3333
MFCC	0.1149	1.3793	3.3333	3.3333	9.6552	20.0000

	Clean	20	15	10	5	0
GFCC	0.1149	1.2644	3.3333	3.3333	12.9885	23.3333
MFCC	0.1149	5.6322	11.0345	16.3218	27.5862	36.6667

Figure 6.6: EER for Speaker Recognition using TIMIT dataset

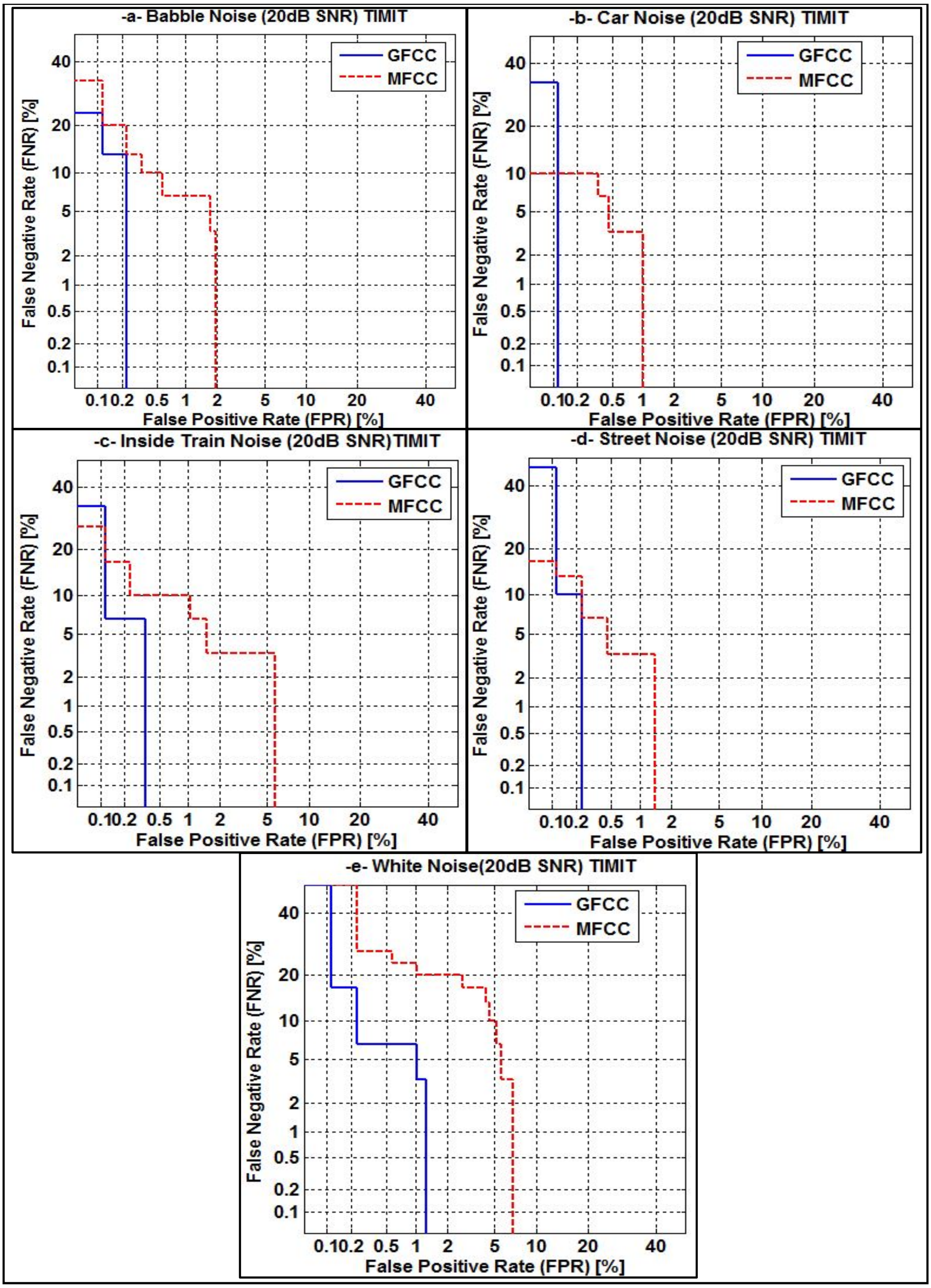


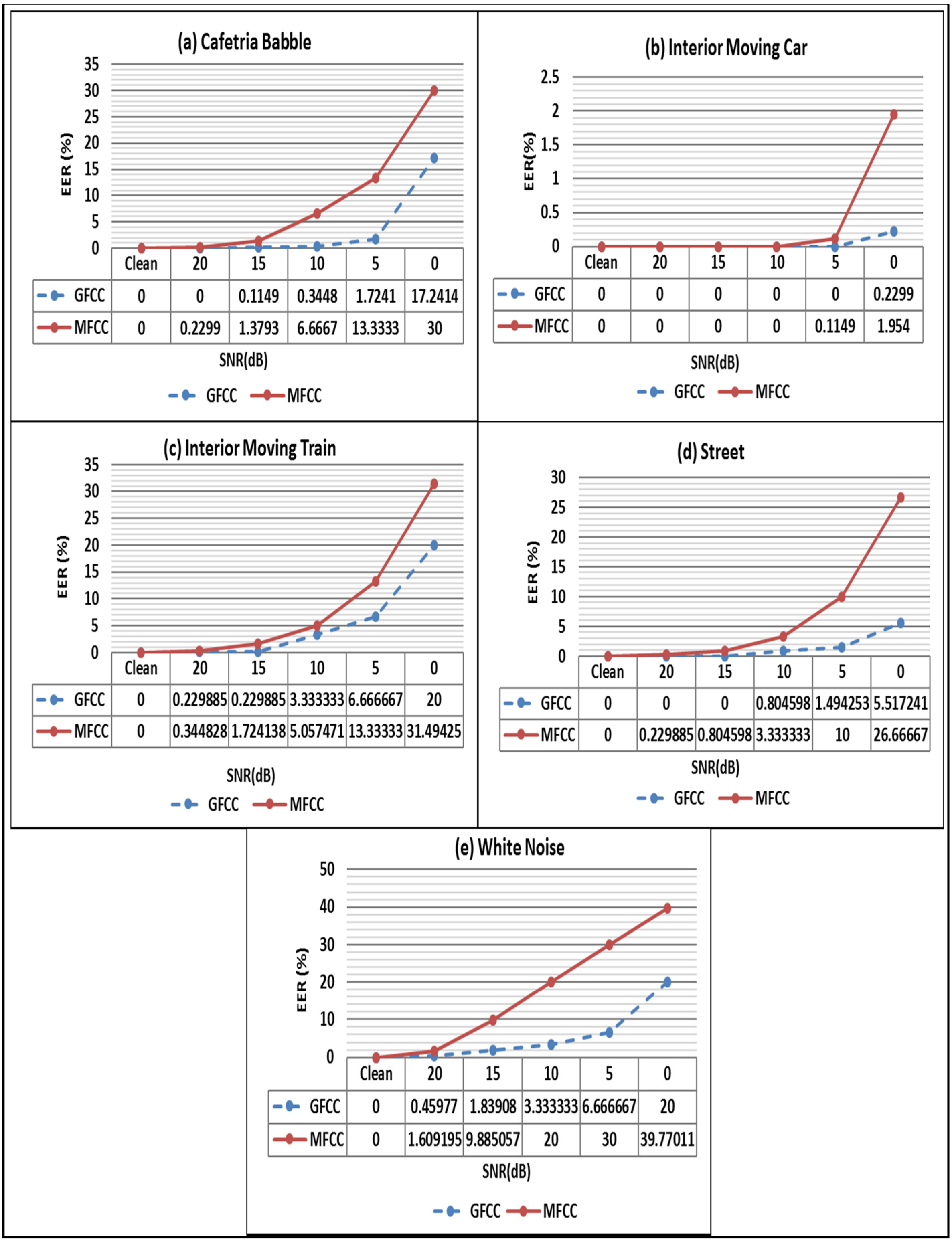
Figure 6.7: DET Graph for GFCC- MFCC based systems in 20 dB using TIMIT

Figure 6.7 illustrates a DET graph for GFCC-MFCC baselines for the five types of noisy speech samples with 20 dB SNR. It is clear that the accuracy of false positive rate FPR (False rejection rate) for the GFCC baseline (the bold blue curve) is better than the MFCC baseline (the dashed red line). False Negative Rate FNR (False Acceptance Rate) for the MFCC baseline, conversely, shows higher performance for some types of noisy samples.

More details about DET graphs for different SNRs are illustrated in Appendix -II-.

When the SALU-AC database was used with the first set of speech samples (a set that included English utterances), both baselines (GFCC and MFCC based speaker recognition) showed significant robustness for the various noise types (Figure 6.8). However, the GFCC baseline still showed greater robustness for various noises than the MFCC baseline, especially when the SNR became too low. For example, in speech samples with street noise (Figure 6.8. d), the GFCC baseline gave 1.49%, 5.52% EER for 5dB, and 0 dB respectively, compared with 10%, 26.67% EER for the MFCC baseline for the same SNRs. Similarly, other noisy speech sample types showed a significant improvement for the GFCC baseline especially for 20 dB and 15 dB with an accuracy of approximately 0% EER as seen with speech samples contaminated with Cafeteria Speech Babble and Interior moving Car noisy speech (Figure 6.8 a, b).

For White noise speech samples (Figure 6.8 e), it is clear that there is a considerable improvement in the accuracy of the GFCC baseline, especially between 15 and 0 dB SNR compared with the MFCC baseline for the same range. As when using the TIMIT database, both baselines show the same accuracy in a clean environment with 0% EER.



(a) Cafetria Babble

	Clean	20	15	10	5	0
GFCC	0	0	0.1149	0.3448	1.7241	17.2414
MFCC	0	0.2299	1.3793	6.6667	13.3333	30

(b) Interior Moving Car

	Clean	20	15	10	5	0
GFCC	0	0	0	0	0	0.2299
MFCC	0	0	0	0	0.1149	1.954

(c) Interior Moving Train

	Clean	20	15	10	5	0
GFCC	0	0.229885	0.229885	3.333333	6.666667	20
MFCC	0	0.344828	1.724138	5.057471	13.33333	31.49425

(d) Street

	Clean	20	15	10	5	0
GFCC	0	0	0	0.804598	1.494253	5.517241
MFCC	0	0.229885	0.804598	3.333333	10	26.66667

(e) White Noise

	Clean	20	15	10	5	0
GFCC	0	0.45977	1.83908	3.333333	6.666667	20
MFCC	0	1.609195	9.885057	20	30	39.77011

Figure 6.8: EER for speaker recognition using English set of SALU-AC

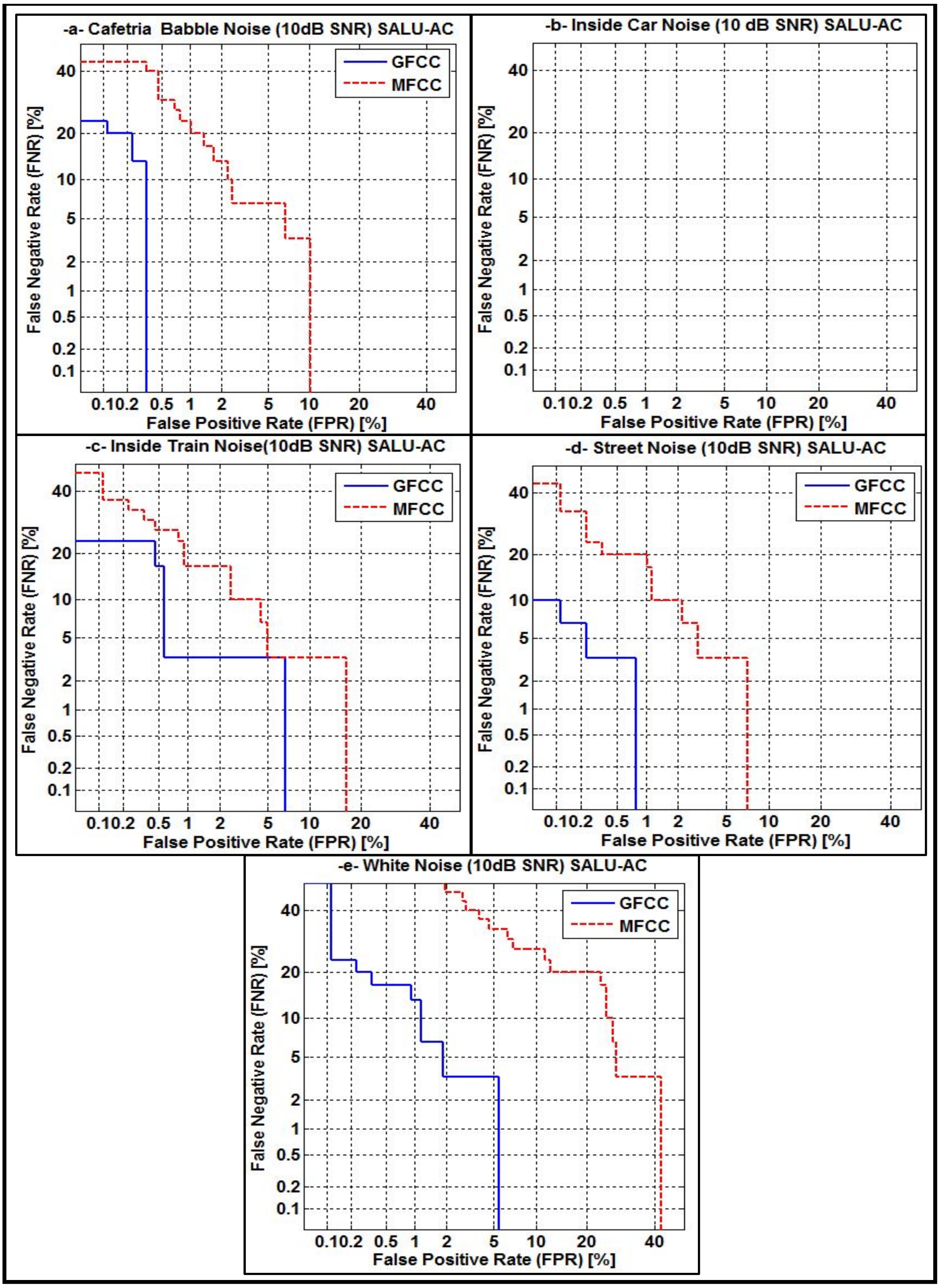
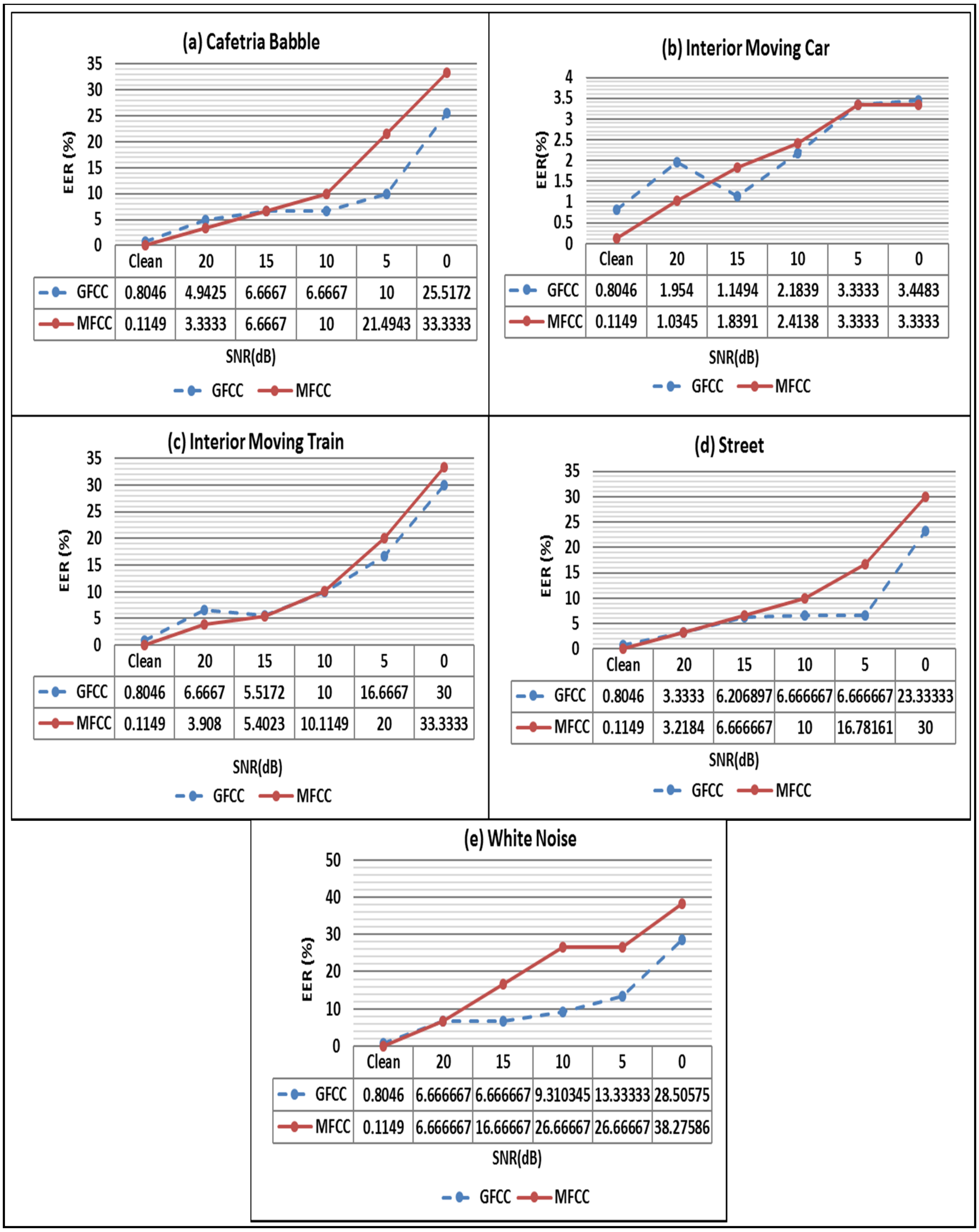


Figure 6.9: DET Graph for GFCC- MFCC based systems in10 dB using SALU-AC (First set)

Figure 6.9 shows DET graphs for GFCC/MFCC baselines using the SALU-AC database with the English utterances set applied for recognition phase in 10 dB SNR. It is noticeable that the GFCC baseline has a better rate for both FPR, FNR than the MFCC baseline. It is also noticeable that the DET Graphs of speech set contaminated with interior moving car noise did not appear, since EER for baselines equals 0 and accuracy of recognition is 100% for both Speaker recognitions.

On the other hand, when the second set of speech samples from the same database are applied (the set that includes utterances from different languages) in the recognition phase (Figure 6.10 a-e), the performance of the GFCC baseline is decreased dramatically, particularly at the clean and high SNRs. At the same time, it can be seen that the change in the accuracy of the MFCC baseline remains almost stable compared with the accuracy of the English utterances set for all types of noises.

It can also be seen that sometimes the performance of the MFCC baseline becomes better than the GFCC baseline, as seen in clean and 20 dB for most types of noisy samples sets (for a clean environment the MFCC baseline has EER 0.114 in contrast to 0.804 EER for the GFCC baseline). However, in some kinds of noise, the GFCC baseline still shows better performance than the MFCC baseline, as seen in speech samples with Cafeteria babble and White noise at 5dB, and 0 dB, with EER 10%, and 25.51% for speech babble and 13.33%, and 28.5% for white noise respectively (Figure 6.10 a, e).



(a) Cafetria Babble

	Clean	20	15	10	5	0
GFCC	0.8046	4.9425	6.6667	6.6667	10	25.5172
MFCC	0.1149	3.3333	6.6667	10	21.4943	33.3333

(b) Interior Moving Car

	Clean	20	15	10	5	0
GFCC	0.8046	1.954	1.1494	2.1839	3.3333	3.4483
MFCC	0.1149	1.0345	1.8391	2.4138	3.3333	3.3333

(c) Interior Moving Train

	Clean	20	15	10	5	0
GFCC	0.8046	6.6667	5.5172	10	16.6667	30
MFCC	0.1149	3.908	5.4023	10.1149	20	33.3333

(d) Street

	Clean	20	15	10	5	0
GFCC	0.8046	3.3333	6.206897	6.666667	6.666667	23.33333
MFCC	0.1149	3.2184	6.666667	10	16.78161	30

(e) White Noise

	Clean	20	15	10	5	0
GFCC	0.8046	6.666667	6.666667	9.310345	13.33333	28.50575
MFCC	0.1149	6.666667	16.66667	26.66667	26.66667	38.27586

Figure 6.10: EER using second set (different language set) for SALU-AC

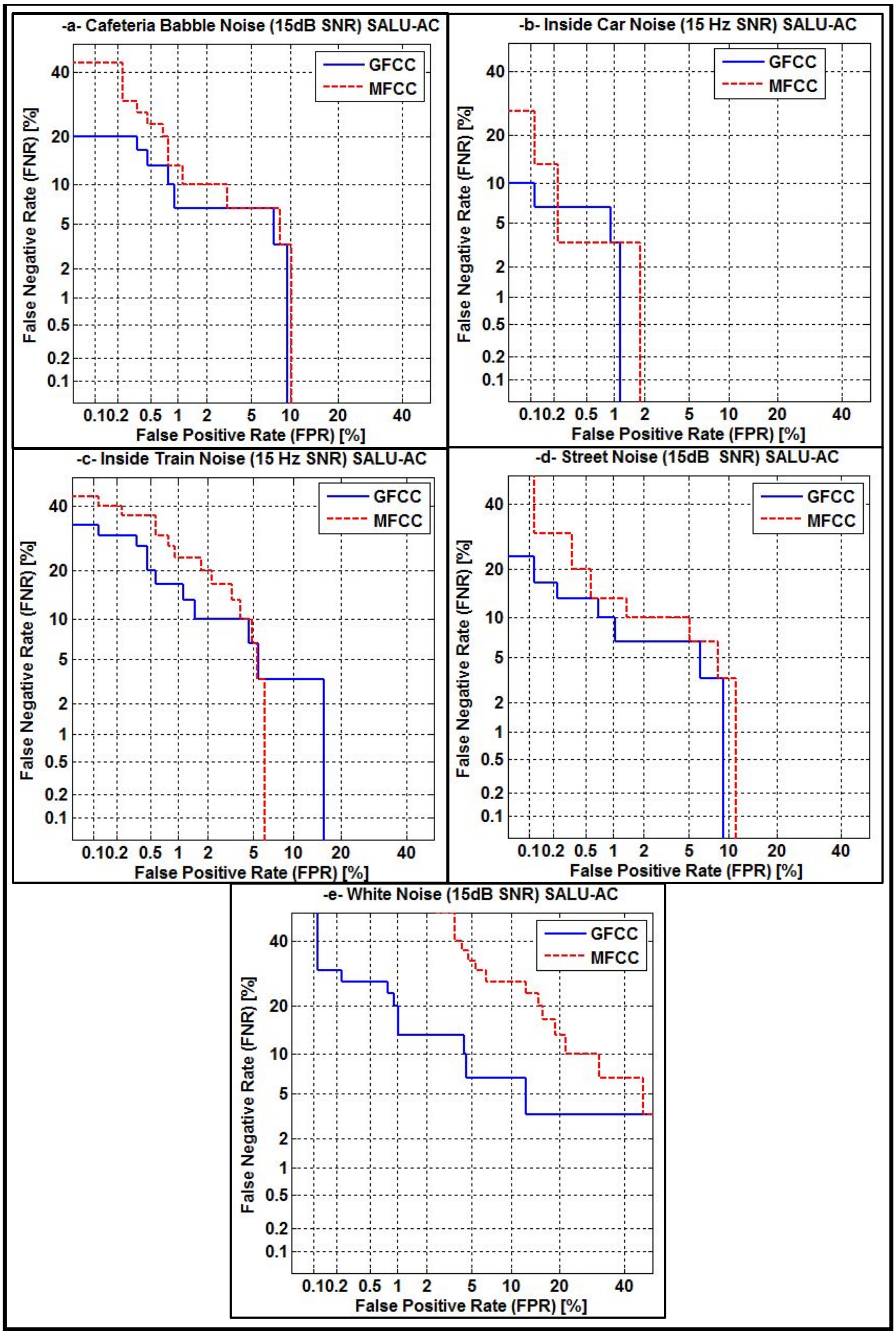


Figure 6.11: DET Graph for GFCC- MFCC based systems in10 dB using SALU-AC (Second set)

In the DET graph for the second recognition set (which represents the performance of systems at 15dB SNR), it is clear that the accuracy of the GFCC baseline is substantially affected by the change of languages in the recognition phase (Figure 6.11). It is clear that FPR for both baselines for speech samples contaminated with cafeteria babble noise are approximately similar (Figure 6.11-a-), while FPR for the MFCC baseline for samples with interior moving train noise is better (Figure 6.11 -c-). Table 6-4 shows the error equal rate for both sets of SALU-AC for different types of noise and with various SNRs. The high effect of changing languages on the GFCC baseline is obvious compared with the MFCC baseline which is affected to a lower degree.

Table 6-4: EER for MFCC–GFCC based systems using the two sets of SALU-AC

Noise type	SNR	MFCC (first set)	MFCC (second set)	Differences	GFCC SALU-AC (First set)	GFCC SALU-AC (second set)	Differences
Cafeteria Babble	Clean	0	0.1149	0.1149	0	0.8046	0.8046
	20dB	0.2299	3.3333	3.1034	0	4.9425	4.9425
	15dB	1.3793	6.6667	5.2874	0.1149	6.6667	6.5518
	10dB	6.6667	10	3.3333	0.3448	6.6667	6.3219
	05dB	13.3333	21.4943	8.161	1.7241	10	8.2759
	0 dB	30	33.3333	3.3333	17.2414	25.5172	8.2758
Interior moving car	Clean	0	0.1149	0.1149	0	0.8046	0.8046
	20dB	0	1.0345	1.0345	0	1.954	1.954
	15dB	0	1.8391	1.8391	0	1.1494	1.1494
	10dB	0	2.4138	2.4138	0	2.1839	2.1839
	05dB	0.1149	3.3333	3.2184	0	3.3333	3.3333
	0 dB	1.954	3.3333	1.3793	0.2299	3.4483	3.2184
Interior moving Train	Clean	0	0.1149	0.1149	0	0.8046	0.8046
	20dB	0.3448	3.908	3.5632	0.2298	6.6667	6.4369
	15dB	1.7241	5.4023	3.6782	0.2298	5.5172	5.2874
	10dB	5.0575	10.1149	5.0574	3.3333	10	6.6667
	05dB	13.3333	20	6.6667	6.6666	16.6667	10.0001
	0 dB	31.4943	33.3333	1.839	20	30	10
Street	Clean	0	0.1149	0.1149	0	0.8046	0.8046
	20dB	0.2299	3.2184	2.9885	0	3.3333	3.3333
	15dB	0.8046	6.666667	5.862067	0	6.206897	6.206897
	10dB	3.3333	10	6.6667	0.8046	6.666667	5.862067
	05dB	10	16.78161	6.78161	1.4943	6.666667	5.172367
	0 dB	26.6667	30	3.3333	5.5172	23.33333	17.81613
White noise	Clean	0	0.1149	0.1149	0	0.8046	0.8046
	20dB	1.6092	6.6667	5.0575	0.4598	6.6667	6.2069
	15dB	9.8851	16.6667	6.7816	1.8391	6.6667	4.8276
	10dB	20	26.6667	6.6667	3.3333	9.3103	5.977
	05dB	30	26.6667	3.3333	6.6667	13.3333	6.6666
	0 dB	39.7701	38.2759	1.4942	20	28.5057	8.5057

The Results of this investigation show that the GFCC features are more robust for noisy environment, but still have limitations in responding to the change of languages between enrolment (Training) and recognitions (Test) phases; MFCC features show more sensitivity to different Environmental noise but are still more active to this change of language. This limitation of GFCC for mismatched languages probably reflects the structure of the Gammatone filter-bank, since it emulates the auditory system of human ear, which makes it more effect for this mismatch. Further investigation on GFCC features and mismatch languages will be subject of future work.

6.3 Chapter Summary

This chapter has investigated the impact of using different features extraction algorithms on the robustness of speaker recognition in different noisy environments. MFCC and GFCC were adopted in this study to find their robustness to different kinds of noisy speech sets. In addition, the experiments included investigating a change of languages of speakers in the test phase on the performance of these features. The results of these experiments show that GFCC has better recognition performance than the conventional MFCC, while both features show the same recognition performance in a clean environment. On the other hand, the modification in languages of the recognition set shows a drop in recognition performance of GFCC both in clean and noisy speech, while the results for MFCC show greater robustness for this modification.

CHAPTER 7

THE IMPACT OF USING SPEECH ENHANCEMENT TECHNIQUES ON PERFORMANCE OF SPEAKER RECOGNITION

Chapter Overview

This chapter describes how different types of speech enhancement approaches were studied and compared for their robustness as a pre-processing stage for speaker recognition. Seven different speech enhancement algorithms were adopted for this task. These algorithms included a spectral subtraction algorithm, statistical based algorithms (e.g. Wiener filtering, and Psychoacoustically motivated algorithm), and a subspace algorithm. First, a Wiener filter was investigated as a pre-processing stage for the recognition phase only and for the enrolment (training) and recognition (test) phases together. Then, the effectiveness of other speech enhancement algorithms on the robustness of speaker recognition under noisy environments was studied. The following section gives a brief review of these algorithms and their classes. Experiment details and results are given in section 7.2. Finally, there is a summary discussion in section 7.3.

7.1 Speech Enhancement Algorithms

The primary goal of speech enhancement is to improve the quality and intelligibility of a speech signal that has been contaminated with environmental noise. This improvement in quality is necessary, especially when the speech signal is generated in a noisy location, or if it is contaminated with noise over the communication channel. This reduction of additive noise from the speech signal always comes at the expense of introducing speech distortion, which in turn may affect speech intelligibility (Loizou, 2013). Therefore, one of the essential challenges in developing effective speech enhancement approaches is to suppress noise without causing any perceptible distortion in the speech signal. Consequently, the main function of speech enhancement approaches is to reduce or suppress the background noise to some degree without affecting speech intelligibility. Therefore these algorithms are also known as ***noise suppression algorithms*** (Loizou, 2013).

Different algorithms have been developed in the literature to deal with clean speech signal and reduce the degree of noise with the primary goal of improving speech quality. These approaches can be classified into three main classes (Loizou, 2013):

1. Spectral Subtraction Algorithms: these algorithms are based on estimating and updating the spectrum of the noise when the speech is absent, then subtracting the result from a noisy signal. These algorithms were first proposed by Weiss et al. (1975) and Boll (1979).
2. Statistical- model-Based Algorithms: in these the cleaning speech problems are modelled in a statistical estimation framework. This is based on a set of measurements, corresponding, for example, to the Fourier transform coefficients of the noisy speech signal, in order to find a linear (or nonlinear) estimator of the parameter of interest, known as the transform coefficients of the speech signal (Loizou, 2013). For these types of approach, the Wiener filter (Lim and Oppenheim, 1978 and Lim and Oppenheim, 1979), and Minimum mean square error (MMSE) algorithms (Ephraim and Malah, 1984) and many others fall into this category.
3. Subspace Algorithms: these are based on linear algebra. The main idea behind these algorithms is that the clean signal might be confined to a subspace of the noisy Euclidean Space. Therefore, based on the process of separating the vector space on a noisy speech

signal into a clean signal subspace, which is primarily occupied by the clean speech, and a noise subspace which is mainly occupied by the noise (Loizou, 2013). Then, the clean speech signal is obtained by nulling the component of noisy vector in the noise subspace. These algorithms were proposed by Dendrinos et al. (1991) and Ephraim and Van Trees (1995).

These algorithms are focused mainly on enhancing speech quality but they failed in improving speech intelligibility. In the next section, the capability of these algorithms as a pre-processing stage was investigated to improve the performance of speaker recognition and to determine how much they help to increase the accuracy of recognition.

7.2 A Comparative Investigation of Speech enhancement for the performance of Speaker Recognition

In this study, seven different speech enhancement algorithms were adopted and investigated as a pre-processing stage for their robustness for speaker recognition in different noisy environments. These algorithms include:

- Wiener filtering based on a priori SNR estimator (Scalart and Filho, 1996): the authors applied Wiener filtering approach based on prior estimation of SNR in order to obtain significant results. According to the authors, this priory estimation permits important noise power suppression without introducing “musical noise” effects.
- Spectral subtraction with adaptive gain averaging (Gustafsson et al., 2001): a proposed approach based on conventional spectral subtraction. This approach applied a noise and speech gain function for each frequency component. This approach, in addition, employs spectrum-dependent adaptive averaging to reduce the variance of gain function.
- Psychoacoustical motivated algorithm (MT-mask) (Hu and Loizou, 2004): A frequency domain optimal linear estimator which includes the masking properties of the human auditory system to make the embedded noise distortion inaudible.
- Audible noise suppression algorithm (Tsoukalas et al., 1997): this technique is dependent on the definition of the psychoacoustically derived quantity of audible noise

spectrum and its subsequent suppressing using optimal nonlinear filter on short time spectral amplitude envelope.

- Log Minimum Mean Square Error (LogMMSE) algorithm (Cohen, 2002): an optimal modification to the log-spectral amplitude estimator, which decreases the mean-square error of log-spectra for speech signal under signal presence uncertainty.
- Bayesian estimator based on weighted-Euclidian distortion measure (Loizou, 2005): this approach represents a development of minimum mean square error (MMSE), which employs Bayesian estimators of short-time spectral magnitude of speech depending on perceptually motivated cost functions.
- Perceptually motivated subspace algorithm (Jabloun and Champagne, 2003): a Frequency to Eigendomain Transformation (FET) which permits measurement of a perceptually based eigenfilter.

First, Wiener filtering as proposed by Scalart and Filho (1996) was applied as a pre-processing stage to filter the noisy speech in the recognition phase only. The same filter was also used as pre-processing for both enrolment and recognition phases. A comparison was then made between the accuracy of speaker recognition in the two cases as well as its accuracy without using this filter. The effect of other aforementioned methods on speaker recognition performance was then investigated.

7.2.A Experimental Setup

As discussed earlier, the aim of these experiments is to investigate the capability of speech enhancement algorithms for improving the robustness of speaker recognition by using them as pre-processing for noisy input signal in the test phase only and in the train-test phases together. The details of these experiments are as follows:

- MFCC including dynamic features (delta and double delta) have been used in these experiments with 39-dimensional space.
- Gaussian mixture model (GMM) was used with 256 mixture.
- The experiments were conducted on a speech samples set from the SALU-AC database. This set included speech samples collected from 60 speakers (30 males and 30 females). For each speaker, a set of 9 utterances were used in the training phase, all spoken in

English, and one utterance for each speaker was used in the recognition phase, also spoken in English. In addition, the recognition set was divided into a number of subsets. Each subset represents the same speech samples mixing with different SNRs (20, 15, 10, 5, and 0 dB) for a specific type of noise. In these experiments, the five types of noise, that mentioned in Chapter 3 were used.

- The evaluation speaker recognition system mainly depends on using Error Equal Rate (EER) to evaluate the performance of the SR system, as well as a Detection Error Tradeoff (DET) graph. The evaluations are computed based on the verification scores (log-likelihood score) based on 60 authentic cases (verified cases) and 3540 unauthentic cases (imposter cases).

Figure 7.1 illustrates how speech enhancement methods are applied as a pre-processing stage for speaker recognition; the blue arrows refer to the test data filtered by speech enhancement algorithms, while the bold red arrows refers to the training data without filtering. The dashed red arrows represent the filtered training data.

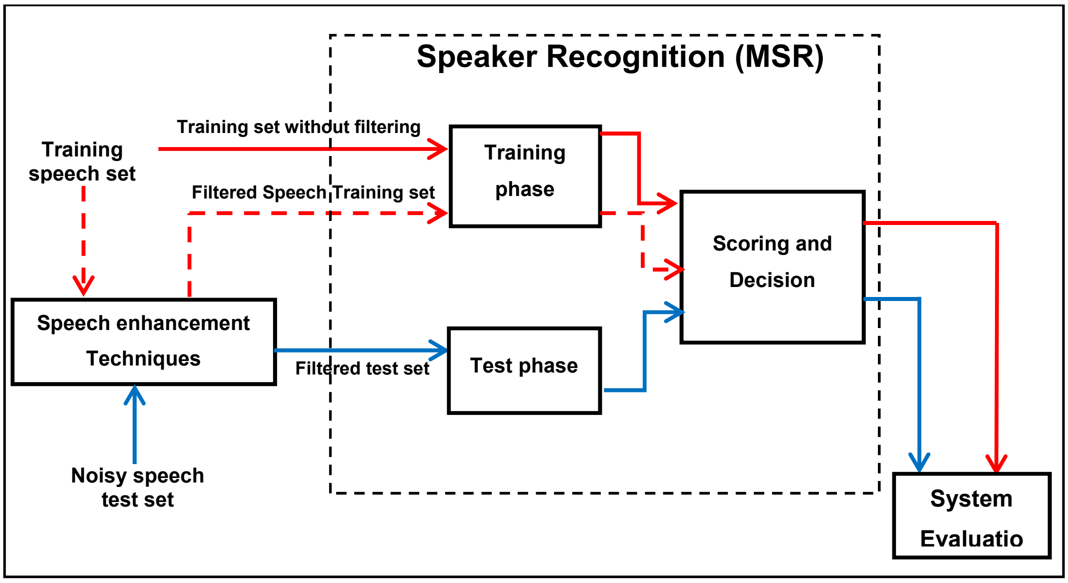
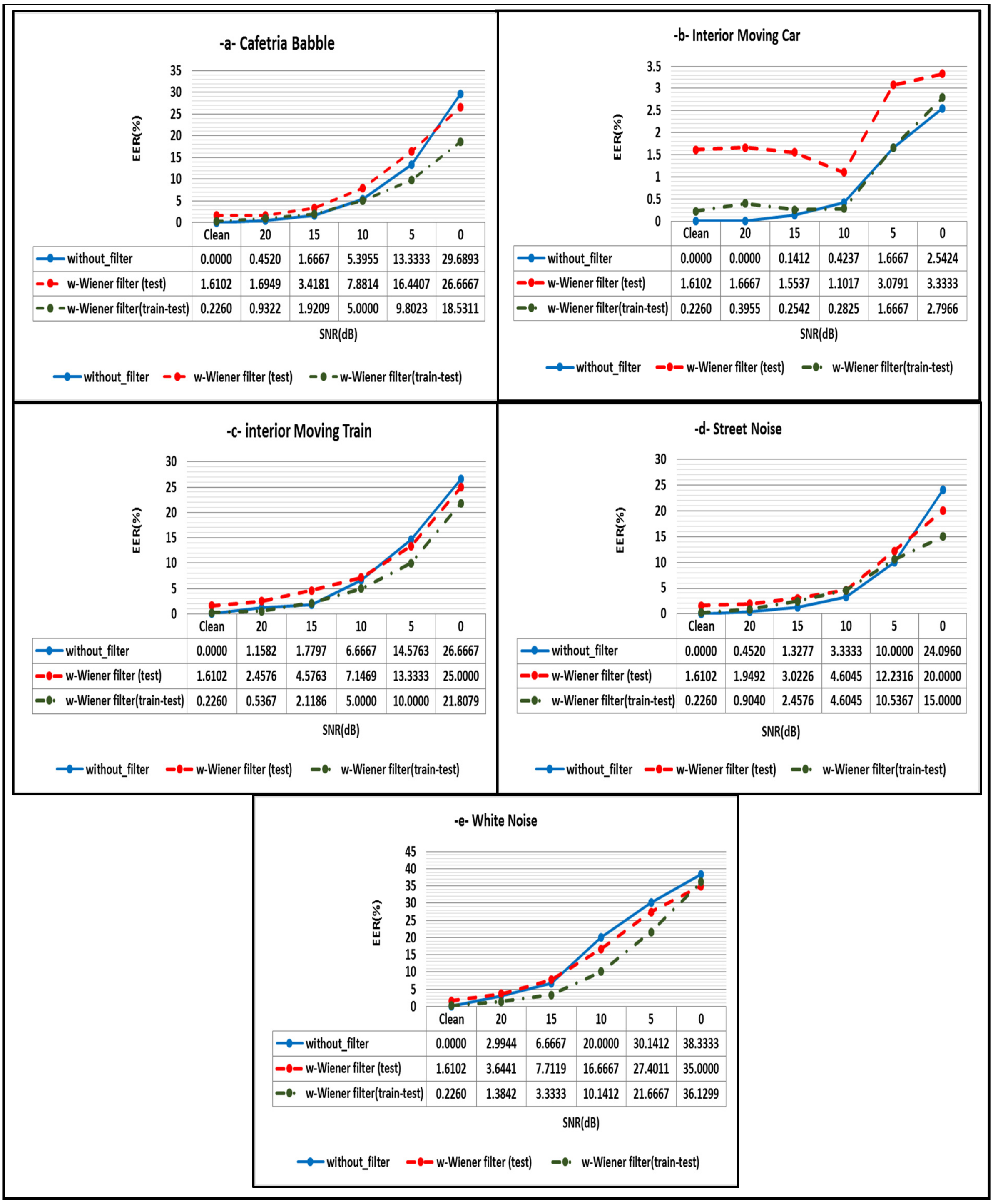


Figure 7.1: Speaker Recognition with speech enhancement Techniques

7.2.B Experimental Results

The aim of these experiments was to investigate how speech enhancement techniques can improve the robustness of speaker recognition under a different type of noise environment. First the effect of applying Wiener filtering developed by Scalart and Filho (1996) as a pre-processing stage for the recognition phase only (test phase) in speaker recognition was studied, and the evaluation results compared with those obtained when the same filter was applied to both enrolment and recognition phases (Figure 7.1).



-a- Cafetria Babble

	Clean	20	15	10	5	0
without_filter	0.0000	0.4520	1.6667	5.3955	13.3333	29.6893
w-Wiener filter (test)	1.6102	1.6949	3.4181	7.8814	16.4407	26.6667
w-Wiener filter(train-test)	0.2260	0.9322	1.9209	5.0000	9.8023	18.5311

-b- Interior Moving Car

	Clean	20	15	10	5	0
without_filter	0.0000	0.0000	0.1412	0.4237	1.6667	2.5424
w-Wiener filter (test)	1.6102	1.6667	1.5537	1.1017	3.0791	3.3333
w-Wiener filter(train-test)	0.2260	0.3955	0.2542	0.2825	1.6667	2.7966

-c- interior Moving Train

	Clean	20	15	10	5	0
without_filter	0.0000	1.1582	1.7797	6.6667	14.5763	26.6667
w-Wiener filter (test)	1.6102	2.4576	4.5763	7.1469	13.3333	25.0000
w-Wiener filter(train-test)	0.2260	0.5367	2.1186	5.0000	10.0000	21.8079

-d- Street Noise

	Clean	20	15	10	5	0
without_filter	0.0000	0.4520	1.3277	3.3333	10.0000	24.0960
w-Wiener filter (test)	1.6102	1.9492	3.0226	4.6045	12.2316	20.0000
w-Wiener filter(train-test)	0.2260	0.9040	2.4576	4.6045	10.5367	15.0000

-e- White Noise

	Clean	20	15	10	5	0
without_filter	0.0000	2.9944	6.6667	20.0000	30.1412	38.3333
w-Wiener filter (test)	1.6102	3.6441	7.7119	16.6667	27.4011	35.0000
w-Wiener filter(train-test)	0.2260	1.3842	3.3333	10.1412	21.6667	36.1299

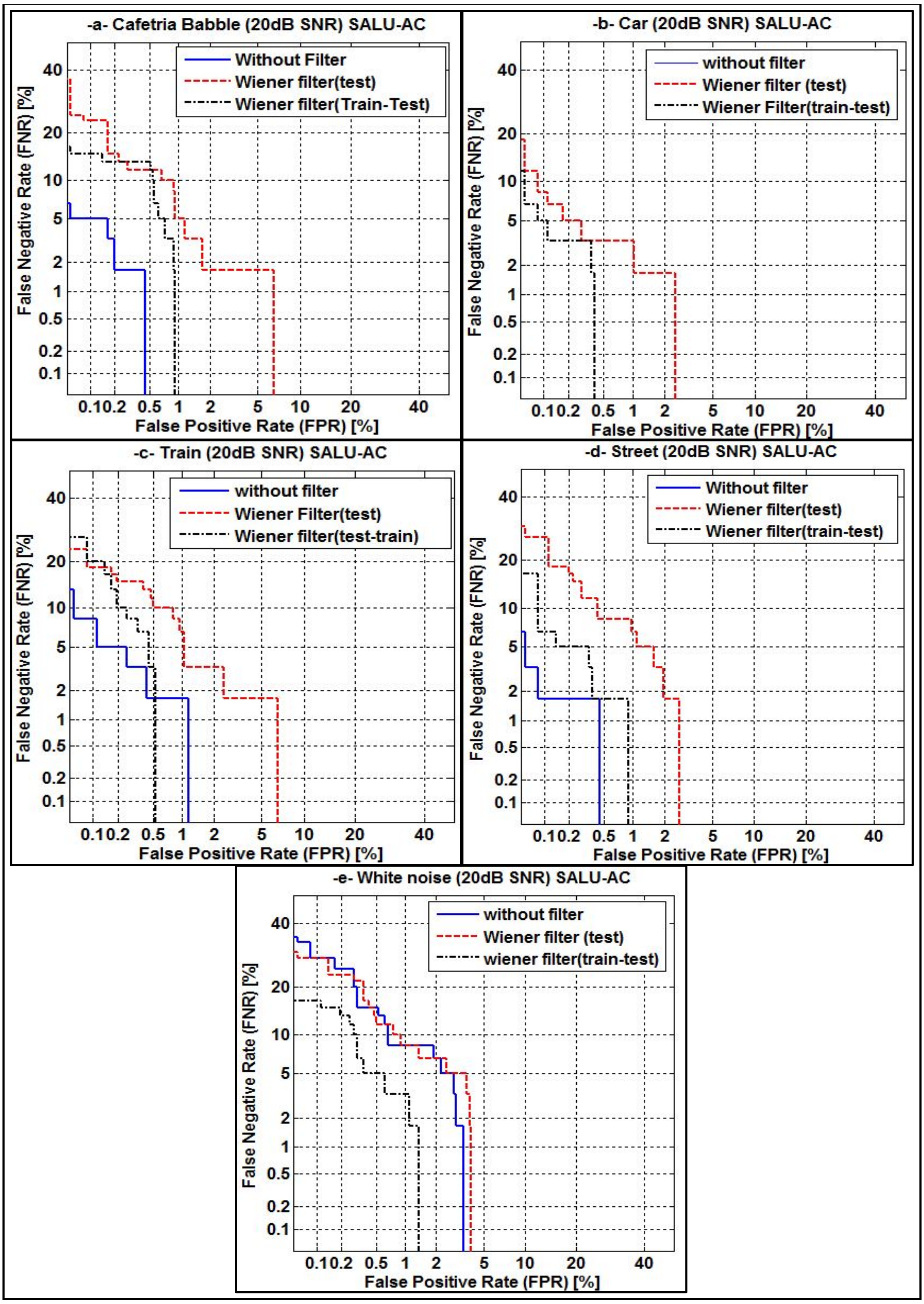
Figure 7.2: Speaker Recognition performance with and without Wiener filter

The accuracy of speaker recognition (based on EER) in both cases and their comparison with the accuracy of the SR without using any filters are demonstrated in Figure 7.2 a-e. The solid blue line refers to the performance of the SR without using any filter and the dashed red line refers to the performance of the SR using Wiener filter for the recognition phase only, while the dashed-dot green line refers to the performance of the SR using Wiener filter for both enrolment and recognition phases.

It is clear that using Wiener filter in the test phase caused degradation in the accuracy of speaker recognition in all types of noisy speech samples except in the case of speech samples contaminated with White noise when SNR became low (i.e. between 10 and 0 dB). For instance, in speech samples with babble noise (Figure 7.2 a) the recognition accuracy was degraded from 1.66% EER to 3.41 and from 5.39 to 7.88 for 15 dB and 10 dB SNR respectively, and similar results were seen for the remaining SNRs for the same noisy speech samples, except at 0dB where there was some slight improvement. Speech samples contaminated with interior moving Car, interior moving Train, and Street noise (Figure 7.2 b, c, and d) showed the same degradation in performance for most SNRs. Conversely, the speech samples with the white noise (Figure 7.2 e) showed a minor improvement in accuracy for the SNR between 10 dB and 0dB when the Wiener filter was applied in the test phase. The SR also shows high degradation in performance for the clean samples.

However, when Wiener filter was used in both training and testing phases, it gave slight improvement in the accuracy of the SR for some types of noisy speech samples and in some points of SNR. But it is also noticeable how this improvement varies from one type of noisy sample to another and from one SNR to another. For example, in speech samples contaminated with moving train noise, there was a slight improvement in accuracy between 10 and 0 dB (Figure 7.2–c-), and the same improvement can seen in cafeteria babble noisy speech in some SNRs. However, noisy speech samples with street noise (Figure 7.2-d-) show degradation in performance for most SNRs with almost the same degradation that was seen when using the filter in test phase only. On the other hand, for speech samples contaminated with white noise, there is an improvement in performance compared with the other two cases (without using a filter and with using it in test phase only).

In summary, using Wiener filtering in recognition phase only (test-phase) caused degradation in the performance of the verification system for most types of noisy speech sample. However, applying the same filter for both phases (training-testing) showed some improvement in performance for some kinds of noisy speech samples and in some SNRs but this improvement was still marginal. Moreover, this improvement was not reliable for all kinds of noises (as seen in the street noisy speech sample). The main reason for this degradation when using speech enhancement algorithm in test phase is that the input speech signal will become distorted when passing through this cleaning processing, so it is necessary to apply these kinds of cleaning algorithms on both enrolment and recognition phases in order to improve the performance of the verification system. However, even that cannot be generalised for all types of noisy speech, since this improvement depends on the type of noise and signal to noise ratio. Figure 7.3 illustrates the DET graph for previous experiments at 20 dB SNR. The solid blue curve refers to the recognition performance without using a filter, and the dashed red curve refers to the performance when using a filter in test phase while the dash-dotted black line refers to the performance when using a filter in training and test phases. According to this figure the negative effect of the Wiener filter when used as pre-processing to the test phase is very clear, while the effects of using this filter in both phases varied from positive impact (such as in interior moving train and White noise) to adverse impact (Cafeteria speech babble, interior moving car, and Street). Furthermore, as seen in the interior moving train, there is an improvement in false positive rate but with degradation in false negative rate.


-a- Cafetria Babble (20dB SNR) SALU-AC

-b- Car (20dB SNR) SALU-AC

-c- Train (20dB SNR) SALU-AC

-d- Street (20dB SNR) SALU-AC

-e- White noise (20dB SNR) SALU-AC

Figure 7.3: DET graphs for Speaker Recognition with and Without Wiener filter at 20 dB

The effects of using the other six speech enhancement algorithms on the performance of speaker recognition for each type of noisy speech samples are presented below. Tables 7-1 to 7-6 show the EER with using these algorithms compared with the EER of the recognition system without using any filter for each type of noisy speech sample. Since applying a filter on the test phase caused degradation in the recognition performance (as seen with Wiener filtering) as a result of signal distortion, this case was ignored in other filtering algorithms, and we focused mainly on the comparison between recognition performance when using the filters in both phases and the performance without using any filter.

Table 7-1: EER of Speaker Recognition using Audible noise suppression Algorithm

Audible Noise Suppression Algorithm										
Noise	**Babble**		**Car**		**Train**		**Street**		**White**	
SNR	**Without filter**	**With filter**	**Without filter**	**With filter**	**Without filter**	**With filter**	**Without filter**	**With filter**	**Without filter**	**With filter**
Clean	0	6.666	0	6.666	0	6.666	0	6.666	0	6.666
20 dB	0.4520	5	0	6.6667	1.1582	10.9887	0.4520	6.6667	2.9944	13.3333
15 dB	1.6667	8.333	0.1412	4.9153	1.7797	13.3333	1.3277	6.6667	6.6667	18.418
10 dB	5.3955	16.66	0.4237	5	6.6667	16.6667	3.3333	15	20.0000	26.61
05 dB	13.3333	26.66	1.6667	8.1356	14.5763	24.0395	10	25	30.1412	26.214
0 dB	29.6893	36.66	2.5424	10	26.6667	35	24.0960	31.5254	38.3333	33.248

Table 7-2: EER of Speaker Verification using Log Minimum Mean Square Error algorithm

Log Minimum Mean Square Error (Log-MMSE) Algorithm										
Noise	**Babble**		**Car**		**Train**		**Street**		**White**	
SNR	**Without filter**	**With filter**	**Without filter**	**With filter**	**Without filter**	**With filter**	**Without filter**	**With filter**	**Without filter**	**With filter**
Clean	0	1.6667	0	1.6667	0	1.6667	0	1.6667	0	1.6667
20 dB	0.4520	1.6667	0	1.6667	1.1582	3.3333	0.4520	2.7119	2.9944	2.2316
15 dB	1.6667	3.0226	0.1412	1.6667	1.7797	5	1.3277	2.7401	6.6667	3.4746
10 dB	5.3955	8.3333	0.4237	1.6667	6.6667	8.9548	3.3333	6.6667	20.0000	11.2429
05 dB	13.3333	17.7119	1.6667	3.3333	14.5763	13.3333	10	13.7006	30.1412	23.3333
0 dB	29.6893	28.3333	2.5424	4.7740	26.6667	26.6667	24.0960	27.9661	38.3333	31.6949

Table 7-3: EER of Speaker Recognition using Psychoacoustical motivated algorithm

Psychoacoustical motivated algorithm (MT-mask)										
Noise	Babble		Car		Train		Street		White	
SNR	Without filter	With filter	Without filter	With filter	Without filter	With filter	Without filter	With filter	Without filter	With filter
Clean	0	0.226	0	0.226	0	0.226	0	0.226	0	0.226
20 dB	0.4520	0.6780	0	0.6215	1.1582	0.6780	0.4520	0.6497	2.9944	1.1582
15 dB	1.6667	0.9322	0.1412	0.9887	1.7797	2.4011	1.3277	1.6667	6.6667	2.3446
10 dB	5.3955	5.1412	0.4237	1.6667	6.6667	5	3.3333	3.3333	20.0000	9.0960
05 dB	13.3333	11.6667	1.6667	1.8362	14.5763	12.0339	10	8.3333	30.1412	21.6667
0 dB	29.6893	23.3333	2.5424	2.5424	26.6667	27.9661	24.0960	16.6667	38.3333	33.9266

Table 7-4: EER of Speaker Recognition using Spectral subtraction with adaptive gain averaging algorithm

Spectral subtraction with adaptive gain averaging algorithms										
Noise	Babble		Car		Train		Street		White	
SNR	Without filter	With filter	Without filter	With filter	Without filter	With filter	Without filter	With filter	Without filter	With filter
Clean	0	1.6667	0	1.6667	0	1.6667	0	1.6667	0	1.6667
20 dB	0.4520	3.3333	0	1.6667	1.1582	4.8305	0.4520	1.6667	2.9944	10
15 dB	1.6667	6.5537	0.1412	3.3333	1.7797	6.6667	1.3277	3.9831	6.6667	17.1751
10 dB	5.3955	10.4520	0.4237	3.3333	6.6667	12.7401	3.3333	8.9831	20.0000	26.6667
05 dB	13.3333	26.6667	1.6667	5.3107	14.5763	19.8588	10	16.6667	30.1412	31.6667
0 dB	29.6893	30	2.5424	8.3333	26.6667	32.7401	24.0960	27.8814	38.3333	38.8983

Table 7-5: EER of Speaker Recognition using Bayesian estimator based on weighted-Euclidian distortion measure

Bayesian estimator based on weighted-Euclidian distortion measure										
Noise	Babble		Car		Train		Street		White	
SNR	Without filter	With filter	Without filter	With filter	Without filter	With filter	Without filter	With filter	Without filter	With filter
Clean	0	3.6441	0	3.6441	0	3.6441	0	3.6441	0	3.6441
20 dB	0.4520	1.3842	0	2.0621	1.1582	4.0960	0.4520	3.3333	2.9944	2.2599
15 dB	1.6667	2.3164	0.1412	1.6667	1.7797	7.1469	1.3277	3.3333	6.6667	5
10 dB	5.3955	6.6667	0.4237	1.3842	6.6667	8.3333	3.3333	5.2260	20.0000	13.3333
05 dB	13.3333	14.6893	1.6667	3.3333	14.5763	13.7571	10	13.3333	30.1412	25
0 dB	29.6893	25	2.5424	5.4802	26.6667	25	24.0960	24.4068	38.3333	35

Table 7-6: EER of Speaker Recognition using perceptually motivated subspace algorithm

Perceptually motivated subspace algorithm										
Noise	Babble		Car		Train		Street		White	
SNR	Without filter	With filter	Without filter	With filter	Without filter	With filter	Without filter	With filter	Without filter	With filter
Clean	0	1.6667	0	1.6667	0	1.6667	0	1.6667	0	1.6667
20 dB	0.4520	1.6667	0	1.6667	1.1582	1.9209	0.4520	1.5537	2.9944	1.2994
15 dB	1.6667	2.0621	0.1412	1.6667	1.7797	2.4011	1.3277	1.3277	6.6667	2.5989
10 dB	5.3955	3.3898	0.4237	2.4011	6.6667	5	3.3333	3.3333	20.0000	7.7119
05 dB	13.3333	13.3333	1.6667	5.0000	14.5763	11.6667	10	10	30.1412	16.6667
0 dB	29.6893	25.4520	2.5424	8.7571	26.6667	23.3333	24.0960	22.3164	38.3333	31.6667

According to the results obtained from these experiments on speech enchantment algorithms the following points emerged from the current investigation:

1. All of these filters cause degradation in recognition performance when the recognition speech samples are clean. However, this degradation varies from one filter to another. For example, when using ***Psychoacoustical motivated algorithm*** (Table 7-3) there is minor increasing in EER from 0 % to 0.22%, while ***Audible noise suppression*** (Table 7-1) and ***Bayesian estimator*** (Table 7-5) algorithms cause a higher increasing in EER from 0% to 6.66 and 3.6 respectively. Figure 7.4 illustrates DET graphs of these six filters on the clean signal. Note the curve of the clean signal without using any filter is not visible in the graph since EER, in this case, is 0 %.

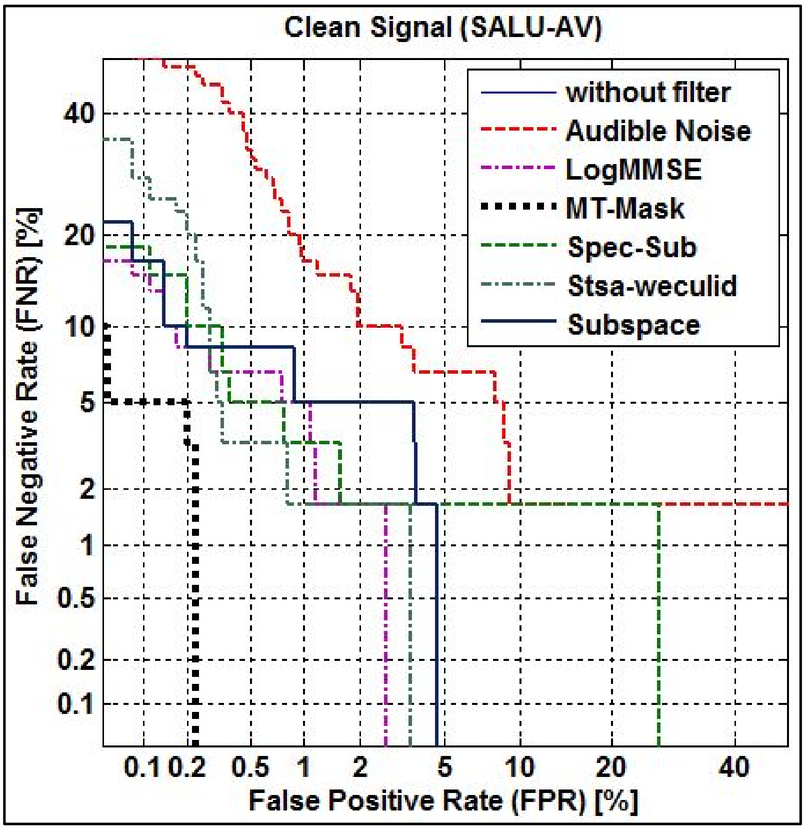


Figure 7.4: DET Graph with /without Speech enhancement approach

2. The effects of these filter algorithms on the performance of speaker recognition are varied from one filter to another and from one noise type to another. For example, the Perceptually motivated subspace approach (Table 7-6) showed some improvement in accuracy of recognition in a speech sample contaminated with interior moving train noise for 10, 5, and 0dB (and the same thing with noisy speech sample contaminated with speech cafeteria speech babble), while speech samples with car noise showed degradation in recognition accuracy for all SNRs when using the same approach. However, audible noise suppression showed degradation in accuracy recognition for all types of noise and for most SNRs.
3. Some of the approaches have different effects on SNRs for the same type of noise. For instance, Psychoacoustical motivated algorithm (Table 7-3) showed improvement in

accuracy at 20 dB and 10 dB of speech samples contaminated with interior moving train from 1.1582% EER to 0.6780 EER and from 6.6667 % EER to 5% EER respectively, while the same approach showed degradation in performance at 15dB for the same type of noise with 1.7797 % EER to 2.4011. This variety of performance makes this kind of approach unreliable for improving the robustness of speaker recognition.

4. The improvements obtained from using these de-noising approaches are almost marginal, which makes employing these kinds of methods for improving the robustness of speaker recognition in noisy environments less useful, particularly if it is known that these approaches may help to increase the speech quality, but have failed so far in improving speech intelligibility. Moreover, using these types of filter to clean the signal causes changes in the speech components in the cleaned signal. These changes will affect directly the extracted features, since the frequency content of speech will change as a result to cleaning this speech signal from noise.
5. Some of these filters are found to be efficient with stationary noise (e.g. white noise) but the same cannot be said with non-stationary noise (e.g. cafeteria speech babble, and interior moving train).

Figure 7.5 shows the DET graphs of speaker recognition using these speech cleaning approaches for the clean speech samples and different noisy speech samples with 10dB SNR. The solid black line represents the performance of the recognition system without using any kind of filter, while the other lines represent the performance of the recognition system using the six approaches. The variable effect of different types of filter on noisy speech samples is clear. ***Perceptually motivated subspace algorithm*** (the dash-dot green curve), for example, shows a slight improvement in both FPR and FNR in speech samples contaminated with Cafeteria babble noise (Figure 7.5 -a-). But the same filter causes degradation in both speech samples contaminated with interior moving car noise and interior moving train (Figure 7.5 b, c). However, some filters show significant improvement with white noisy speech samples (Figure 7.5 -e-).

More details about DET graphs for different SNRs are illustrated in Appendix -III-.

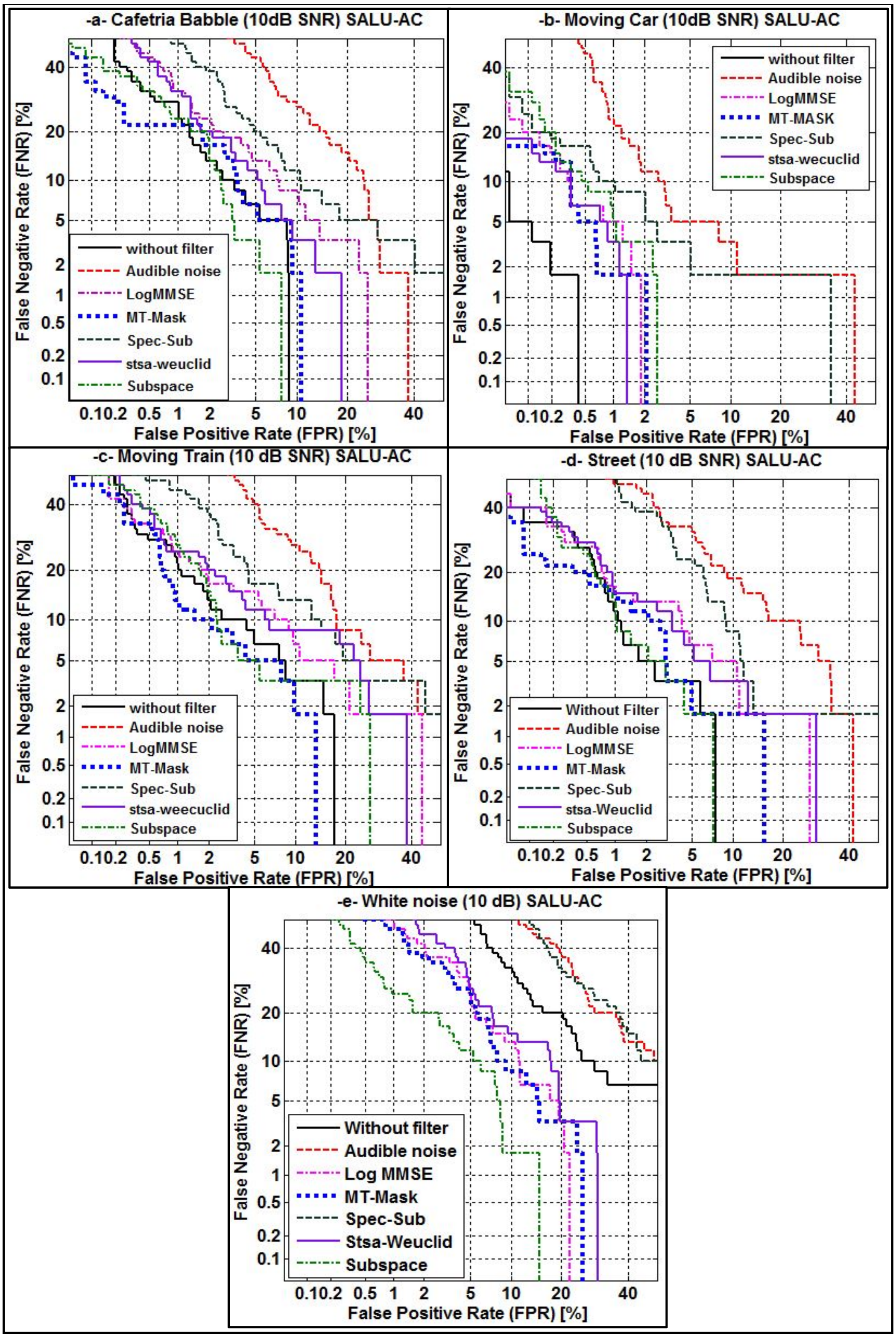


Figure 7.5: DET graph with /without speech enhancement approach at 10 dB SNR

7.3 Chapter Summary

This Chapter has dealt with investigating the impact of different speech enhancement approaches on the the robustness of speaker recognition under additive environmental noise. The cleaning approaches studied include Wiener filtering, audible noise suppression, LogMMSE, Psychoacoustical motivated algorithm, Spectral subtraction, Bayesian estimator, and Subspace Algorithms. First, we studied the effect of using Wiener filtering on the performance of speaker recognition by applying this filter on recognition-phase only and on recognition and enrolment phases. Then we examined the impact of using the other filters on the performance of recognition. The results obtained from these experiments indicate that these approaches caused degradation in accuracy of speaker recognition when they were used with recognition phase only. However, the experimental results indicate that these speech filtering approaches provide very limited robustness for the performance of speaker recognition, and at times can even have an adverse effect on the recognition system.

Therefore, this kind of approach is unreliable for robustness of speaker recognition in different noisy environments. In Chapter 9, a different purpose to employ these types of approach will be proposed. This proposal includes applying these algorithms to extract the noise from the noisy signal instead of focusing on using them to clean the signal.

CHAPTER 8
TRAINING ON THE FLY
FOR ROBUST SPEAKER RECOGNITION

Chapter Overview

*The results of the experiments in Chapter 5 show significant improvement in accuracy of Speaker Recognition when noise reference models have been used. These results suggest that inclusion of noisy conditions in the reference models can mitigate, to some extent, the performance degradation, particularly when sufficient information is available about the identity of the noise and its level (i.e. SNR) of contamination of the input signal. In other words, if the noisy conditions can be somehow estimated based on one of the SNR estimation algorithms, and if the identity of the noise can be extracted from an input contaminated signal, then it would be possible to provide a suitable re-trained model in order to improve the robustness of speaker recognition. Some researchers (i.e. Yoshida et al., 2004 and Pullella et al., 2008) proposed adopting noisy speech models to reduce the mismatch between the reference model for registered speaker and recognition materials. These multiple models represent different noisy conditions for speech signals belonging to each registered speaker. During the recognition phase, the reference model that gives the closest matches to the characteristics of the input speech signal is selected. One of the major limitations of this type of system is that it relies on prior information of the noise sources. Therefore, it is difficult for this type of system to handle arbitrary environmental noise. In this chapter, a new approach to real world Speaker recognition has been proposed. This technique includes re-creating the enrollment models of the speakers based on the signal to noise ratio (SNR) and noise identity estimated from signals submitted to the system for recognition. This technique is called “**Training on the fly**” for speaker recognition or “ **Adaptive noisy training model**” for speaker recognition. The next section provides a description of the proposed system. In Section 8.2 an experimental setup and result to evaluate the proposed technique has been presented. Finally, a summary and discussion are given in Section 8.3. This work is published by the author at AES International Conference on Audio Forensics-Finding Signal in the Noise. Audio Engineering Society (Al-Noori et al., 2017).*

8.1 Proposed Techniques Description

To deal with the mismatches that occur between the reference models and recognition input signal as a result of additive environmental noise, a "Training on the Fly" approach has been proposed (Al-Noori et al., 2017). This technique is based on generating an adaptive noisy model to become close enough to the recognition signal, depending on signal to noise ratio (SNR) and identity of noise that contaminated this recognition signal. The adaptive and creative model will have the same noise characteristics that are embedded with the recognition signal to decrease the mismatch between them in recognition phase and give more accurate recognition decisions. Figure 8.1 gives a block diagram of the proposed method.

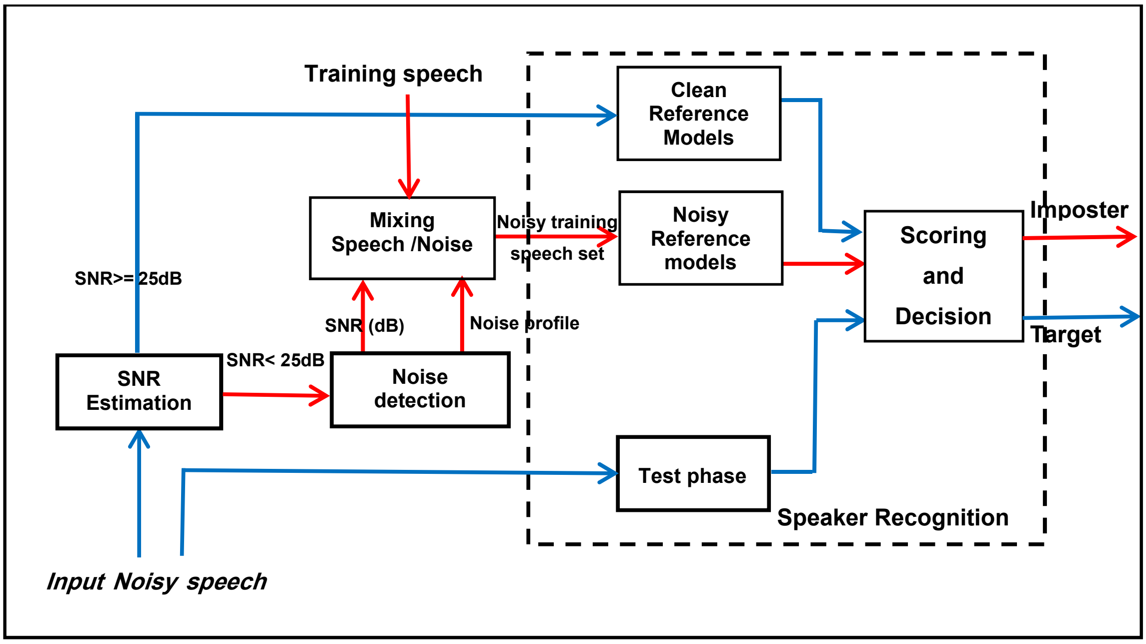


Figure 8.1: Training on the fly Block Diagram

The steps of Training on the Fly are as follows:

1. The input signal (the recognition signal) is first passed through an SNR estimator to find the signal to noise ratio of this signal.
2. If SNR of input signal is higher than the specified SNR threshold (in this work the threshold is specified to 25 dB since it is noticed that the accuracy of the speaker recognition starts decreasing from 20 dB SNR for different types of noise; however, this threshold may vary

from one type of noise to another) then the speech signal is identified as being clean, thus the baseline model(s) trained on clean speech is used.

3. Otherwise, if the SNR of the signal is less than the specified threshold (in this case less than 25dB) the noisy input signal is passed through a noise signal detection to determine the identity of the noise that is contaminating the speech signal. For this process, different methods have been adopted. The first one is based on detecting the profile of the noise and then generating an emulated noise close enough to the one that contaminates the speech signal as described in the next section. The second method is based on extracting the noise directly from the noisy input signal using one of the speech cleaning algorithms; more details of this method are given in the next chapter.
4. The detection noise (the emulated or extracted one) is then mixed with the enrolment signals (the training signal) for the claimed speaker based on the estimated SNR.
5. Finally, the new mixture signals are passed through Feature extraction and modelling stages to generate a noisy enrolment model for the claimed speaker. After that, this model is used in the recognition phase to match the input signal to find whether this signal belongs to the claimed person or an imposter.

   In the next section, further details about the SNR estimation and noise detection are presented.

8.1.A Signal to Noise Ratio (SNR) Estimation

There are several techniques to estimate the signal to noise ratio (SNR). A simple method to estimate SNR based on the time domain envelope of the noisy speech signal was adopted. This was done by measuring the distance of the envelope from the time axis (x axis) since the time envelope is shifted away from the time axis when the speech signal is noisy, whereas it is close to the time axis when the signal is clean. Figure 8.2 shows the envelope of 6 seconds of the same speech signal in clean and noisy environments. The first graph, Figure 8.2 -A, is the signal in a noise free environment and Figure 8.2-B represents the extracted envelope. Figure 8.2-C is the same signal contaminated with cafeteria speech babble noise and Figure 8.2-D represents its envelope. The increased distance between the signal envelope and the time axis can be clearly seen.

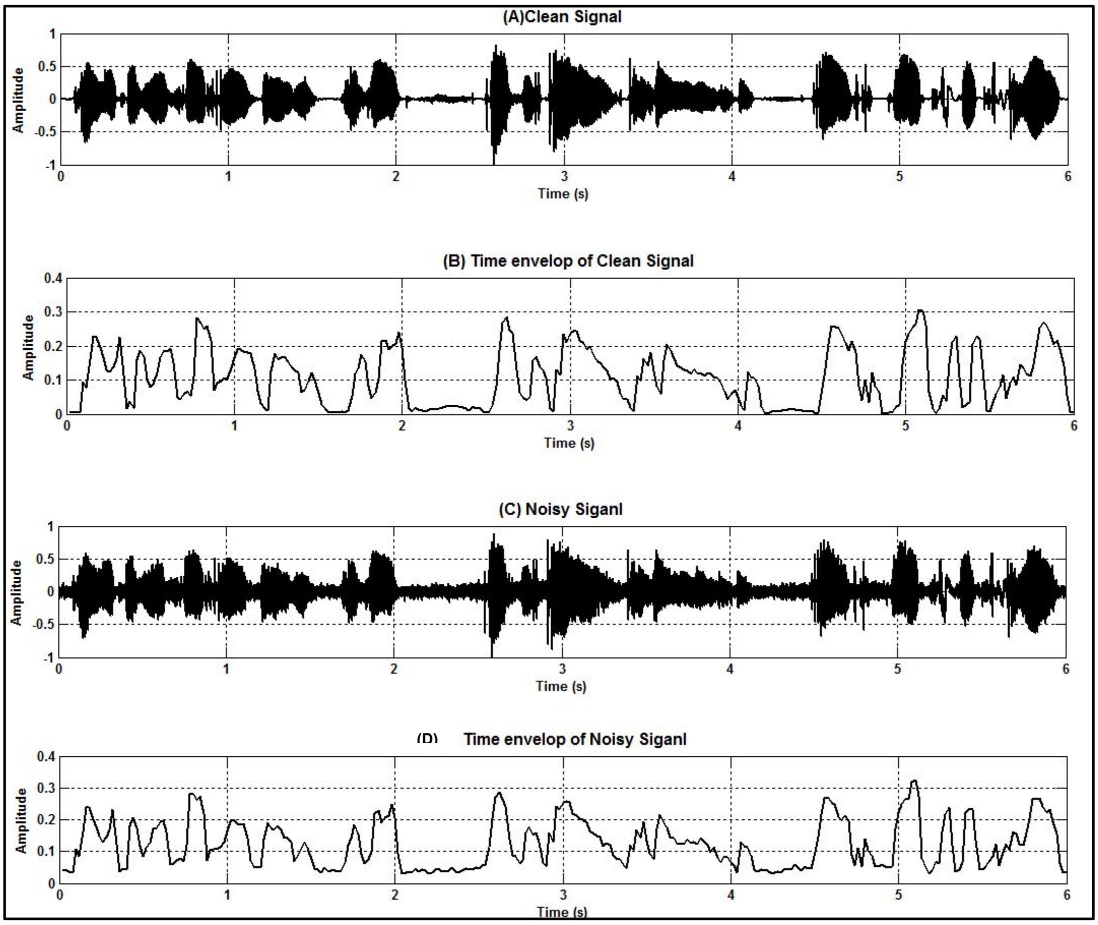


Figure 8.2: Envelope Detection for clean/Noisy signal

To estimate the SNR of the speech signal based on the time envelope the following steps are followed:

1. The envelope ev(t) of the speech signal x(t) is determined based on the Hilbert transform envelope detector as follows:

$$\boldsymbol{ev(t) = \sqrt{x^2(t) + \hat{x}(t)^2}} \qquad \textbf{(8.1)}$$

where $\hat{x}(t)$ is the Hilbert transform of the speech signal $x\ (t)$ defined by:

$$\boldsymbol{\hat{x}(t) = H[x(t)] = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(t-t')}{t'}\, dt'} \qquad \textbf{(8.2)}$$

Then, a low-pass filter is applied to decimate the sample of the envelope signal. Figure 8.2-A- shows 6 seconds of the speech signal and its decimated envelope is shown in Figure 8.2-B-.

2. The extracted envelope is then divided into *N* non-overlapping frames of size *M* and the minimum sample value for each frame is found (in this work the number of frames N is 10 and the size of each frame M is 30 samples). Finally, the mean of these minimum values is calculated namely $\boldsymbol{\alpha}$. This value represents the threshold between the speech and the noise such that any sample less than or equal to $\boldsymbol{\alpha}$ belongs to the noise region, while samples greater than $\boldsymbol{\alpha}$ belong to the speech region.

$$\boldsymbol{S} \in \begin{cases} \boldsymbol{S_{noise}} & ,s\leq\alpha \\ \boldsymbol{S_{speech}} & ,s>\alpha \end{cases} \qquad \textbf{(8.3)}$$

where ***s*** represent the samples of the envelope.

3. The power of the noise, *P*noise, based on samples from noise region S_{noise} is found where the power of the noise is the sum of the absolute squares of its time-domain samples divided by the signal length, or, equivalently, the square of its RMS level:

$$\boldsymbol{P_{noise}} = \frac{\boldsymbol{1}}{\boldsymbol{T}} \times \sum_{j=1}^{J} |\boldsymbol{S_{noise}(j)}|^2 = \boldsymbol{RMS(S_{noise})}^2 \qquad \textbf{(8.4)}$$

where ***T*** is the length of the signal, ***J*** no. of samples belonging to noise.

4. The power of the whole signal, *P*, is found; then *P*noise is subtracted from it to obtain the power of the speech P_{Speech} .

$$\boldsymbol{P} = \frac{\boldsymbol{1}}{\boldsymbol{T}} \sum_{t=1}^{T} \left|\boldsymbol{S_{signal}(t)}\right|^2 = \boldsymbol{RMS\left(S_{signal}\right)}^2 \qquad \textbf{(8.5)}$$

$$\boldsymbol{P_{speech}} = \boldsymbol{P} - \boldsymbol{P_{noise}} \qquad \textbf{(8.6)}$$

5. Finally, to calculate the SNR in dB scalse, the following equation is used:

$$\boldsymbol{SNR_{dB}} = \boldsymbol{10\log\left(\frac{P_{speech}}{P_{noise}}\right)} \qquad \textbf{(8.7)}$$

8.1.B Noise Identity Detection

The next step in the 'Training on the fly' approach is how to detect the type of noise (the identity of the noise) from the input recognition signal. As mentioned earlier, two different techniques are proposed for this purpose. This chapter is focused on describing the first technique, while the second technique will be described later in Chapter 9. The first technique is based on estimating the noise profile from a noisy signal using 1/3 octave filter to generate an emulated noise close to the original which contaminated the input signal. Figure 8.3 illustrates the block diagram of the procedure by which the noise profile is obtained:

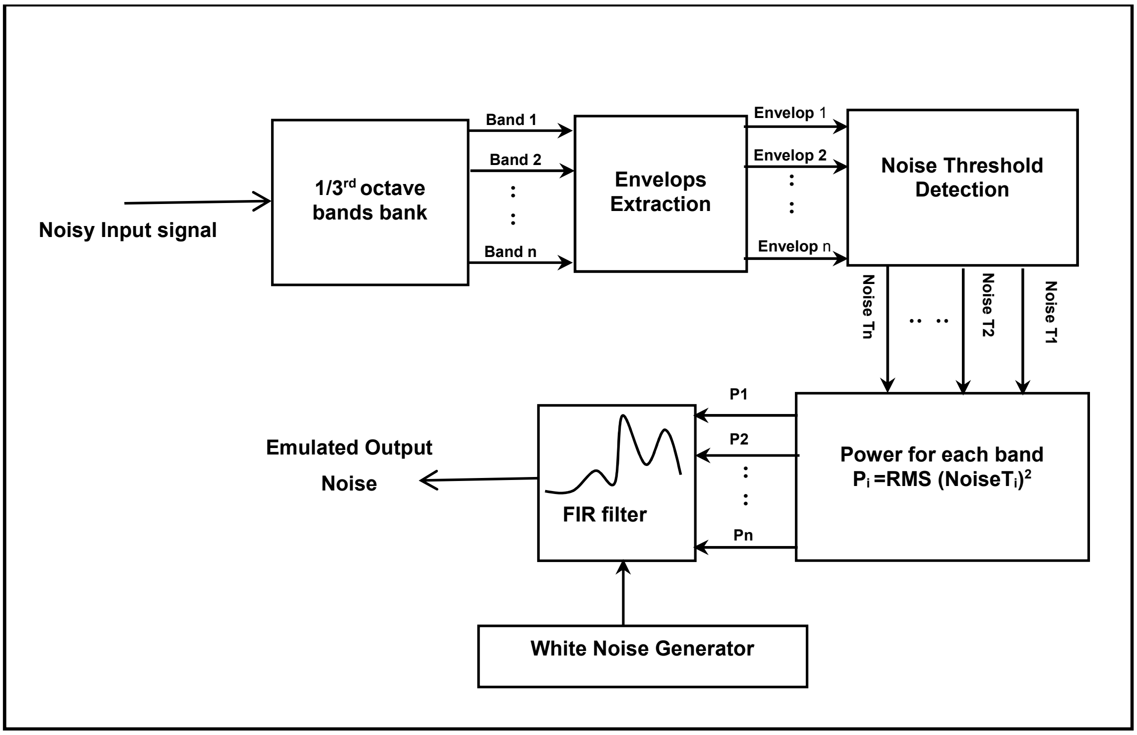


Figure 8.3: Noise Profile Estimation block diagram

The noise profile is obtained via the following procedure:

1. The noisy signal is passed through a 1/3-octave filter bank, giving M band-limited signals $f_i(t)$ where i represents the i^{th} band filter ($i=1...M$). The 1/3-octave band filter bank (or fractional octave band Bank) is a bank of bandpass filters where each filter determined by its centre frequency and its order (from 25Hz to 16 KHz). The magnitude attenuation limits are defined in the ANSI® S1.11-2004 standard. 1/3-octave filters are

usually employed to carry out spectral analysis for noise control since they offer a meaningful measure of the power of noise in various frequency bands. Since the sample rate of speech samples in this work is 16 kHz, the number of bands (M) used is 25 bands (from 25 Hz to 6.350 kHz). Figure 8.4 illustrates the filtered signal in each band.

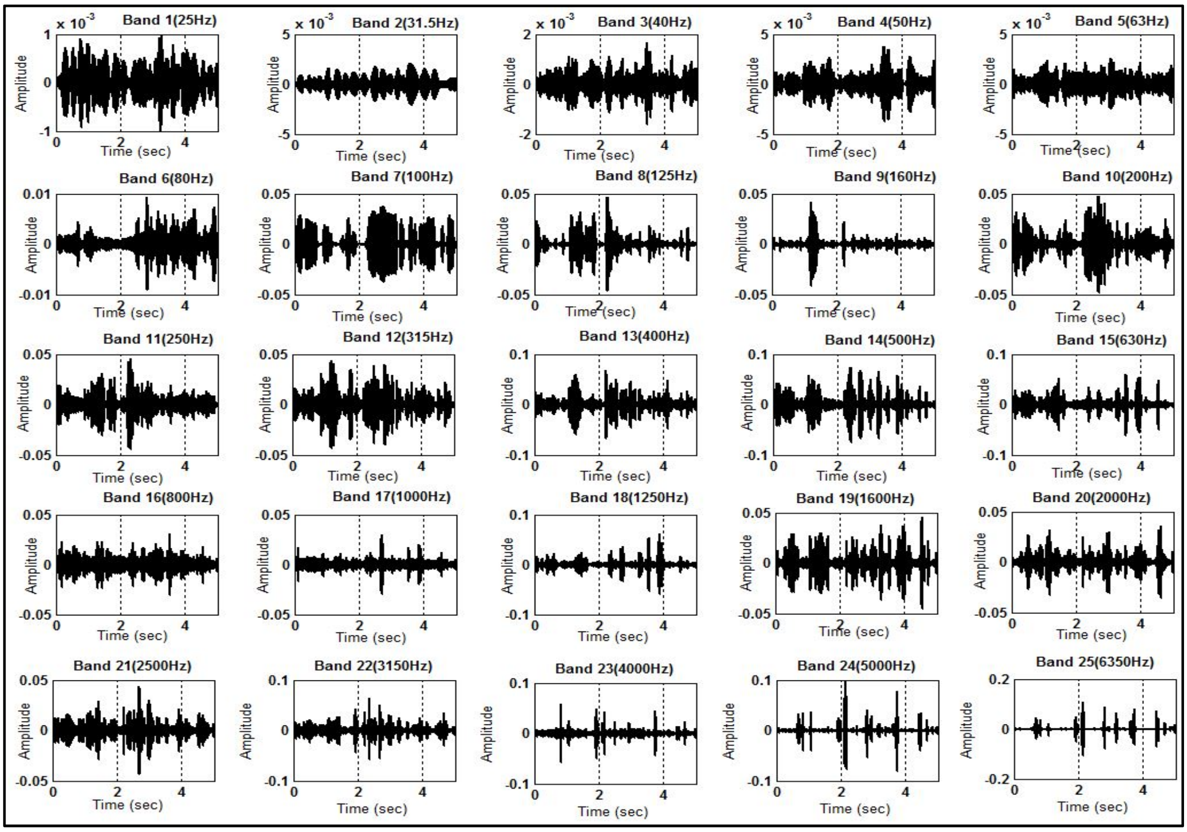


Figure 8.4: Filtered signals through 1/3 octave bands

2. The envelope ***e_i(t)*** for each filtered signal ***f_i(t)*** is extracted using Hilbert transform and low pass filter.

3. By using the same method in a SNR estimator, the noise threshold, **α**, is detected for each envelope $e_i(t)$, and then the samples that belong to the noise section can be detected. Figure 8.5 shows the extracted envelope for each band and threshold for each envelope (the red line).

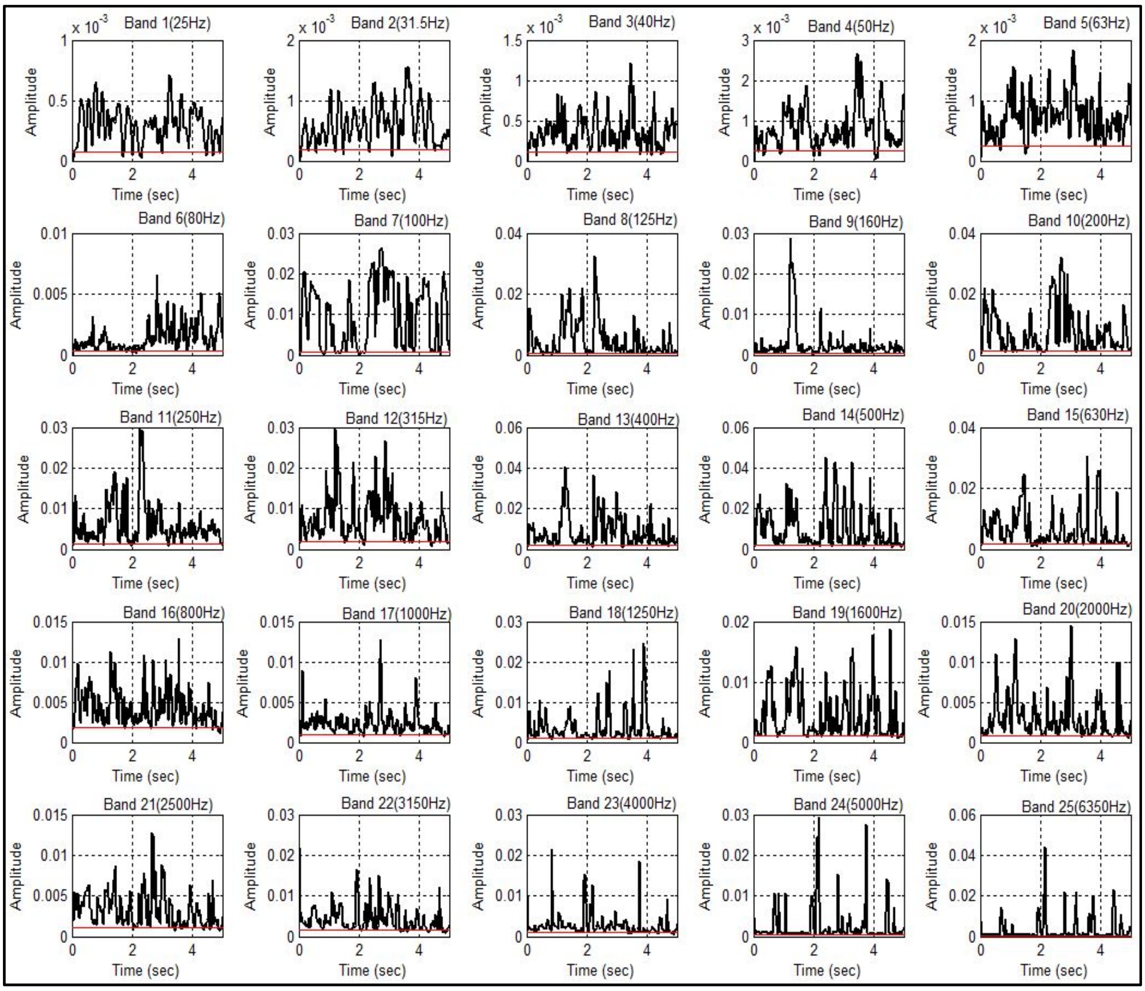


Figure 8.5: Extracted envelope for each band

4. Depending on the noise samples detected in step 3, the power for each band is found by squaring the root mean square (RMS) of noise samples such that:

$$\boldsymbol{P_i = RMS(d_i(u))^2} \qquad \textbf{(8.8)}$$

Where $d_i(u)$ represents the samples belonging to the noise region in each envelope $e_i(t)$. Then $\mathrm{P_i}$ is normalised using RMS normalisation.

The power spectrum *P* represents the estimated power of the noise within the signals $f_i(t)$. Figure 8.6 demonstrates the estimated power bands from a noisy signal contaminated with cafeteria babble (red bars) against the power bands of a reference noise sample (the blue bar) with the difference between them (the green bar) in linear and dB scale respectively.

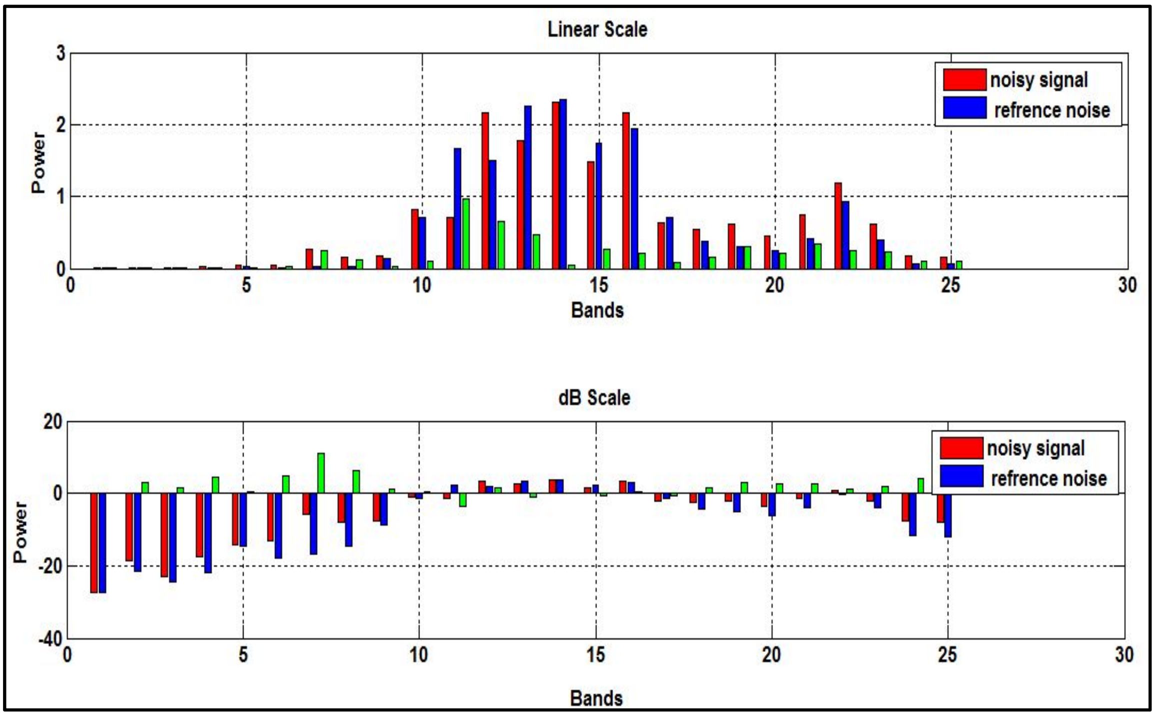


Figure 8.6: Power-band for Babble noisy signal (red bar) and reference signal (blue bar) with error ratio (green bar)

5. Finally, a bank of 1/3 octave FIR bandpass filters, designed to have a summative frequency response as defined by the spectrum *P*, is employed to shape white noise to generate an emulated noise that has the similar shape of original noise that contaminated the input signal as described in the following equation:

$$\boldsymbol{d_{emulated}(t) = \sum_{i=1}^{M} w_i(i).P_i} \qquad \textbf{(8.9)}$$

where $w_i(t)$ is, white noise filtered by 1/3 octave band *i*, $d_{emulated}$ is the output noise. This noise is used later in mixing procedure.

The main goal of using white noise is the characteristic of having the constant power spectrum density, which makes easy to shape this noise to create an emulated noise which is close to the power density of the original noise. Figure 8.7 shows the output emulated noise in time and frequency domain.

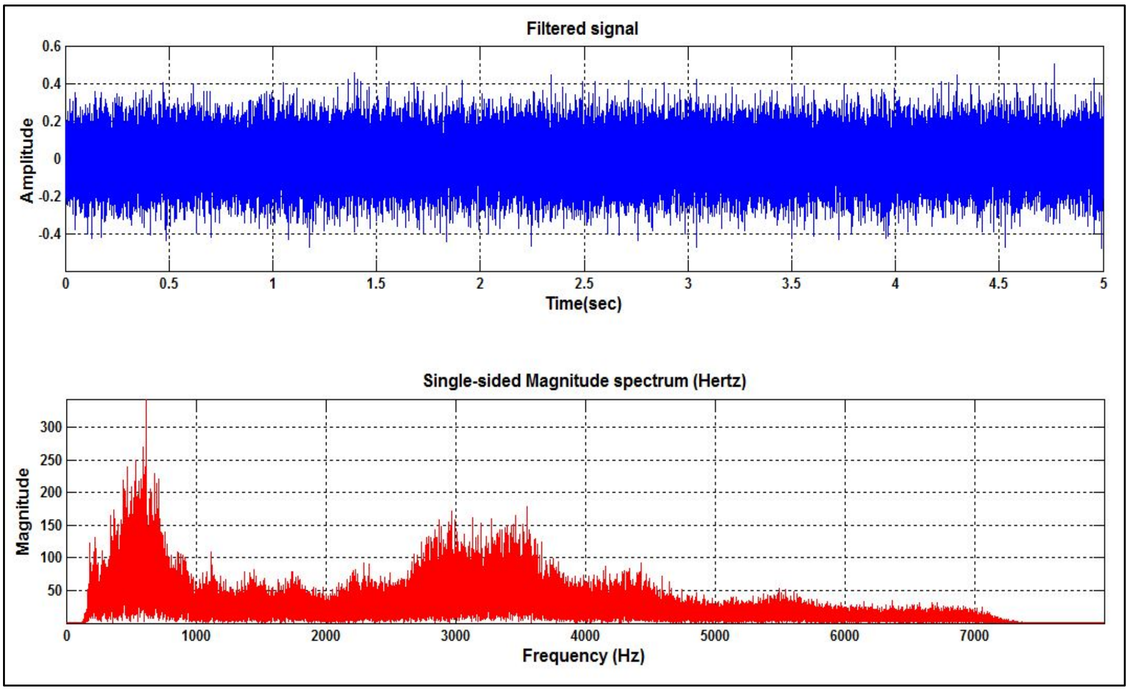


Figure 8.7: The emulated noise signal in Time/Frequency domains

8.1.C Speech/ Noise mixing procedure

After SNR and emulated noise are provided. the next procedure is to produce the noisy signals to create the noisy reference models. For this purpose, the enrolled speech signals for each claimed person are mixed with the emulated noise depending on the estimated SNR. The mixing procedure is similar to that described in Chapter 3 (Figure 3.2) to provide speech samples that are contaminated with noise with specific SNR. The output noisy signals are then passed through training phase stages (features extraction followed by modelling stage) to create a new model to use in matching with input recognition signal.

8.2 Proposal approach evaluation Experiments

The main aim of these experiments is to show to what extent the proposed approach can achieve significant progress in speaker recognition performance. To evaluate the efficiency of the Training on the Fly approach, two types of speech database have been adopted: TIMIT and SALU-AC. Furthermore, two kinds of features techniques are used, Mel Frequency Cepstrum Coefficients MFCC, which represents the most commonly used features in Speaker recognition fields, and Gammatone Cepstrum Coefficients, which represents one of the noise robust features as mentioned before.

8.2.A Experimental Setup

As mentioned before, the main purpose of these experiments is to show the improvements obtained from the proposed system compared with traditional speaker recognition systems (based on clean speech samples only). Furthermore, this study focuses mainly on showing that this type of technique can achieve a significant improvement in SR accuracy in a noisy environment using a suitable SNR estimation and noise profile detection. In this chapter the method explained in section 8.1B has been adopted, while a different method is adopted in the next chapter. The details of these experiments are as follows:

- Two types of features spaces are adopted for these experiments, MFCC including dynamic Features (delta and delta delta) with 39-dimensional space, and GFCC with 23-dimensional space, as mentioned in chapter 6.
- Gaussian Mixture Model was used with 256 mixture. The main reason for using GMM instead of using GMM-UBM is the requirement to create the model from scratch depending on the characteristics of additive noise (SNR and Noise profile) instead of adapting from a Background model. Otherwise, the whole background model would need retraining to make it possible to adapt a suitable model.
- To check the performance of 'on the fly' training, two datasets were used from two different databases, TIMIT and SALU-AC. Each dataset included speech samples from 60 speakers (30 male, 30 female). For each speaker, a set of 9 utterances were used in the training phase, all spoken in English, and one utterance for each speaker was used in the recognition phase, also spoken in English. In addition, the recognition set was divided into a number of subsets. Each subset represents the same speech samples mixed with different SNRs (20, 15, 10, 5, and 0 dB) for a specific type of noise. The noise types adopted in these experiments are Cafeteria speech babble, interior moving Car, interior moving train, Street, and White noise.
- To evaluate our approach in speaker recognition, two different metrics are used, Error Equal rate and detection Error trade-off. The evaluation is computed on the recognition scores (log-likelihood scores) based on 60 authentic cases (verified cases)

and 3540 inauthentic cases (imposter cases). The results of degradation through different SNR for each type of noise are compared with those obtained from the clean training mentioned in chapter 4.

8.2.B Experimental Results

A. MFCC Baseline Results

Figure 8.8 (a-e) shows the degradation of accuracy of SR in different noisy conditions with various SNRs based on the EER for the proposed method based Speaker recognition and conventional speaker recognition (based on reference model from clean signals) using SALU-AC. The abscissa represents the SNR in dB (20, 15, 10, 5, and 0 dB) while the ordinate represents EER (in %). It is clear that there is improved robustness in the performance of the proposed approach based speaker recognition to various noise types with different SNRs compared with the performance of the traditional system, specifically for car interior and white noise contaminated speech having low SNR. For instance, in the range between 15 and 5 dB in both types of noise (car interior, and white noise) there is greater improvement in EERs for proposed approach (0%, 0.02%, and 0.08%, for car interior noise and 3.33%, 5.31%, and 12.74% for white) compared to the traditional Speaker Recognition (0.16%, 0.48%, and 1.6% for car interior ,and 8.3%, 20.7%, 31.6 for white noise), while for speech samples contaminated with cafeteria babble and street noise there is still clear improvement, especially for the range between 10dB, and 0dB. However, the speech sample contaminated with train interior noise shows only slight improvement when compared to the other types of noise, possibly because this noise sample contains a number of sources which contribute to the noise signal, making the estimation of this type difficult compared to other types of noise.

Figure 8.9 (a to e) illustrates a DET graph for the proposed approach and traditional based SR for the five types of noisy speech samples with 10 dB SNR. It can be seen that the accuracy of false positive rate (FPR) for the training on the fly approach (the solid blue line) shows significant improvement compared to the traditional speaker recognition, especially for car interior and White Noise. Conversely, train interior noise shows the same accuracy for both systems since, as mentioned before, the estimation of this kind of noise is difficult.

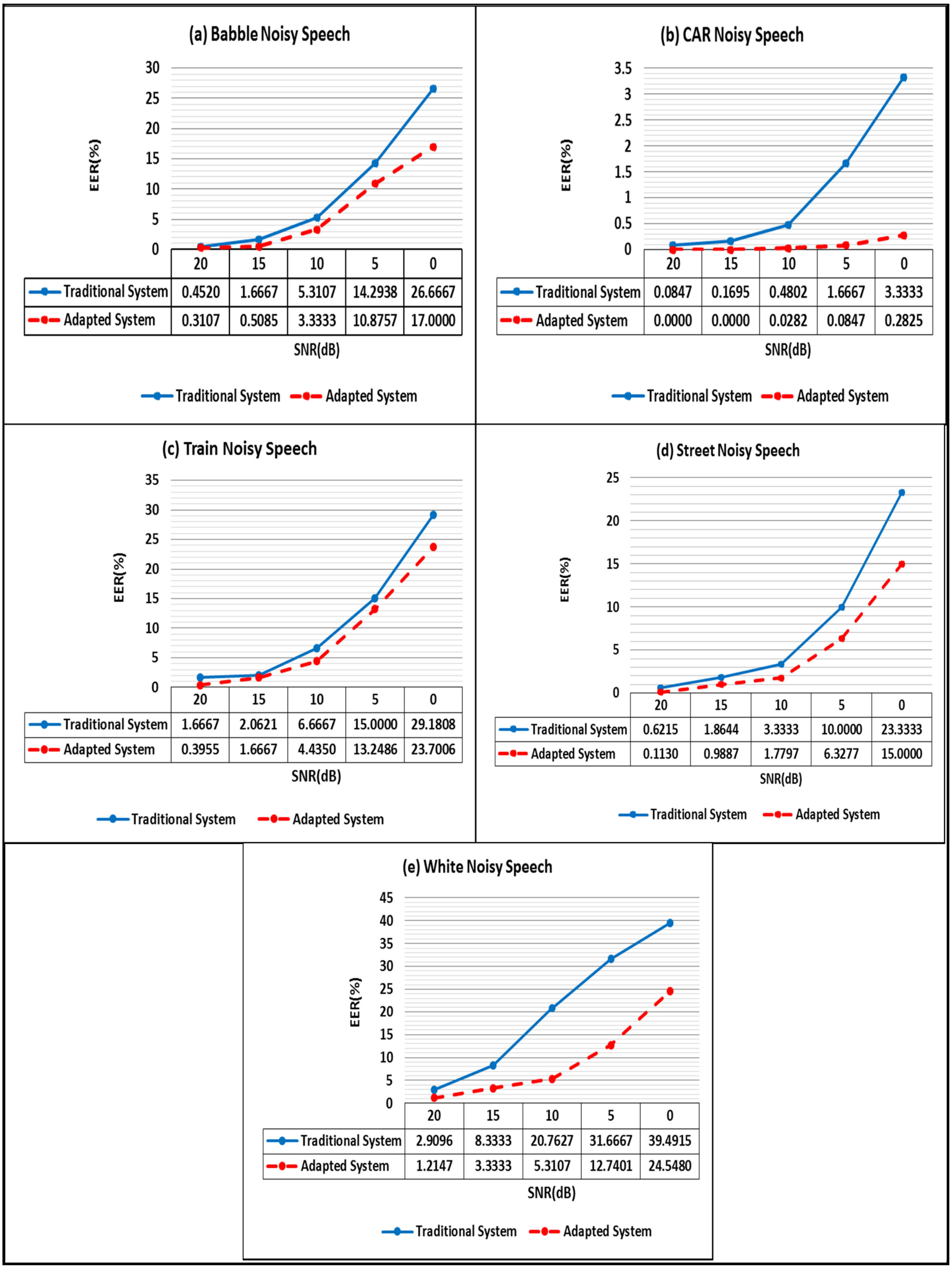


Figure 8.8: EER of proposed approach and clean training based SR using SALU-AC

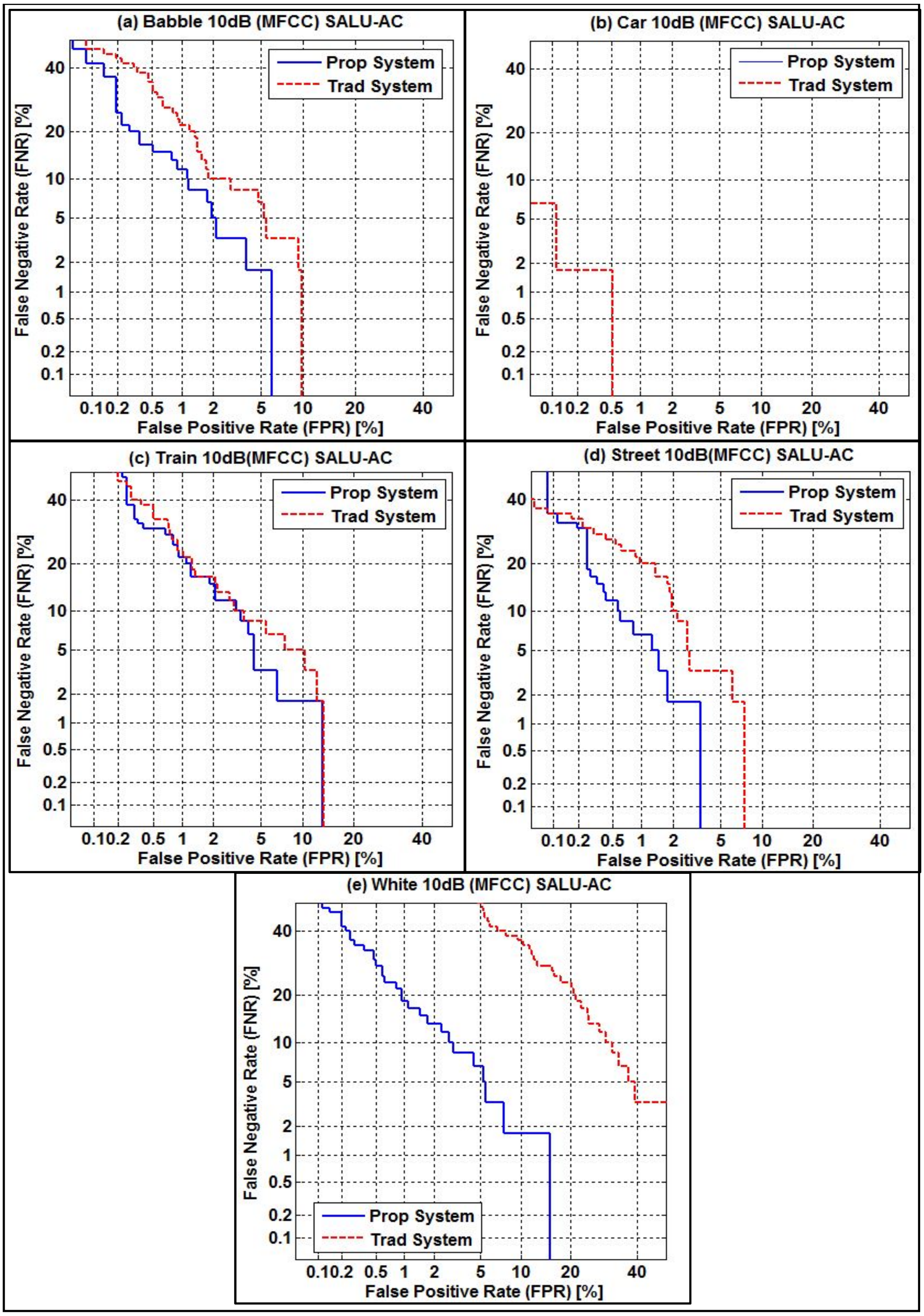


Figure 8.9: DET graph for proposed approach based SR and Traditional SR in 10 dB SNR using SALU-AC

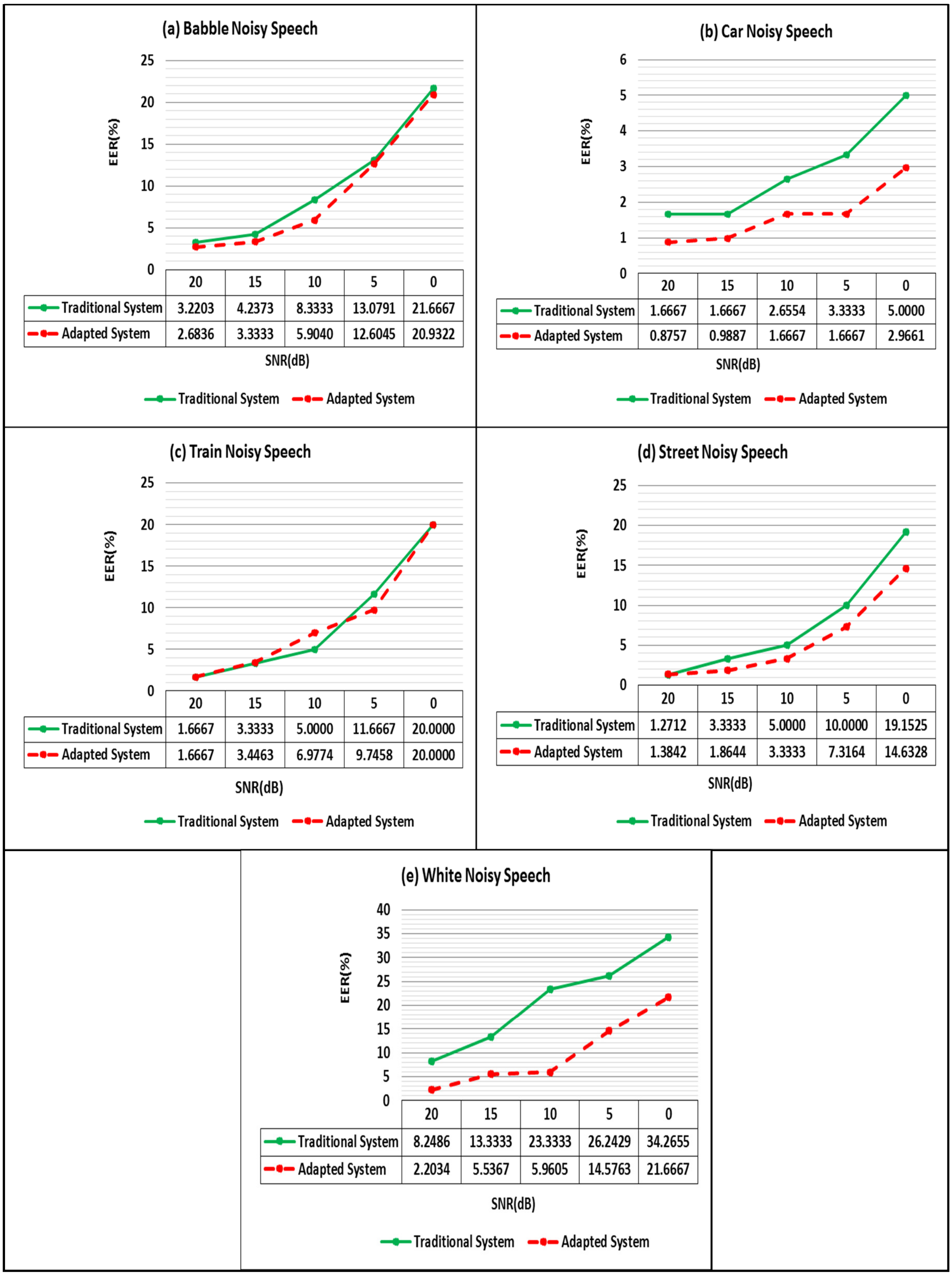
Experimental results using the TIMIT dataset are shown in Figure 8.10 and it is clear that the training on the fly approach shows improvement in the accuracy of the speaker recognition for both speech samples contaminated with car interior and white noise over the range of SNRs, particularly for low SNR, while both babble and street noise show promising progress in EER of this approach, specifically in range 10 and 0 dB when compared with SR based on clean training.

The main reason that the improvements involving these types of noise are less dramatic when compared to car interior and white noise can be attributed to the difficulty of detecting the profile of these types of noise, since they contain different sources of noises, compared to the detection of noise profiles such as car interior and white noise, which have very limited sources contributing to the noise signal. Furthermore, we can see there are still some limitations in EER for train interior noise for the same reasons.

Figure 8.11 shows DET graphs for the five types of noise in 15dB SNR for TIMIT database where the solid green line represents the training on the fly approach and the dashed red line represents the conventional SR.

It is clear that there is a noticeable improvement in four types of noise (Cafeteria speech babble, interior car, street and white noise), especially in FPR. In this Figure, it can also be seen that improvement involving some kinds of noise (car, and white noise) is better compared with these that include different resources (babble, street, and train). However, there is still some limitation in FNR of some types of noisy data (such as in car interior and train interior); but not substantially different, since the effect of environmental noise is focusing more on False positive rate than False negative rate.

From these experiments, we conclude that given a good SNR estimation and noise profile estimation detection from the input recognition signal, an adaptive model can be re-trained using these parameters in order to improve the performance of SR system in noisy environments. We call this kind of system a ‘training on the fly speaker recognition system’.



(a) Babble Noisy Speech

	20	15	10	5	0
Traditional System	3.2203	4.2373	8.3333	13.0791	21.6667
Adapted System	2.6836	3.3333	5.9040	12.6045	20.9322

(b) Car Noisy Speech

	20	15	10	5	0
Traditional System	1.6667	1.6667	2.6554	3.3333	5.0000
Adapted System	0.8757	0.9887	1.6667	1.6667	2.9661

(c) Train Noisy Speech

	20	15	10	5	0
Traditional System	1.6667	3.3333	5.0000	11.6667	20.0000
Adapted System	1.6667	3.4463	6.9774	9.7458	20.0000

(d) Street Noisy Speech

	20	15	10	5	0
Traditional System	1.2712	3.3333	5.0000	10.0000	19.1525
Adapted System	1.3842	1.8644	3.3333	7.3164	14.6328

(e) White Noisy Speech

	20	15	10	5	0
Traditional System	8.2486	13.3333	23.3333	26.2429	34.2655
Adapted System	2.2034	5.5367	5.9605	14.5763	21.6667

Figure 8.10: EER of proposed approach and clean training based SR using TIMIT

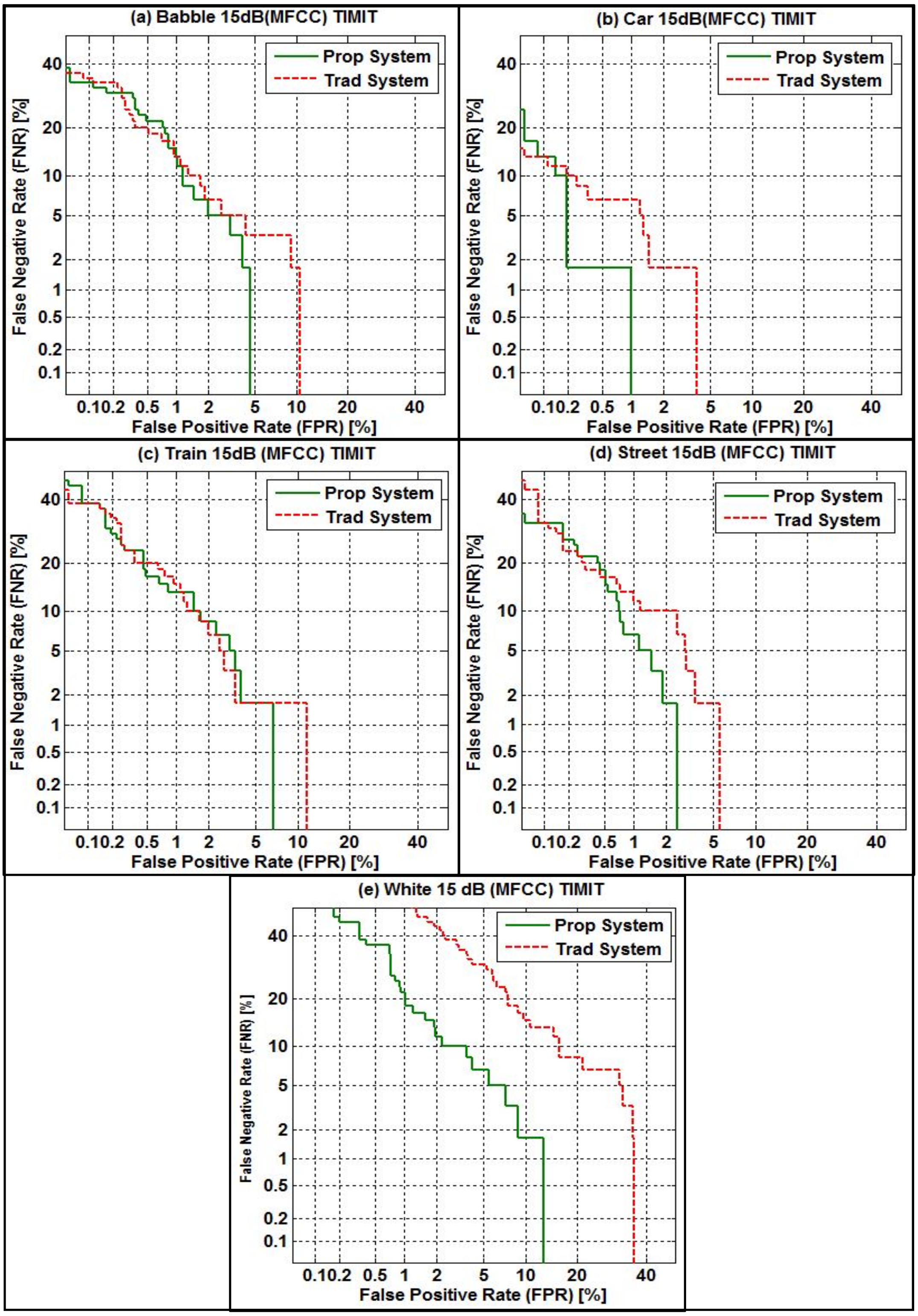


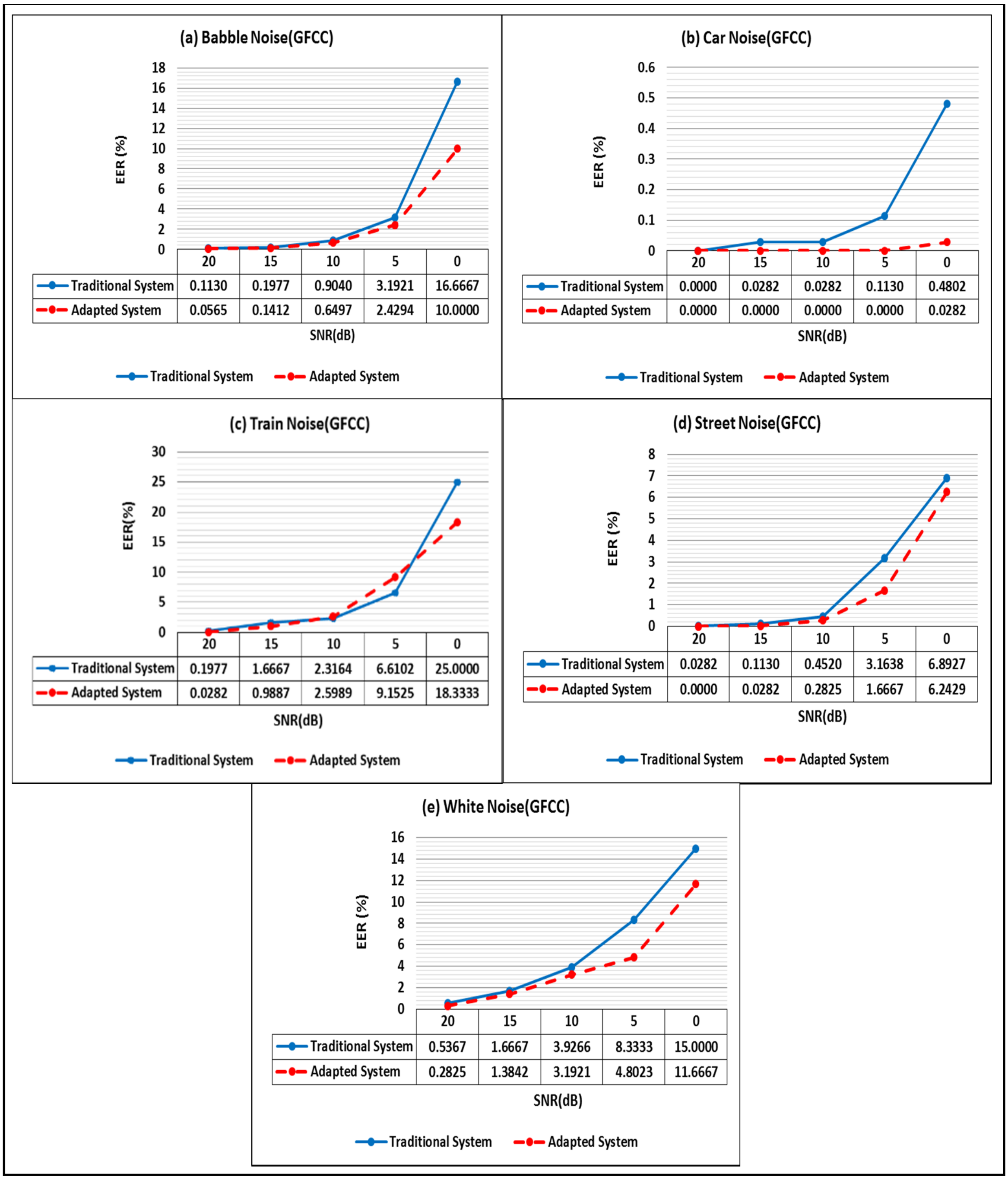
Figure 8.11:DET graph for proposed approach based SR and Traditional SR in 15 dB SNR using TIMIT

B. GFCC Baseline Results

When a Training on the fly approach is used with GFCC baseline SR on SALU-AC dataset, the EER results show limited improvement, especially in high SNRs, over conventional speaker recognition, compared with those obtained from MFCC based Speaker recognition. On the other hand, the proposed approach shows promising progress in low SNR (especially in 5 and 0 dB) for some types of the noisy speech sample. Figure 8.12 a-e show the degradation of EER through SNRs for the proposed approach (the dashed red line) and conventional speaker recognition (based on clean training) (the bold blue line) using GFCC based Speaker recognition and SALU-AC as a dataset. It is clear that there is a minor improvement of EER for the adapted approach over clean training based speaker recognition in the SNR range between 20 dB and 10 dB for the four types of noisy speech samples (Cafeteria speech babble, car interior, street, and white noise) in Figure 8.12 a, b, d, and e. However, the EERs at 5dB and 0dB SNRs show promising improvement for speech samples contaminated with car interior, street and white noise.

Conversely, the EERs for speech samples contaminated with train interior noise show that the proposed approach has failed to achieve improvement in performance of speaker recognition since the results obtained from these samples are variant from one SNR to another. For example, although there is very limited improvement in 20 dB and 15 dB with 0.02% and 0.9% EER respectively over clean training SR with 0.19% and 1.66% EER respectively for the same SNRs, the EER at 5dB SNR shows the clean training speaker recognition outperforming the proposed approach with 6.6 % EER, with just 9.15% for Training on the fly based SR. The main reason for this limitation is the difficulty in estimating the profile of this type of noise, since it includes different sources which make the created emulated noise not close enough to the original one, as mentioned earlier.

Figure 8.13 a to e illustrate the detection error trade-off (DET) at 05 dB for the five types of noisy speech sample using GFCC features extraction approach and SALU-AC dataset. It is clear there is limited improvement for the proposed approach in False positive rate (FPR) and False negative rate (FNR) (the bold blue curve) over conventional Speaker recognition (dashed red curve) in noise samples contaminated with cafeteria speech babble (Figure 8.13 a). Conversely the speech samples contaminated with car interior, Street, and white noise (Figure 8.3 b, d, and e) show significant improvement in both FPR and FNR, especially in the car interior noise sample. However, the train interior noisy speech sample showed degradation in FPR and very limited improvement in FNR (Figure 8.13 c).


(a) Babble Noise(GFCC)

SNR(dB)	20	15	10	5	0
Traditional System	0.1130	0.1977	0.9040	3.1921	16.6667
Adapted System	0.0565	0.1412	0.6497	2.4294	10.0000

(b) Car Noise(GFCC)

SNR(dB)	20	15	10	5	0
Traditional System	0.0000	0.0282	0.0282	0.1130	0.4802
Adapted System	0.0000	0.0000	0.0000	0.0000	0.0282

(c) Train Noise(GFCC)

SNR(dB)	20	15	10	5	0
Traditional System	0.1977	1.6667	2.3164	6.6102	25.0000
Adapted System	0.0282	0.9887	2.5989	9.1525	18.3333

(d) Street Noise(GFCC)

SNR(dB)	20	15	10	5	0
Traditional System	0.0282	0.1130	0.4520	3.1638	6.8927
Adapted System	0.0000	0.0282	0.2825	1.6667	6.2429

(e) White Noise(GFCC)

SNR(dB)	20	15	10	5	0
Traditional System	0.5367	1.6667	3.9266	8.3333	15.0000
Adapted System	0.2825	1.3842	3.1921	4.8023	11.6667

Figure 8.12: EER of proposed approach and clean training based SR using SALU-AC (GFFC based)

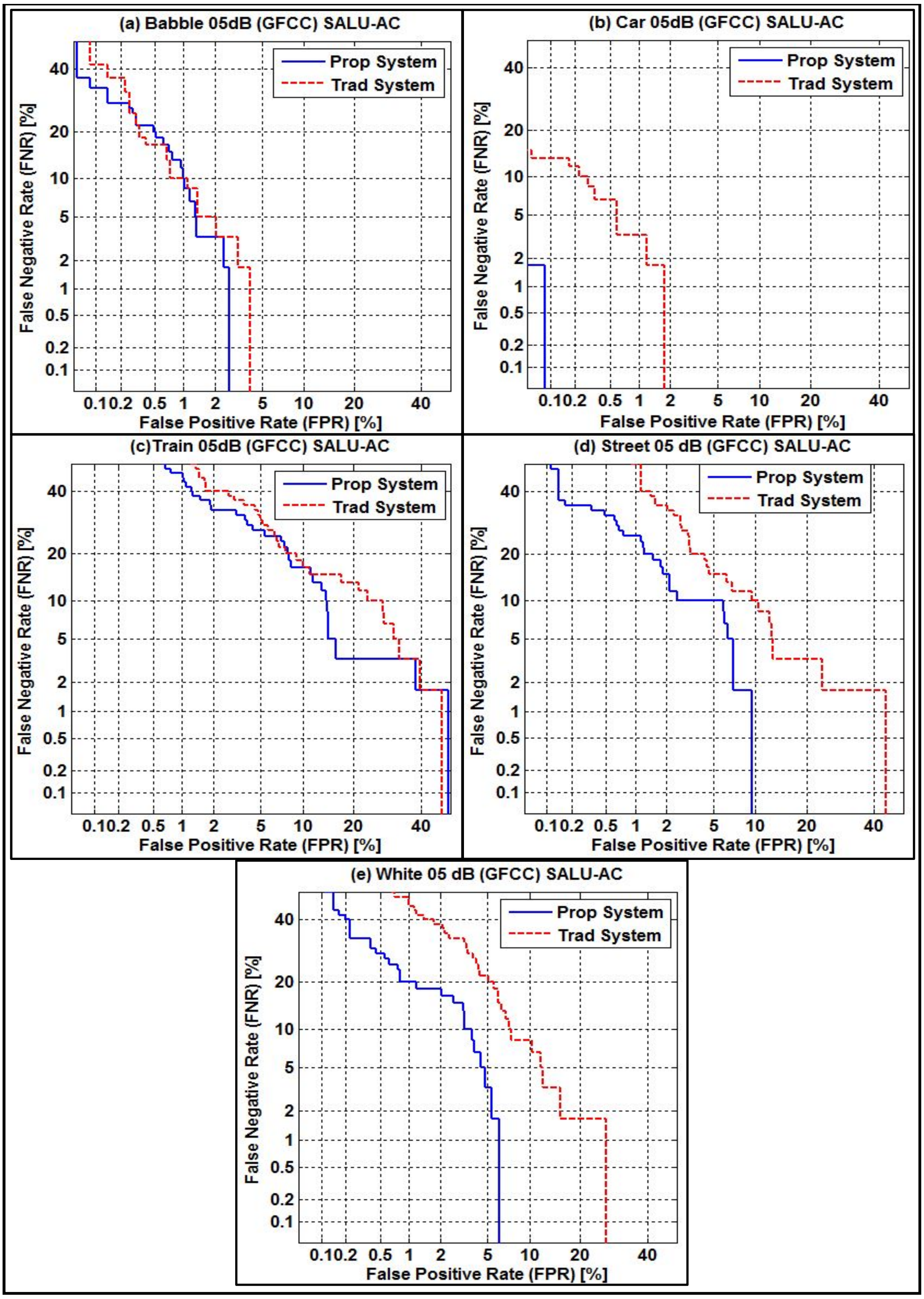
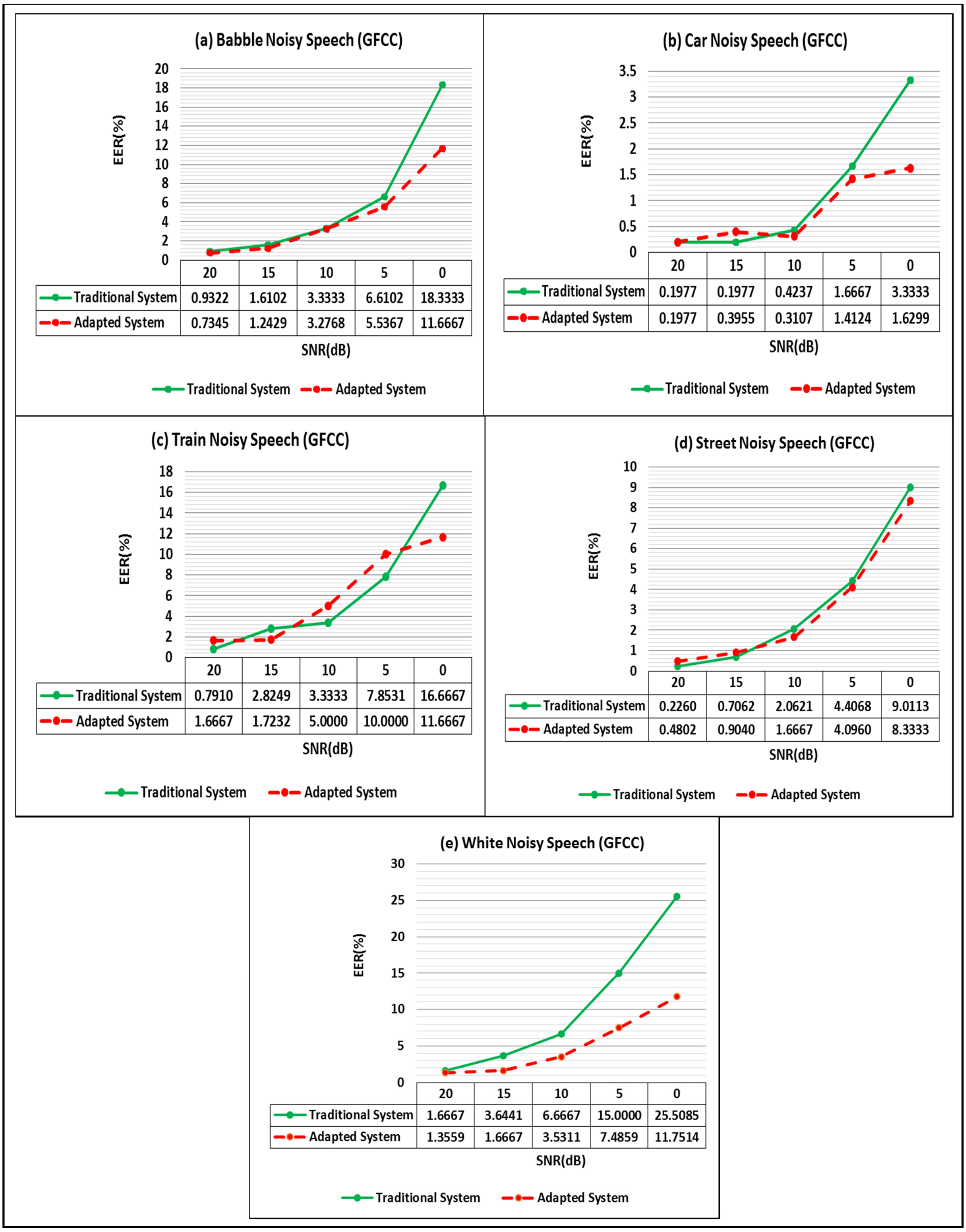


Figure 8.13: DET graph for proposed approach based SR and Traditional SR in 05 dB SNR using SALU-AC (GFCC Based SR)

When the TIMIT dataset was used (Figure 8.14 a to e) , the EER results showed variation from one noisy speech sample to another. In general, however, there are still the same limitations in the performance of the proposed approach in the range between 20 dB and 10 dB for whole types of noise. Furthermore, in some cases the speaker recognition based clean training outperformed the proposed approach based SR, as seen in 20 and 15 dB for street noise with 0.22 % and 0.7% EER respectively for conventional SR, against 0.48 % and 0.9% EER respectively for the proposed approach at the same SNRs (Figure 8.14 d). Again, the speech samples contaminated with train interior noise show the same variant in EERs through the different SNRs.

Figure 8.15 a to e show DET curve for training on the fly speaker recognition (the bold green curve) and clean training based speaker recognition (the dashed red curve) at 10 dB using GFCC based Features and TIMIT dataset. As is clear in this figure, the FPR improvements for the proposed approach are very limited for speech samples contaminated with cafeteria speech babble, car interior, and street noise, while the speech samples contaminated with the white noise show significant improvement in FPR and FNR.

The main reason that Training on the Fly is more efficient with MFCC features than GFCC feature is the robustness of GFCC to different environmental noises, which means this type of feature can deal with noisy speech without the requirement for the noisy training model, especially for high SNRs (20,15 and 10 dB). However, for low SNRs, there is still a significant improvement in speaker recognition performance which makes this approach with GFCC possible with low SNRs by changing the SNRs threshold as mentioned in section 8.1.



(a) Babble Noisy Speech (GFCC)

	20	15	10	5	0
Traditional System	0.9322	1.6102	3.3333	6.6102	18.3333
Adapted System	0.7345	1.2429	3.2768	5.5367	11.6667

(b) Car Noisy Speech (GFCC)

	20	15	10	5	0
Traditional System	0.1977	0.1977	0.4237	1.6667	3.3333
Adapted System	0.1977	0.3955	0.3107	1.4124	1.6299

(c) Train Noisy Speech (GFCC)

	20	15	10	5	0
Traditional System	0.7910	2.8249	3.3333	7.8531	16.6667
Adapted System	1.6667	1.7232	5.0000	10.0000	11.6667

(d) Street Noisy Speech (GFCC)

	20	15	10	5	0
Traditional System	0.2260	0.7062	2.0621	4.4068	9.0113
Adapted System	0.4802	0.9040	1.6667	4.0960	8.3333

(e) White Noisy Speech (GFCC)

	20	15	10	5	0
Traditional System	1.6667	3.6441	6.6667	15.0000	25.5085
Adapted System	1.3559	1.6667	3.5311	7.4859	11.7514

Figure 8.14: EER of proposed approach and clean training based SR using TIMIT (GFFC based)

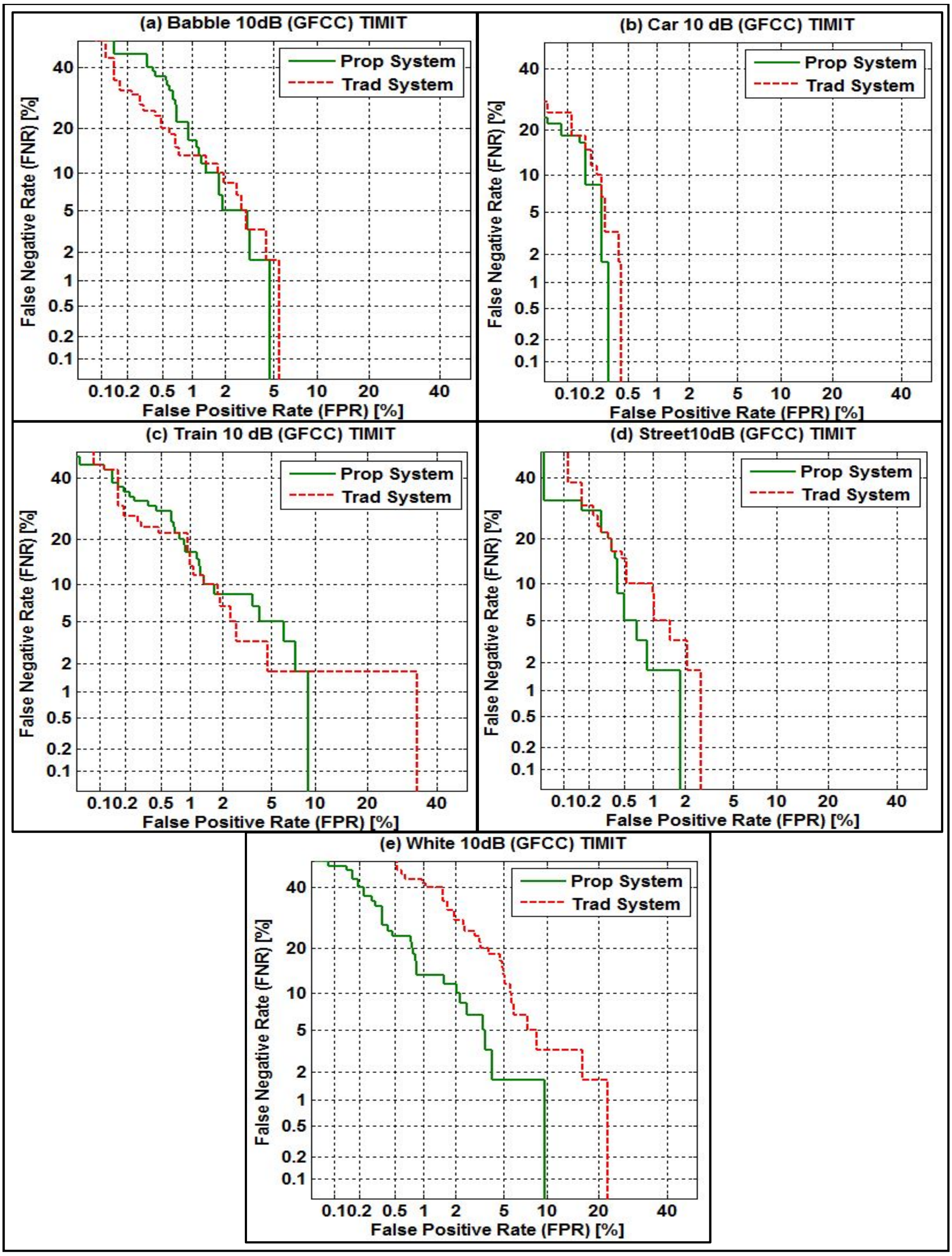


Figure 8.15: DET graph for proposed approach based SR and Traditional SR in 10 dB SNR using TIMIT (GFCC Based)

Another important point from the previous experiments is specification the best SNRs that used to provide the adapted noisy reference model. In section 8.1.A simple method are proposed to estimate the SNRs of the noisy recognition signal, but the question is how can provide the best noisy reference model that gives the best performance based on the suitable SNR. This value is based mainly on the type of noise and the level of SNRs estimated from the signal. In general, however, based on the experiments made in Chapter 5 (Figure 5.7) and the experiments of **Training on the Fly,** if estimated SNR is high (between 20 and 10 dB) then the best speaker recognition performance can be obtained when the SNR used to produce the noisy reference model is equal or close to estimated SNR. If the estimated SNR is low (between 10 and 0) then the best performance can be obtained when the SNR used to generate the adapted reference model is greater than the estimated SNR. For example, if estimated SNR for input signal is 17dB then the adapted reference noisy model should ideally have the same SNR, while if the estimated SNR is 3 dB then the adapted model is generating with 5 or 6 dB SNR.

Table 8-1 and Table 8-2 present the summary of EERs of previous experiments using MFCC and GFCC based speaker recognition for each data set.

Table 8-1: EER for MFCC and GFCC based Speaker recognition using SALU-AC dataset

Noise type	SNR	Est. SNR (dB)	Train SNR (dB)	MFCC Clean Training	MFCC Training on the fly	GFCC Clean Training	GFCC Training on the fly
Cafeteria Babble	20dB	21.5dB	21d.B	0.45	0.31	0.11	0.05
	15dB	17.3	17	1.66	0.5	0.19	0.14
	10dB	12.5	15	5.31	3.33	0.9	0.64
	05dB	7.2	10	14.29	10.87	3.19	2.42
	0 dB	4.1	5	26.6	17	16.66	10
Interior moving car	20dB	21.4	20	0.08	0	0	0
	15dB	16.7	17	0.169	0	0.022	0
	10dB	12.5	15	0.48	0.02	0.028	0
	05dB	8.1	10	1.66	0.08	0.11	0
	0 dB	4.3	5	3.3	0.28	0.48	0.02
Interior moving Train	20dB	19.3	19	1.6667	0.39	0.19	0.02
	15dB	16.1	16	2.06	1.66	1.66	0.98
	10dB	11.4	15	6.66	4.43	2.31	2.59
	05dB	7.6	10	15	13.24	6.61	9.15
	0 dB	4.2	5	29.18	23.7	25	18.33
Street	20dB	19.3	19	0.62	0.11	0.028	0
	15dB	15.4	15	1.86	0.98	0.11	0.02
	10dB	11.3	15	3.33	1.77	0.45	0.28
	05dB	6.7	10	10	6.32	3.16	1.66
	0 dB	3.5	5	23.3	15	6.89	6.24
White noise	20dB	19.8	20	2.9	1.21	0.53	0.28
	15dB	14.7	15	8.33	3.33	1.66	1.38
	10dB	9.9	15	20.76	5.31	3.92	3.19
	05dB	5.4	10	31.66	12.74	8.33	4.80
	0 dB	1.1	5	39.49	24.54	15	11.66

Table 8-2: EER for MFCC and GFCC based Speaker recognition using TIMIT dataset

Noise type	SNR	Est. SNR	Train SNR	MFCC Clean Training	MFCC Training on the fly	GFCC Clean Training	GFCC Training on the fly
Cafeteria Babble	20dB	19.7	20	3.22	2.68	0.93	0.73
	15dB	15.4	15	4.23	3.33	1.61	1.24
	10dB	11	15	8.33	5.9	3.33	3.27
	05dB	6.6	10	13.07	11.66	6.61	5.53
	0 dB	3.8	5	21.66	19.03	18.33	11.66
Interior moving car	20dB	19.9	19	1.66	0.87	0.19	0.19
	15dB	16.3	16	1.66	0.98	0.19	0.39
	10dB	12.3	15	2.65	1.66	0.42	0.31
	05dB	7.8	10	3.33	1.66	1.66	1.41
	0 dB	3.9	5	5	2.96	3.33	1.62
Interior moving Train	20dB	19.7	20	1.66	1.66	0.79	1.66
	15dB	15.3	15	3.33	3.44	2.82	1.72
	10dB	10.5	15	5	6.97	3.33	5
	05dB	5.6	10	11.66	9.74	7.85	10
	0 dB	3.7	5	20	20	16.66	11.66
Street	20dB	19.7	20	1.27	1.38	0.22	0.48
	15dB	15.1	15	3.33	1.86	0.7	0.9
	10dB	11.6	15	5	3.33	2.06	1.66
	05dB	7.2	10	10	7.31	4.4	4.09
	0 dB	2.2	5	19.15	14.63	9.01	8.33
White noise	20dB	19.6	20	8.24	2.20	1.66	1.35
	15dB	15.4	15	13.33	5.53	3.64	1.66
	10dB	9.8	15	23.33	5.96	6.66	3.53
	05dB	5.2	10	26.24	14.57	15	7.48
	0 dB	1.6	5	34.26	21.66	25.5	11.75

Based on the results of previous experiments, the following conclusions can be summarised as follows:

- Training on the Fly shows significant improvement in speaker recognition performance when compared with conventional speaker recognition, especially with stationary noise (such car interior and white noise). For non-stationary noise, the proposed approach showed promising improvement.
- The results show that the training on the fly approach is more efficient when used with MFCC based speaker recognition than GFCC based speaker recognition, since the GFCC features are more robust to noise conditions. Hence, the efficiency of the proposed method with GFCC based SR is limited (or sometimes negative) with high SNRs. However, the proposed method still shows promising results in low SNRs. To deal with this limitation, it easy to change the SNR threshold used to select between the clean reference model and adapted model to better fit the type of features used.
- Two types of dataset were used to evaluate the training on the fly. In general, the EER results from SALU-AC outperformed those results obtained from TMIT. The main

reason was that the duration of the signal in SALU-AC is longer than in TIMIT, which makes the estimation of noise profile more efficient in SALU-AC. Furthermore, the noisy signal in SALU-AC includes only the speech signal and contaminated noise, which makes the estimation easier than TIMIT which contained other adverse conditions.

- For noise profile estimation, simple techniques are used to detect the noise in the signal. This technique shows efficient performance when used with speaker recognition for different types of noise. However, this technique still shows some limitation in some types of noise (such as train interior noise). Therefore, it is recommended to improve this technique to get the best performance for training on the fly based systems. In the next chapter, different techniques are proposed based on extracting the noise from the signal. The results are compared with those obtained from recent techniques as well as those obtained from clean training.
- The estimated SNR used to check whether the signal is noisy or clean (threshold SNR) is not necessarily the same SNR used to create the reference adapted model; but it is still important to specify the value of SNR used to create the model, such that if the estimated SNR is almost high (between 20 and 10) then it better to use the same SNR to create the adapted model, while if the estimated SNR for input signal is low (between 10 and 0) then it is better to use a higher SNR value than estimated to obtain a better performance for Speaker recognition. However, the type of noise still plays the main role in specifying the best SNR.

8.3 Chapter Summary

In this Chapter, a new proposed approach to enhance the performance of speaker recognition in different noisy environments has been presented. This approach re-trains the enrolment model on the fly, depending on the SNR and Noise profile, which are estimated from submitted signals for recognition, to decrease the mismatch between enrolment models and the recognition signal. For noise profile detection, 1/3 octave band filter bank and time envelope are used to create an emulated noise close enough to the original. The EER and DET results of MFCC based speaker recognition show promising progress over clean training based SR for the four types of noise. On the other hand, when training on the fly is applied with GFCC based Speaker recognition the improvement is limited to low SNRs while the high SNRs show only slight progress.

The next chapter focuses on using different approaches to detect the noise in the signal. This technique is based on using a speech enhancement approach and spectral subtraction to extract noise from the signal.

CHAPTER 9
TRAINING ON THE FLY USING NOISE EXTRACTION TECHNIQUE

Chapter Overview

In the previous chapter, a new approach has been proposed to tackle the environmental noise issue in speaker recognition application. This approach, in contrast to other approaches, is based on creating an instant adaptive reference model that includes noise with the same characteristics included in the noisy recognition signal. For this purpose, a technique is proposed to estimate the profile of noise using 1/3 octave filter bank to create an emulated noise that is used to mix with reference signals for the claimed person. Furthermore, in Chapter 7, various speech cleaning algorithms were used to investigate their impact on performance on speaker recognition in noisy conditions, and the experimental results show these types of algorithms are unreliable for speaker recognition applications. In this chapter, a new proposed technique for extracting the noise from the input recognition signal is presented. This technique is based on employing speech cleaning algorithms to extract the noise from the signal contaminated with noise instead of directly cleaning the noisy signal, and then using this extracted noise to mix with reference signals and generate the adapted models. The next section describes the noise extraction technique. Section 9.2 includes the details of the experimental setup to evaluate this technique, with 'Training on the fly' being compared with the results in the previous chapter and conventional Speaker recognition. Finally, there is a summary and discussion in Section 9.3.

9.1 Noise Extraction Technique

This technique is based mainly on trying to split the noise signal from the speech signal contaminated with that noise and then using the extracted noise directly in the mixing procedure. This technique is based on the concept that the distorted noise signal represents a combination of the clean speech signal and additive noise source such that:

$$\boldsymbol{y(n) = x(n) + d(n)} \qquad \textbf{(9.1)}$$

where $\boldsymbol{x(n)}$ refers to clean speech signal, $\boldsymbol{d(n)}$ is the additive noise source, and $\boldsymbol{y(n)}$ is the corrupted noise signal. The essential goal of a speech enhancement algorithm is to improve the quality in the noisy speech signal $\boldsymbol{y(n)}$ by suppressing the background noise $\boldsymbol{d(n)}$ to obtain output signal $\boldsymbol{x'(n)}$ which is much closer to $\boldsymbol{x(n)}$. One of the major drawbacks of these types of approach is that the improvement in quality always comes at the expense of distortion of the speech signal, which in turn may affect speech intelligibility (Loizou, 2013). According to the previous experiments described in Chapter 7, these types of algorithms give very limited improvement in speaker recognition performance, and at times can even have an adverse effect on the performance. Furthermore, this improvement is limited to specific types of noise which cannot be generalized to all types of noise.

Now, instead of employing these algorithms to clean speech signal, these approaches are applied to extracting the noise from a noisy speech signal. In other words, suppose $\boldsymbol{x'(n)}$ is the output signal from applying speech cleaning algorithms, and $\boldsymbol{y(n)}$ represents the corrupted speech signal with additive noise. Then, the extracted noise signal $\boldsymbol{d'(n)}$ can easily be obtained by using a spectral subtraction of the cleaning signal $\boldsymbol{x'(n)}$ from the original signal $\boldsymbol{y(n)}$, that is:

$$\boldsymbol{d'(n) = y(n) - x'(n)} \qquad \textbf{(9.2)}$$

Taking the discrete-time Fourier transform DFT (Equation 6.3) of both sides gives:

$$\boldsymbol{D'(k) = Y(k) - X'(k)} \qquad \textbf{(9.3)}$$

where $\boldsymbol{D'(k)}$, $\boldsymbol{Y(k)}$, and $\boldsymbol{X'(k)}$ represent the magnitude spectrum of each extracted noise signal, original signal, and cleaning signal respectively. Figure 9.1 illustrates the block diagram of the extracted noise from signal technique.

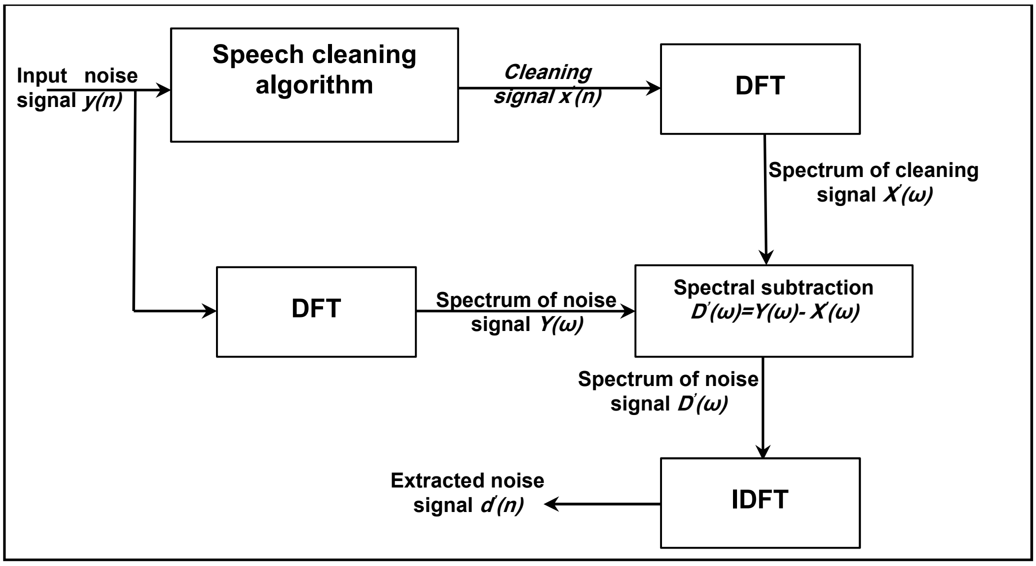

Input noise signal y(n)
Speech cleaning algorithm
Cleaning signal x(n)
DFT
Spectrum of cleaning signal X'(ω)
DFT
Spectrum of noise signal Y(ω)
Spectral subtraction D'(ω)=Y(ω)- X'(ω)
Spectrum of noise signal D'(ω)
IDFT
Extracted noise signal d'(n)

Figure 9.1: Block Diagram for noise Extraction Technique

In order to extract the noise $\boldsymbol{d'(n)}$ signal from a noisy speech signal $\boldsymbol{y(n)}$ the next procedure is followed:

1. noisy signal $\boldsymbol{y(n)}$ is passed through a speech enhancement algorithm (in this case the Wiener filter algorithm by Scalart and Filho (1996) is adopted, giving the cleaned signal $\boldsymbol{x'(n)}$ as seen in Figure 9.2.
2. For each $\boldsymbol{y(n)}$ and $\boldsymbol{x'(n)}$, Discrete Fourier Transform has been used to convert the signals from time domain to spectral domain to obtain $\boldsymbol{Y(k)}$ and $\boldsymbol{X'(k)}$ respectively.

$$\boldsymbol{Y(k) = DFT\big(y(n)\big) = \sum_{n=0}^{N-1} y(n)e^{-i\frac{2\pi kn}{N}}} \qquad \textbf{(9.4)}$$

$$\boldsymbol{X'(k) = DFT(x'(n)) = \sum_{n=0}^{N-1} x'(n)e^{-i\frac{2\pi kn}{N}}} \qquad \textbf{(9.5)}$$

where N is the number of samples in the signal

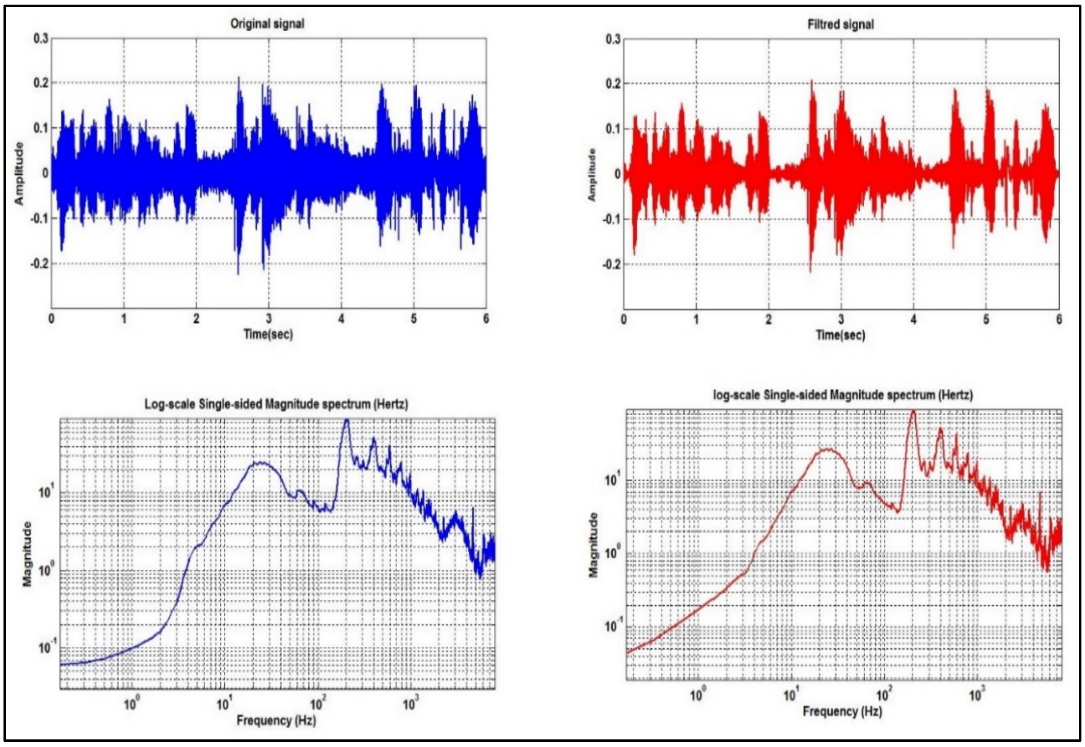


Figure 9.2: Original and filtered signals in Time and Frequency domains

3. Apply spectral subtraction to subtract the cleaned signal $\boldsymbol{X'(k)}$ from $\boldsymbol{Y(k)}$ using Equation 9.2 to obtain $\boldsymbol{D'(k)}$ which represents the spectrum of the noise signal.
4. Finally, the extracted signal $\boldsymbol{d'(n)}$ is obtained using Inverse Discrete Fourier transform IDFT for $\boldsymbol{D'(k)}$ such that:

$$\boldsymbol{d'(n) = IDFT\big(D'(k)\big) = \frac{1}{N}\sum_{k=0}^{N-1} D'(k)e^{+i\frac{2\pi kn}{N}}} \qquad \textbf{(9.6)}$$

Figure 9.3 shows the output signal in time and spectral domains.

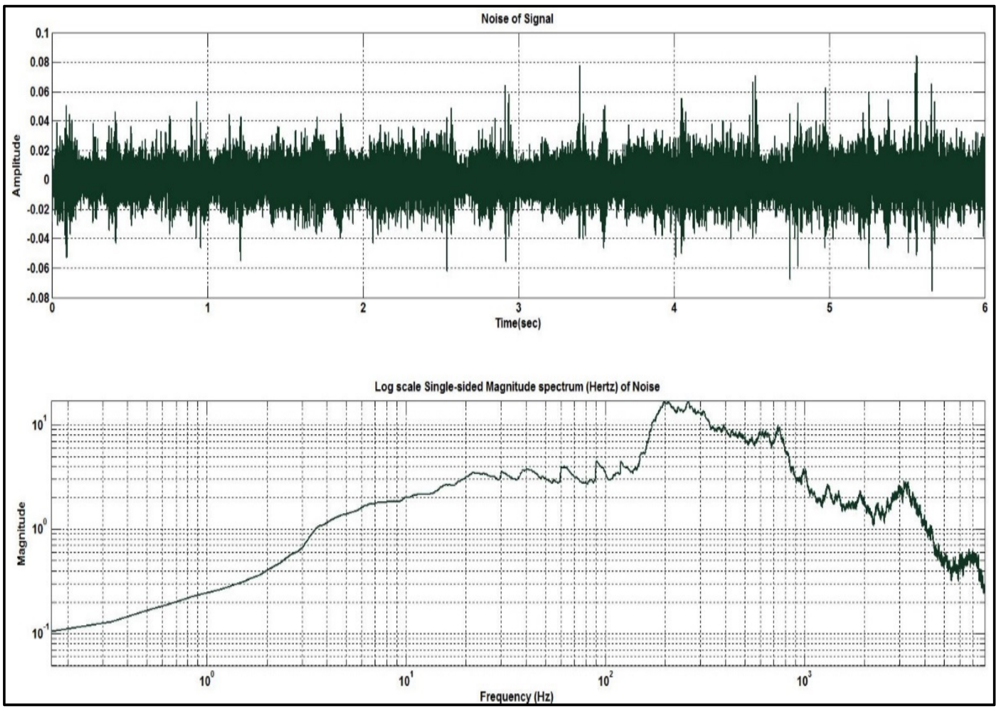


Figure 9.3: Output Noise Signal in Time and Frequency Domains

9.2 Training on the Fly using noise Extraction Technique

To investigate the training on the fly using noise extraction technique, the following experiments were made on speaker recognition to see what degree of progress could be achieved in the accuracy of speaker recognition. The results of these experiments are compared with those obtained in the previous chapter as well as those obtained from conventional speaker recognition. As in the previous chapter, two types of speech data have been adopted for this purpose TIMIT and SALU-AC. In addition, two kinds of features techniques are used, Mel Frequency Cepstrum Coefficients MFCC and Gammatone Cepstrum Coefficients GFCC. The Experiments are focusing on the most three challenging noise types (Cafeteria speech babble, interior moving train, and Street noise).

9.2.A Experimental Setup

The details of these experiments are the same as used in the previous chapter in order to see to what extent training on the fly using noise extraction can improve the performance of speaker recognition. The accuracy of the results is compared with the results from the previous chapter in addition to results from clean training based speaker recognition. The following are the details for these experiments: -

- MFCC including dynamic features (delta and delta delta) with 39-dimesional space and GFCC with 23- dimensional space were adopted.
- Gaussian Mixture model was used with 256 mixture. As mentioned before, the main reason for using GMM with training on the fly is the need to build the new reference model instead of adapting it from a universal model.
- To check the performance of training on the fly with noise extraction technique two datasets were adopted in this research. Each dataset included speech samples from 60 speakers (30 male, 30 female). 9 utterances were used in enrolment (training) phase for every speaker, all spoken in English. For recognition phase (training phase) one English utterance was used for each speaker. As before, this recognition set was divided into a number of subsets. Each subset represents the same speech samples mixed with different SNRs (20, 15, 10, 5, and 0 dB) for a specific type of noise to check the degradation in accuracy of speaker recognition.
- Three types of noise were adopted in these experiments: Cafeteria speech babble, interior train, and street noise, since these represent the most challenging types of noise. Furthermore, the results from the previous chapter show there is some limitation with speech sample contaminated with train interior noise.
- As before, error equal rate and detection error tradeoff were adopted to evaluate the results. The EER results degradation through different SNRs for aforementioned noises were compared with results from training on the fly using emulated noise in addition to the EER results from conventional speaker recognition.

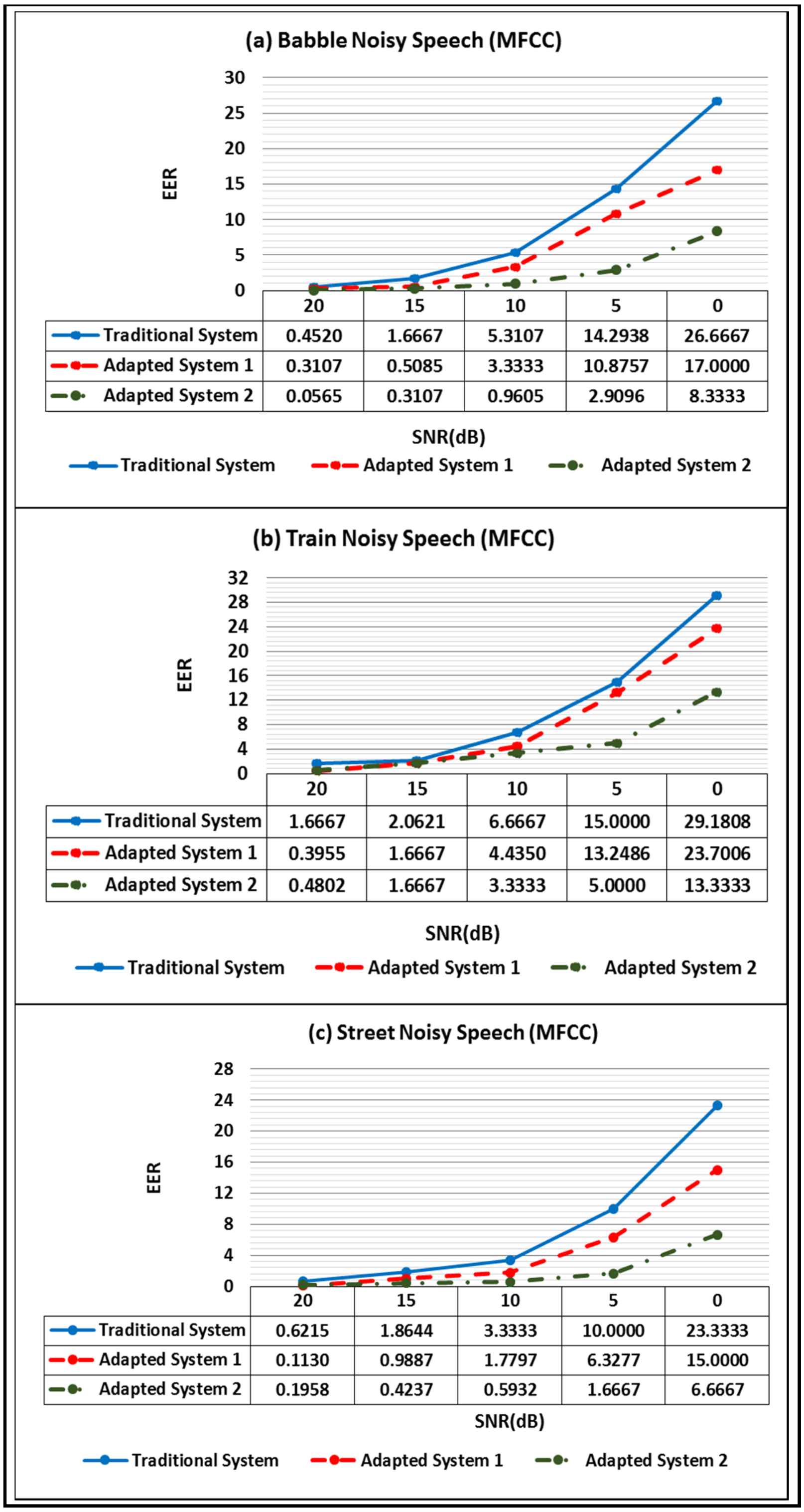
9.2.B Experimental Results

1. MFCC Baseline Results

In general, the EER results of Training on the fly using noise extraction technique (for simplicity this approach will be called TOTF-2 and Training on the fly using emulated noise is called TOTF-1) show significant improvement in the performance of speaker recognition under different SNRs for the specified types of noise when compared with results obtained from TOTF-1 and conventional speaker recognition (based on clean training signal). Figure 9.4 a to c demonstrate the degradation of performance (based on EER) through different SNRs for the three types of noise using the SALU-AC data set for MFCC based Speaker recognition. The dashed dot green line represents TOTF-2 and dashed red line represents TOTF-1, while the bold blue line represents conventional speaker recognition.

There is clearly a significant enhancement in speaker recognition accuracy when noise extraction techniques are applied with training on the fly, particularly in the range between 10 and 0 dB for the three types of noisy speech sets. For interior train (Figure 9.4 b), for instance, the EERs improve rapidly at 10, 05, and 0 dB from 4.4%, 13.2%, and 23.7% EERs in TOTF-1 to 3.3%, 5% ,13.3% ERRs in TOTF-2 respectively, which represents very substantial improvement for these speech samples contaminated with train interior noise. The other two types of noisy data (Cafeteria speech babble, and street noise) show the same improvement in EERs for most SNR values except at 20dB where both approaches (TOTF-1 and TOTF-2) have closely similar EER accuracy (Figure 9.4 a and Figure 9.4 c).

In DET Graphs, Figure 9.5 a to c show the performance of speaker recognition for the three cases (where the dashed-dot green curve represents TOTF-2 and dashed red curve represents TOTF-1, while the blue bold curve represents the conventional SR) for 5 dB SNR using SALU-AC and MFCC features based SR. It is clear that TOTF-2 outperforms TOTF-1 and conventional speaker recognition in both FPR and FNR for speech data contaminated with cafeteria speech babble, and street noise (Figure 9.5 a and Figure ***9.5*** b). Speech samples contaminated with interior train noise show the same FPR as conversational SR but still have the best FNR compared with the other two cases (Figure 9.5 c).



	20	15	10	5	0
Traditional System	0.4520	1.6667	5.3107	14.2938	26.6667
Adapted System 1	0.3107	0.5085	3.3333	10.8757	17.0000
Adapted System 2	0.0565	0.3107	0.9605	2.9096	8.3333

	20	15	10	5	0
Traditional System	1.6667	2.0621	6.6667	15.0000	29.1808
Adapted System 1	0.3955	1.6667	4.4350	13.2486	23.7006
Adapted System 2	0.4802	1.6667	3.3333	5.0000	13.3333

	20	15	10	5	0
Traditional System	0.6215	1.8644	3.3333	10.0000	23.3333
Adapted System 1	0.1130	0.9887	1.7797	6.3277	15.0000
Adapted System 2	0.1958	0.4237	0.5932	1.6667	6.6667

Figure 9.4: EER of two proposed approaches and clean training with MFCC based SR using SALU-AC

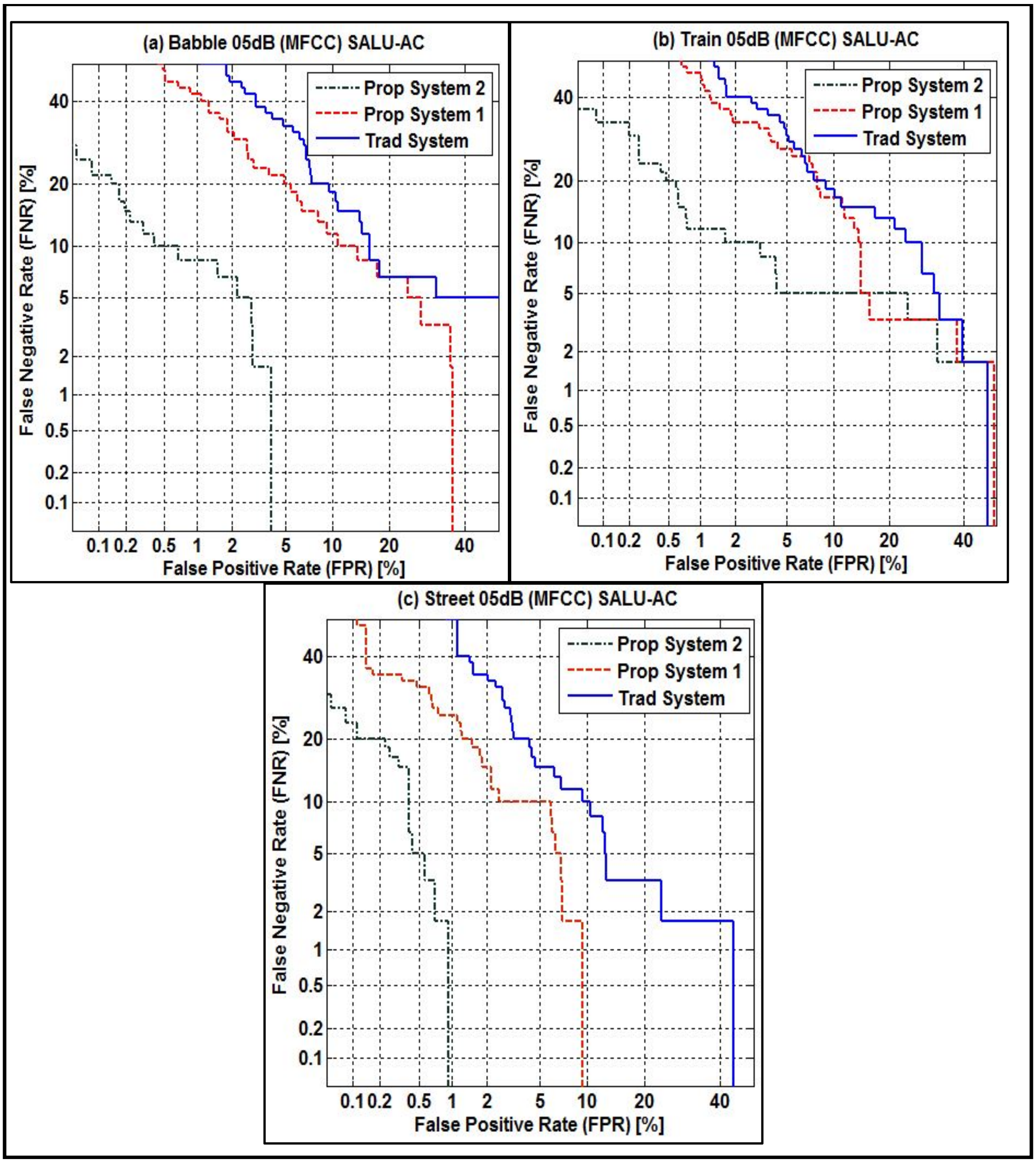
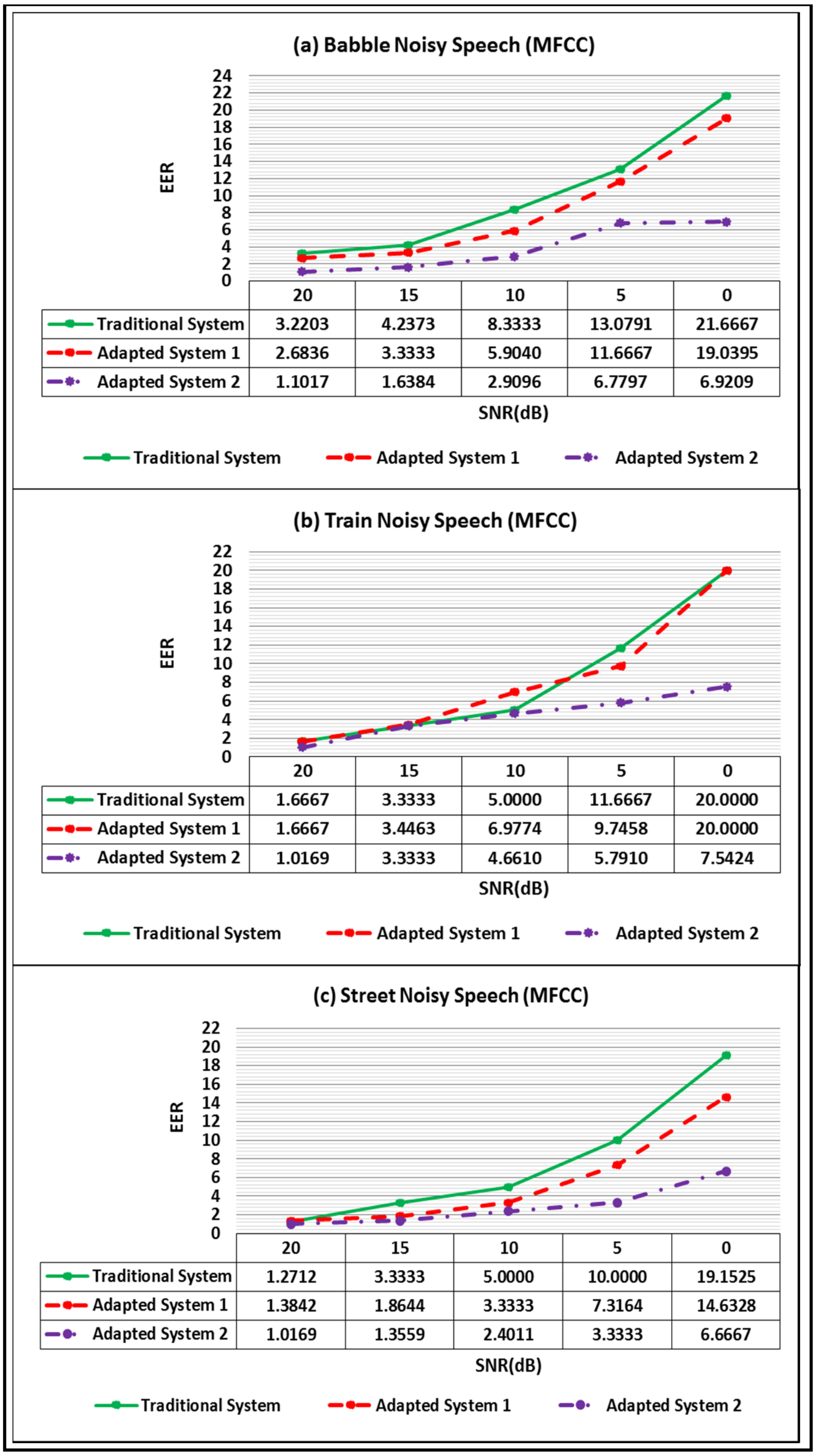


Figure 9.5: DET graphs for two proposed approaches based SR and Conventional SR in 05 dB SNR using SALU-AC (MFCC based)

When the TIMIT datasets were applied (Figure 9.6 a to c), the training on the fly based noise extraction technique (the dashed dot pink line) still showed the same improvement in EERs for all SNRs and for the three types of noisy speech. As in the SALU-AC dataset, this enhancement in accuracy became very high at 10, 5, and 0 dB, particularly for the speech data contaminated with interior train noise (Figure 9.6 b) which showed significant improvement if compared with EERs from TOTF-1 described in the last chapter, which showed limited improvement for this type of noise. Furthermore, the improvement in EERs for low SNRs (5 and 0 dB) show high improvement in values of EERs when compared with the EERs values from TOTF-1 and conventional SR.

In the Detection error trade-off (DET) graphs Figure 9.7 a to c represent the performance of the three cases at 10dB SNR for the three types of noise (the dashed dot pink curve represents the TOTF-2, the dashed red curve represents the TOTF-1, while the bold green curve represents the conventional speaker recognition). As seen in these figures, SR based on TOTF-2 outperformed TOTF-1 and conversational SR in both FPR and FNR for speech samples contaminated with speech babble and street noise. Speech samples contaminated with interior train noise showed the same performance in FPR for adapted approaches (TOTF-1 and TOTF-2) but still outperformed the conventional SR (Figure 9.7 b).

The main reason that noise extraction based training on the fly outperformed noise emulated based training on the fly is that the noise used in the first techniques to build the adapted reference model is similar to the noise embedded in the input recognition signal, since it is extracted directly from the signal; while the second type of technique is based mainly on the accuracy of estimations of the profile noise to create an emulated noise close to the original.



(a) Babble Noisy Speech (MFCC)

	20	15	10	5	0
Traditional System	3.2203	4.2373	8.3333	13.0791	21.6667
Adapted System 1	2.6836	3.3333	5.9040	11.6667	19.0395
Adapted System 2	1.1017	1.6384	2.9096	6.7797	6.9209

(b) Train Noisy Speech (MFCC)

	20	15	10	5	0
Traditional System	1.6667	3.3333	5.0000	11.6667	20.0000
Adapted System 1	1.6667	3.4463	6.9774	9.7458	20.0000
Adapted System 2	1.0169	3.3333	4.6610	5.7910	7.5424

(c) Street Noisy Speech (MFCC)

	20	15	10	5	0
Traditional System	1.2712	3.3333	5.0000	10.0000	19.1525
Adapted System 1	1.3842	1.8644	3.3333	7.3164	14.6328
Adapted System 2	1.0169	1.3559	2.4011	3.3333	6.6667

Figure 9.6: EER of two proposed approaches and clean training with MFCC based SR using TIMIT

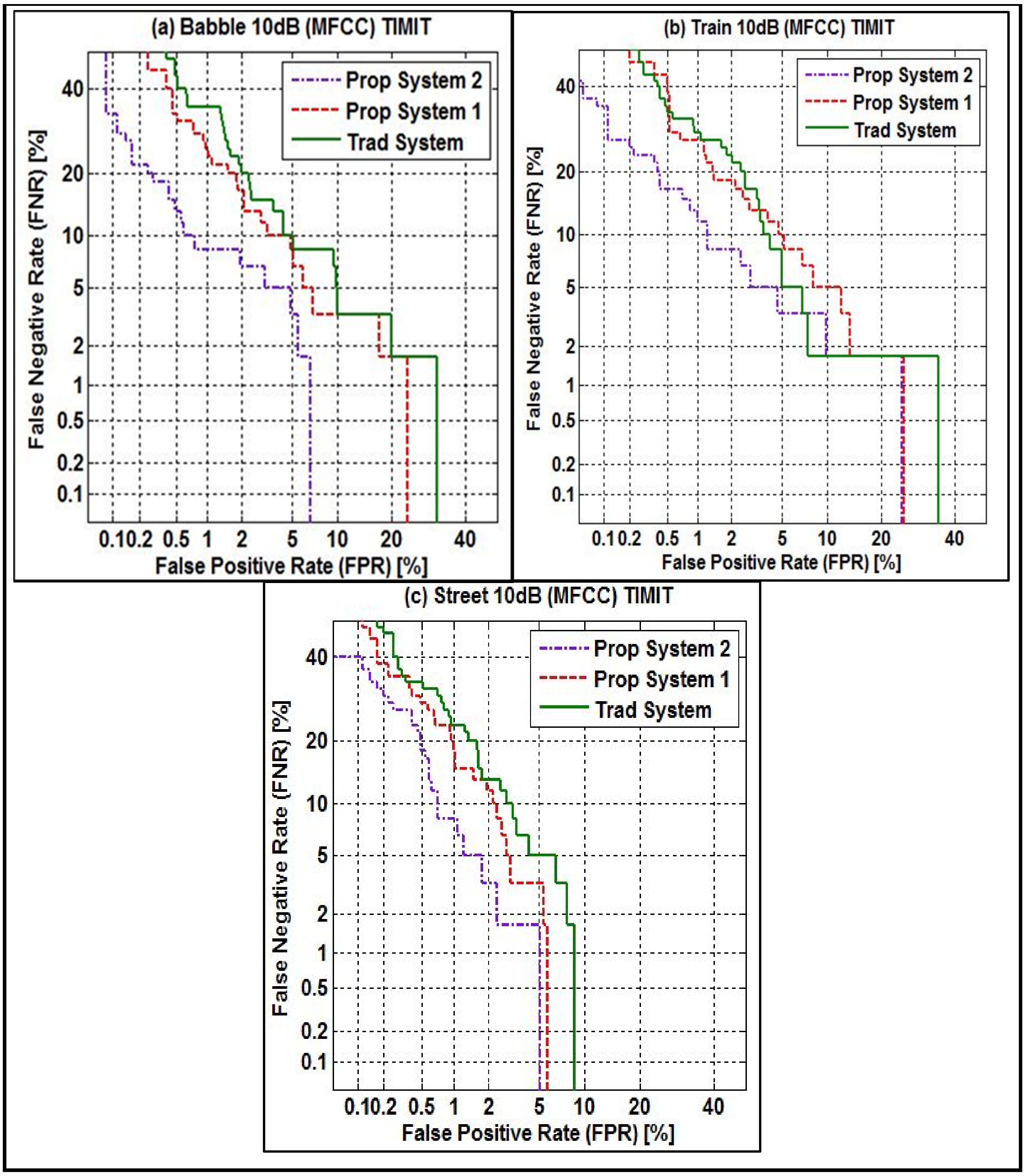


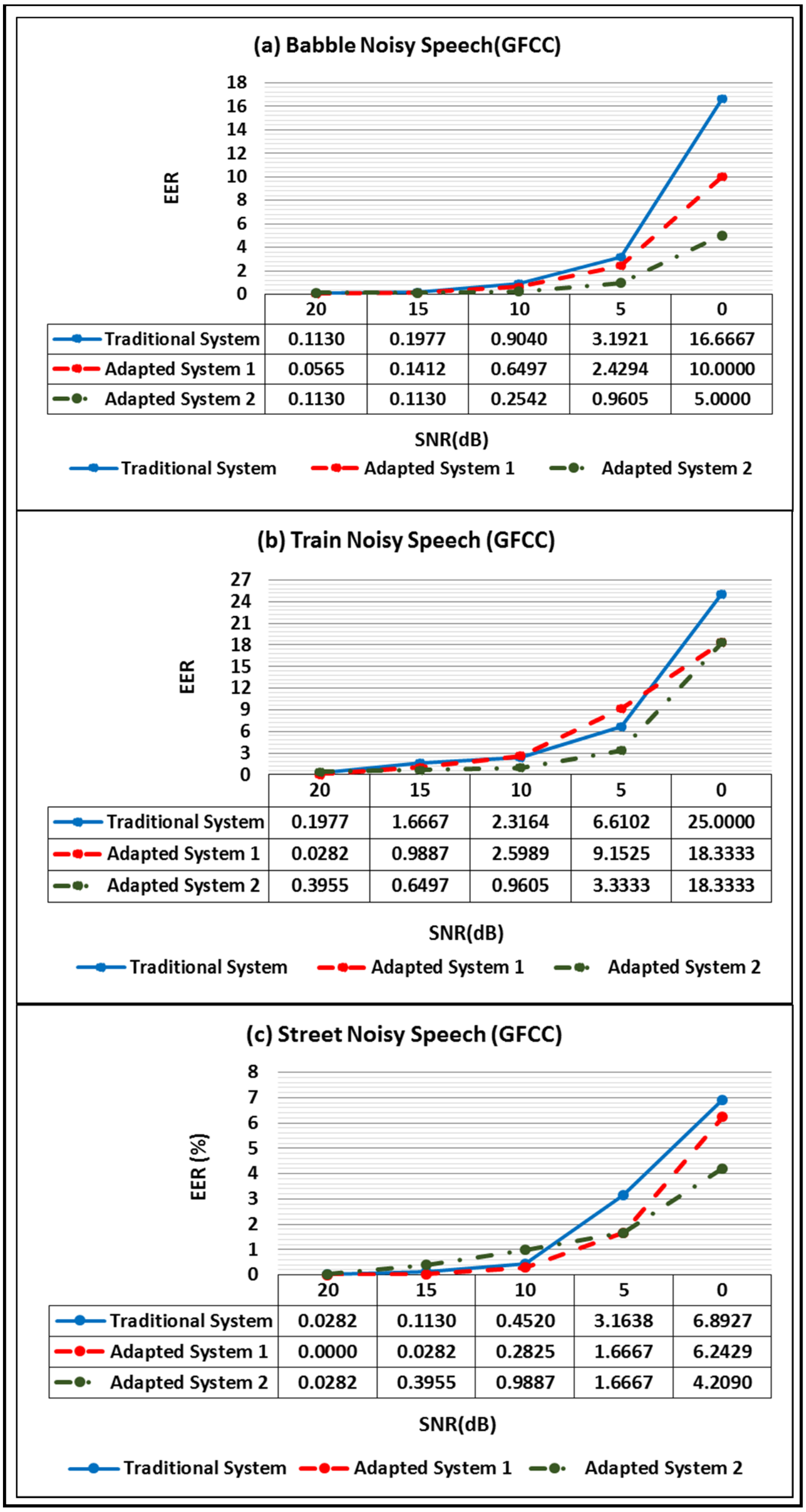
Figure 9.7: DET graph for two proposed approaches based SR and Conventional SR in 10 dB SNR using TIMIT (MFCC based)

2. GFCC Baseline Results

When Training on the fly using noise extraction technique is applied in GFCC based speaker recognition using SALU-AC datasets, the EER results, as in the previous approach, show very limited improvement (even in some cases a slight degradation) over conventional speaker recognition and TOTF-1 for 20, 15, and 10 dB SNRs. On the other hand, the low SNRs (5 and 0 dB) show significant enhancement in EER values over TOTF-1 and conventional SR, as seen in Figure 9.8 a to c, where the dashed dot green line represents the training on the fly based noise extraction technique (TOTF-2), dashed red represents training on the fly based on estimated emulated noise (TOTF-1), and the bold blue line represents the conventional speaker recognition.

For example, the three cases have close EERs at 20 dB SNR for the three types of noise. Conversely, the TOTF-2 shows good improvement at 5, and 0 dB SNR for speech samples contaminated with babble noise, and interior train noise, especially in the case of interior train noise where the EERs decreased from 6.6 % in conversational SR to 3.3 in TOTF-2 at 5 dB SNR, while the TOTF-1 failed to achieve any improvement in this SNR (Figure 9.8 b).

The DET graphs (Figure 9.9 a to c) which represent the performance of the three cases at 10 dB SNR for the three types of noisy speech data, show the improvement in accuracy of TOTF-2 (the dashed dot shade green curve) over the other two cases (the red dashed curve represents TOTF-1, and the bold blue line represents the conventional SR) for speech samples contaminated with cafeteria babble and interior train in both FPR and FNR (Figure 9.9 a, b). Conversely, the DET graph of speech data contaminated with street noise shows a decrease in FPR for TOTF-2, while the best performance is for TOTF-1 (Figure 9.9 c).



	20	15	10	5	0
Traditional System	0.1130	0.1977	0.9040	3.1921	16.6667
Adapted System 1	0.0565	0.1412	0.6497	2.4294	10.0000
Adapted System 2	0.1130	0.1130	0.2542	0.9605	5.0000

	20	15	10	5	0
Traditional System	0.1977	1.6667	2.3164	6.6102	25.0000
Adapted System 1	0.0282	0.9887	2.5989	9.1525	18.3333
Adapted System 2	0.3955	0.6497	0.9605	3.3333	18.3333

	20	15	10	5	0
Traditional System	0.0282	0.1130	0.4520	3.1638	6.8927
Adapted System 1	0.0000	0.0282	0.2825	1.6667	6.2429
Adapted System 2	0.0282	0.3955	0.9887	1.6667	4.2090

Figure 9.8: EER of two proposed approaches and clean training with GFCC based SR using SALU-AC

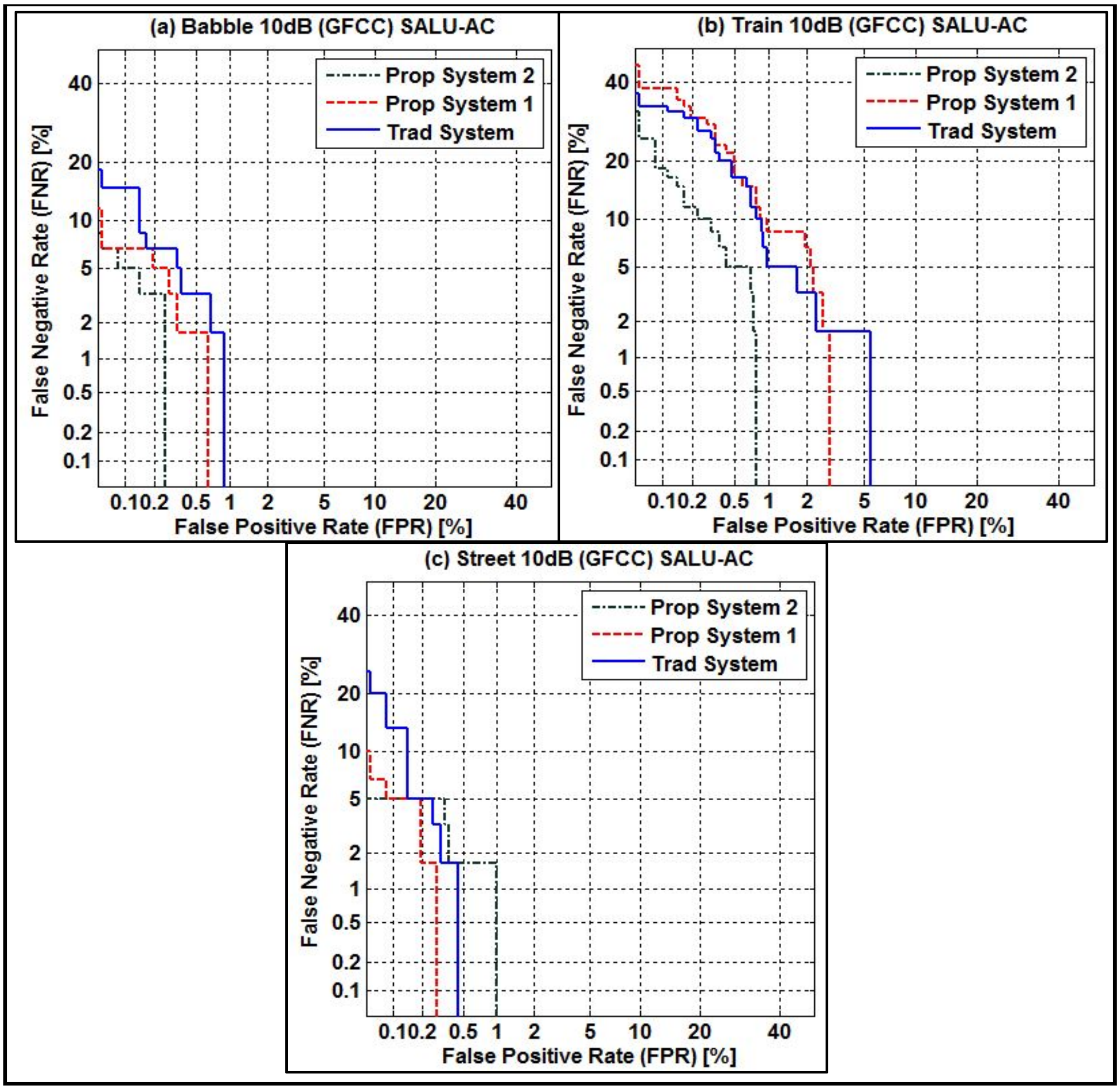
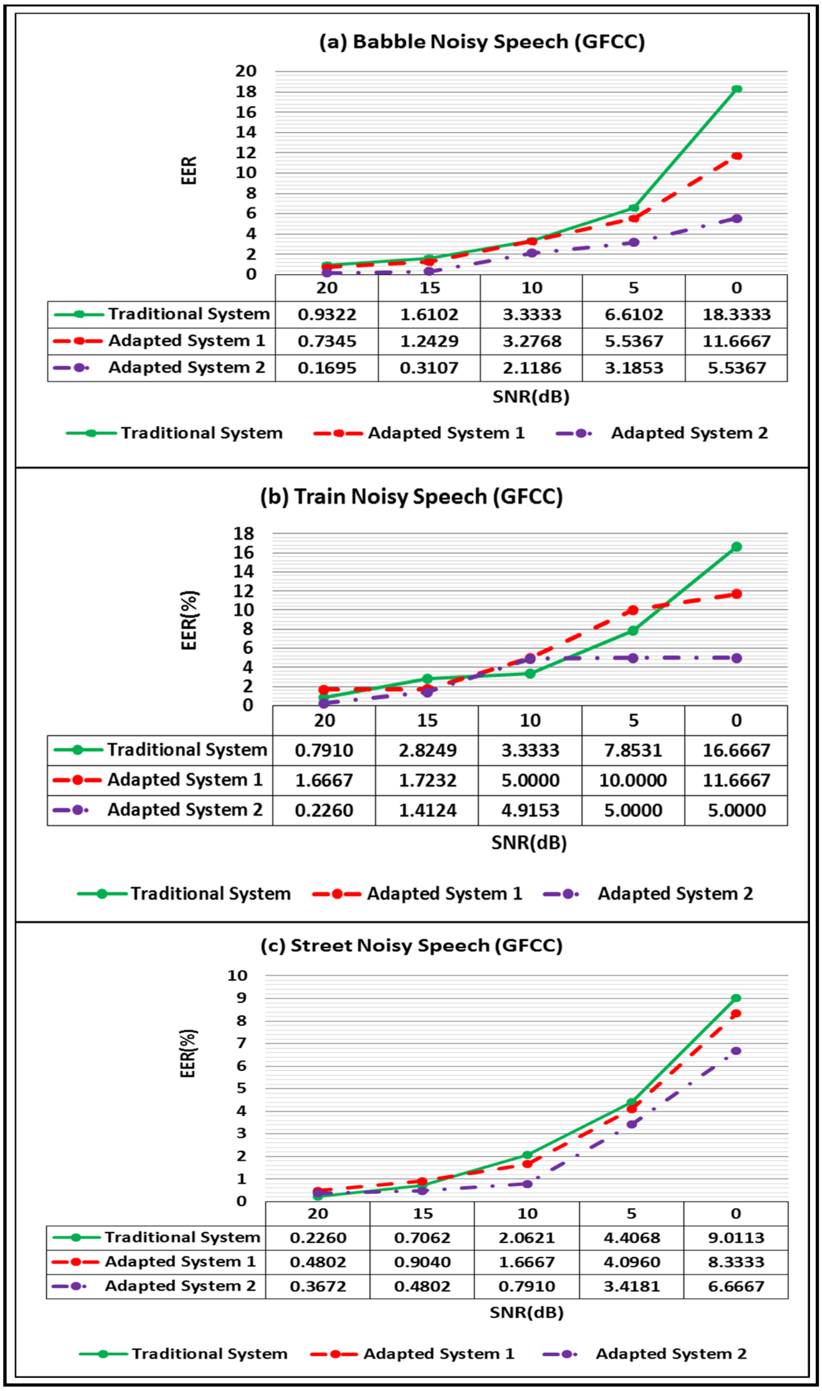


Figure 9.9: DET graphs for two proposed approaches based and Conventional SR in 10 dB SNR using SALU-AC (GFCC based)

As with the SALU-AC, when the TIMIT datasets are used (Figure 9.10 a to c), the EERs for TOTF-2 (the dashed dot pink line) show significant improvement over the other two cases (the dashed red for TOTF-1 and bold green for conventional SR) at 5 and 0 dB SNR, while the EERs at 20, 15 and 10 vary from one type of noise to another.

The DET graphs (Figure 9.11 a to c) show the performance of three cases using the TIMIT datasets with GFCC-based speaker recognition at 5 dB SNR. The TOTF-2 (the dashed dot pink curve) outperforms the adapted approach-1(the dashed red curve) and conventional SR (the bold green curve) for cafeteria babble, and train interior noisy data in both FPR and FNR (Figure 9.11 a, c) . However, TOTF-2 shows the same FPR for conventional SR when speech samples are contaminated with street noise.



(a) Babble Noisy Speech (GFCC)

	20	15	10	5	0
Traditional System	0.9322	1.6102	3.3333	6.6102	18.3333
Adapted System 1	0.7345	1.2429	3.2768	5.5367	11.6667
Adapted System 2	0.1695	0.3107	2.1186	3.1853	5.5367

(b) Train Noisy Speech (GFCC)

	20	15	10	5	0
Traditional System	0.7910	2.8249	3.3333	7.8531	16.6667
Adapted System 1	1.6667	1.7232	5.0000	10.0000	11.6667
Adapted System 2	0.2260	1.4124	4.9153	5.0000	5.0000

(c) Street Noisy Speech (GFCC)

	20	15	10	5	0
Traditional System	0.2260	0.7062	2.0621	4.4068	9.0113
Adapted System 1	0.4802	0.9040	1.6667	4.0960	8.3333
Adapted System 2	0.3672	0.4802	0.7910	3.4181	6.6667

Figure 9.10: EER of two proposed approaches and clean training with GFCC based SR using TIMIT

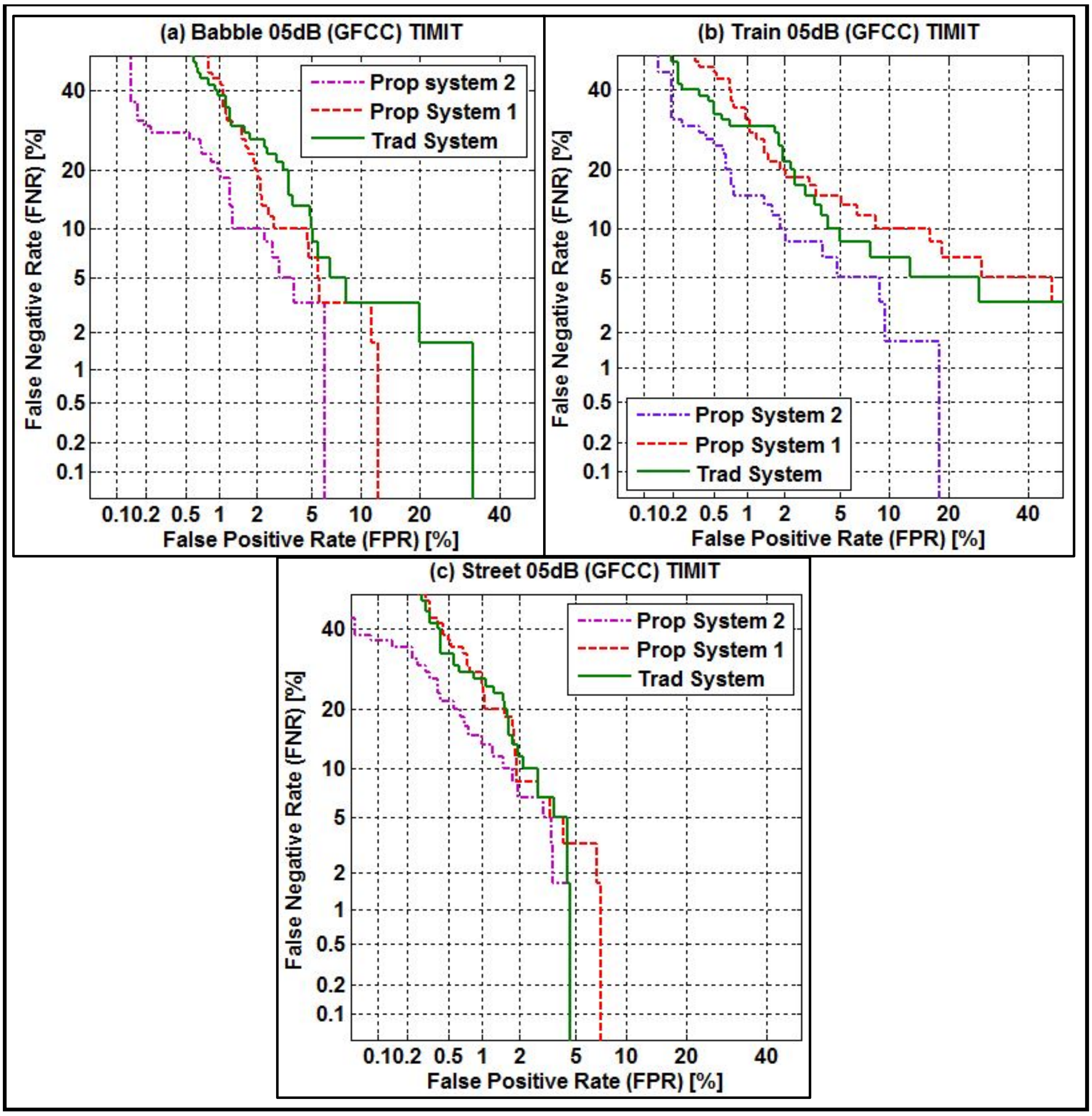


Figure 9.11: DET graph for two proposed approaches based and Conventional SR in 05 dB SNR using TIMIT (GFCC based)

In summary, according to the results from previous experiments, the findings are as follows:

- Training on the fly with noise extraction techniques described at the beginning of this chapter is more efficient with speaker recognition than method based on estimated emulated noise, since the noise used to create the adapted reference model is the same noise that contaminates the input recognition signal, which makes the mismatch between the enrolment and recognition phases smaller. However, the extracted noise signal still contains a low level of speech from the original signal, but this level of speech is too low to affect the mixing signal used to build the reference model. Furthermore, this remaining speech, when mixed with the reference signal for the claimed speaker, will increase the probability that the input signal belongs to the claimed person if the recognition signal belongs to the target person. Otherwise, if the recognition signal belongs to an imposter, then the remaining speech will help to increase the mismatch between the created model and input signal.

- Similar to the finding from the previous chapter, training on the fly based noise extraction was shown to be more efficient with MFCC- baseline speaker recognition than GFCC- baseline. However, the results still show significant improvement for training on fly with GFCC based Speaker recognition at low SNRs.
- As described in the last chapter, the value of SNR used to create an adapted noisy signal depends mainly on estimated SNR from input signal; but not necessarily the same value, because this SNR varies from one type of noise to another and also the level of estimated SNR.

Table 9-1 and Table 9-2 show the summary of EERs for training on fly approaches using the two types of dataset and two types of features extraction.

Table 9-1: EER for MFCC and GFCC based Speaker recognition using SALU-AC dataset

Noise type	SNR	Est. SNR (dB)	Train SNR (dB)	MFCC TOTF-2	MFCC TOTF-1	GFCC TOTF-2	GFCC TOTF-2
Cafeteria Babble	20dB	21.5dB	21d.B	0.05	0.31	0.11	0.05
	15dB	17.3	17	0.3	0.5	0.11	0.14
	10dB	12.5	15	0.96	3.33	0.25	0.64
	05dB	7.2	10	2.9	10.87	0.96	2.42
	0 dB	4.1	5	8.3	17	5	10
Interior moving Train	20dB	19.3	19	0.48	0.39	0.39	0.02
	15dB	16.1	16	1.66	1.66	0.64	0.98
	10dB	11.4	15	3.33	4.43	0.96	2.59
	05dB	7.6	10	5	13.24	3.33	9.15
	0 dB	4.2	5	13	23.7	18.33	18.33
Street	20dB	19.3	19	0.19	0.11	0.028	0
	15dB	15.4	15	0.42	0.98	0.39	0.02
	10dB	11.3	15	0.59	1.77	0.98	0.28
	05dB	6.7	10	1.66	6.32	1.66	1.66
	0 dB	3.5	5	6.66	15	4.2	6.24

Table 9-2: EER for MFCC and GFCC based Speaker recognition using TIMIT dataset

Noise type	SNR	Est. SNR	Train SNR	MFCC TOTF-2	MFCC TOTF-1	GFCC TOTF-2	GFCC TOTF- 2
Cafeteria Babble	20dB	19.7	20	1.1	2.68	0.16	0.73
	15dB	15.4	15	1.63	3.33	0.31	1.24
	10dB	11	15	2.9	5.9	2.11	3.27
	05dB	6.6	10	6.77	11.66	3.18	5.53
	0 dB	3.8	5	6.92	19.03	5.53	11.66
Interior moving Train	20dB	19.7	20	1.01	1.66	0.22	1.66
	15dB	15.3	15	3.33	3.44	1.41	1.72
	10dB	10.5	15	4.66	6.97	4.91	5
	05dB	5.6	10	5.79	9.74	5	10
	0 dB	3.7	5	7.54	20	5	11.66
Street	20dB	19.7	20	1.01	1.38	0.36	0.48
	15dB	15.1	15	1.35	1.86	0.48	0.9
	10dB	11.6	15	2.4	3.33	0.79	1.66
	05dB	7.2	10	3.33	7.31	3.41	4.09
	0 dB	2.2	5	6.66	14.63	6.66	8.33

9.3 Chapter Summary

In this Chapter, Training on the fly using noise extraction techniques to improve the noise robustness of speaker recognition has been presented. This technique includes using one of the speech enhancement algorithms (in this case Wiener filter was adopted) as well as a spectral subtraction algorithm to extract the noise from the input speech signal and then using this noise to mix with the reference signals to build the adapted reference models. The EERs and DET results from this technique have been compared with those obtained in the previous chapter in addition to results from conventional speaker recognition (based on clean training). The results of MFCC- baseline speaker recognition show the proposed techniques increase the efficiency of training on the fly over those based on estimated emulated noise since this technique deals with noise directly instead of estimating it from input noisy signal.

CHAPTER 10
DISCUSSION,CONCLUSIONS
AND RECOMMENDATION FOR FUTURE WORKS

The primary goals of this research study can be categorised into two aspects: the general aspect, on the one hand, concerned with adopting different approaches to investigate their effectiveness in improving the robustness of speaker recognition in the presence of different environmental noise sources. In addition, the impact of language mismatch between the enrolment and recognition phases on the accuracy of speaker recognition was studied under the same noisy conditions. This was done by using a new database (called SALU-AC) containing audio speech samples recorded in different languages in addition to English. These samples were recorded without being limited to particular text messages.

The specific aspect, on the other hand, focused on developing a new approach to improve robust text-independent speaker recognition systems for real world application under various environmental noises, especially when the information provided about the noise is limited or absent. This approach is based on instantly re-training the enrolment models, depending on Signal to noise ratio and the identity of noise, which are estimated from signals submitted to recognition, to decrease the mismatch between the newly created enrolment models and the recognition signal. The summary and conclusions of this research, in addition to some suggestions for future work, are presented in sections 10.1 and 10.2 respectively.

10.1 Discussion

Automatic Speaker Recognition has recently become one of the most common aspects of biometric authentication fields, due to the simplicity of capturing the speech samples from speakers compared with other types of biometrics. Furthermore, this type of biometric does not require special devices to be implemented, such as in thumbprint and iris authentications. As a result, this biometric has recently been receiving a great deal of attention from many researchers. One of the major challenges of speaker recognition is that associated with its performance against background environmental noise, since these types of applications perform poorly under noisy conditions.

This study deals with noise challenges to improve the speaker recognition under these adverse conditions. For this purpose, different approaches have been adopted. These approaches, as mentioned before, can be categorised into general and specific aspects.

The general aspect, which concerns different approaches to deal with noise issues, can be categorised into three scenarios:

1. To handle the degradation in speaker recognition, an investigation has been made on enrolment phase in speaker recognition. This investigation involved training the speaker recognition with various types of noisy speech and with controlled SNRs to create noisy reference models. The results of using noisy utterances in the enrolment phase showed significant improvement in the accuracy of speaker recognition if compared with clean speech enrolment. However, it was important to specify the SNR of the noisy speech used in the enrolment phase to ensure good speaker recognition performance in noisy conditions. Furthermore, the results show that the identity of the noise plays a major role in providing a suitable noisy reference model because each type of noise has a different power spectrum which represents a unique identity for each type of noise. As a result, to create suitable noisy reference model two factors should be taken into account: SNR level, and noise identity with the same characteristics as included in the noise that contaminated the recognition signal.
2. The robustness of different features extraction algorithms in the performance of speaker recognition was investigated. Conventional MFCC, including dynamic features, and GFCC were adopted in this study to investigate their robustness to different kinds of noise, in addition to their sensitivity to a variety of languages of recognition samples sets. The results of these experiments showed that GFCC baseline models have better performance than conventional MFCC baseline under different additive noise conditions. Conversely, changing the languages of the recognition speech set adversely affected the performance of the GFCC baseline, while the MFCC baseline showed more robustness for this change. The reason for this sensitivity of GFCC features for language mismatch is probably the design of the Gammatone filter-bank, representing an emulation of the human auditory system, which makes it more sensitive to changing the language used in enrolment and recognition phases. For this point, two options are suggested, first undertake further investigation on

the Gammtone filter-bank in order to improve its robustness to mismatched language. Second, investigate the performance of using logarithmic operation instead of cubic root to extract GFCC features.

3. The impact of speech enhancement algorithms as a pre-processing stage on the performance of speaker recognition was investigated under various noisy conditions. The cleaning algorithms adopted in this study involved a spectral subtraction algorithm, statistical based algorithms, and a subspace algorithm. First, the effect of Wiener filtering algorithms was considered when applied as a pre-processing stage for test-phase and training-test phases together on the performance of speaker recognition under additive noise conditions. Then the impact of the other cleaning algorithms was examined when they were applied as a pre-processing stage for both training and test phases on the performance of the verification system in the same environments. The results obtained from these experiments indicate that these approaches caused degradation in accuracy of speaker recognition when used with the recognition (test) phase only. On the other hand, these noise cleaning methods showed very limited improvement and sometimes had an adverse impact on speaker recognition performance. Therefore, these kinds of approaches are unreliable for robustness of speaker recognition in different noisy conditions. The main reason that these types of algorithms are unreliable with speaker recognition is that most of them focus on using the same filtering techniques to tackle broadband noise so that when the noise is removed from the speech signal this causes alterations in speech components. As a result, this change will affect the frequency range of speech after cleaning. This change in frequency components will affect the features that are extracted from this filtering signal. It is therefore recommended to focus on using speech enhancement algorithms that deal with enhancing both quality and intelligibility of the speech signal.

For the specific aspect, and based on findings in noisy training in chapter 5, a ‘Training on the fly’ approach (Al-Noori et al., 2017) has been proposed to improve the performance of speaker recognition in the presence of environmental noise. This approach is based on estimating the SNR and finding or estimating the noise in the input recognition signal. For SNR estimation, a time domain envelope has been employed. For finding the identity of noise in the signal, two techniques are proposed: the first is based on estimating the power spectrum on the noise using 1/3 octave band filter bank and then creating an emulated noise much closer to the original using white noise. The second technique is based on attempting to extract the noise by filtering the noisy signal using one of the speech enhancement approaches adopted in Chapter 7 (in this research Wiener filters were adopted) and then using spectral subtraction to subtract the filtered signal from the original to obtain the noise used later to generate the noisy reference model. From experiments on the two types of features adopted in this research with GMM-based Speaker recognition, and by using the two types of speech database to evaluate the performance of training on the fly approach, the following findings can be drawn:

1. The 'Training on the fly' approach based on the two detection noise techniques shows significant improvement in performance of MFCC-baseline when compared with the performance of conventional speaker recognition (based on a clean reference model). For training on the fly based on emulated noise, the results show significant improvement for speech samples contaminated with the interior car and white noise and promising improvement for speech samples contaminated with street and cafeteria babble noise. However, Training on the fly shows limitation with speech samples contaminated with interior train noise, because this type of noise has different sources of noise for which the estimation of the power spectrum is difficult.
2. Training on the fly based on noise extraction technique shows significant improvement over the one based on emulated noise and conventional speaker recognition in MFCC-GMM baseline speaker recognition for all types of the noisy speech sample. The main reason that training on the fly based noise extraction outperforms the one based on estimated noise is that the characteristics of extracted noise are the same as those included in the original signal, which makes the reference model created from mixing this noise with the reference speech signal more accurate than one created from the estimated noise. However, the extracted noise may still contain a ratio of speech from the original signal, but this ratio is too small to affect the created model. In addition, the remaining speech when mixed with speech signal belongs to the authorised person, making the model more robust.
3. The evaluation results show that the training on the fly approach is more efficient with MFCC-baseline than GFCC-baseline. The improvement of GFCC-baseline is limited in low SNR. This leads us to conclude that training on the fly with GFCC-baseline is not efficient when SNR is high (i.e. between 20 and 10 dB).

4. In order to get the best performance for speaker recognition using training on the fly, it is necessary to specify the SNR for the SNR threshold and for speech-noise mixing, since the impact of the noise on the accuracy of speaker recognition varies from one type to another (as seen in Chapter 4). Therefore, it is necessary to calibrate the signal to noise ratio with real world noise to give the best performance accuracy for speaker recognition.
5. The Training on the fly approach may cause a delay in recognition process since it requires extra time to create the adapted model (on the fly model). This delay varies depending on the type of speaker recognition system applied for. In real time speaker verification systems this delay is not a serious issue since the adapted models will be created only for the claimed person. However, in the case of speaker identification this delay becomes very high because, in this case, all enrolment models will be re-trained to find the closest model. For that reason, training on the fly is limited to a small number of speakers if used with speaker identification.
6. In this thesis, GMM based speaker recognition is used to evaluate the training on the fly approach, but that does not mean it cannot be applied with universal background model (UBM). In this case, the UBM model can be used if the input signal is clean (i.e. when the recognition signal is identified as clean). Otherwise, the UBM is re-trained using the signals of the claimed person only and then this model is adapted as a speaker model (in this case the speaker model is adapt from itself instead of independent models). The aim of using the UBM adaptation, in this case, is to build a robust supervector model. This step will be reserved for future research.

10.2 Conclusions

From the previous discussion section the conclusions can be summarised as follows:

1. The impacts of different types of noise on the performance of Speaker recognition in the context of accuracy are varied from one type of noise to another. For instance, the effects of Train interior and babble noise on the SR accuracy are very high compared with the effects of car interior.
2. To provide suitable adapted noisy models that closely resemble the noisy input signal, two factors should be taken into account: first, signal to noise ratio (SNR), which represents the main factor to specify whether the input signal is noisy or clean. In addition, this factor

specifies the level of noise in a noisy signal through the mixing process. Second, the profile of the noise, which represents the unique identity of the noise in noisy signal.

3. The value of estimated SNR used to specify whether the recognition signal is noisy or clean (threshold SNR) is not necessarily the same SNR value that is used to create the reference adapted model (on the fly model) ; but it is still important to specify the value of SNR used in the mixture process. However, the profile of the noise still plays the main role in specifying the best SNR for the mixture process.
4. Although the results show that GFCC features are more robust than conventional MFCC in noisy environments, mismatched language between the elemental and recognition phase affects the performance of GFCC based SR features more than one based on MFCC features.
5. Although the results show that using speech enhancement algorithms is unreliable for robustness of speaker recognition, these algorithms showed efficiency when used to extract the noise for the input signal to create noisy adapted models for the Training on the fly approach.
6. Since the GFCC features are more robust to noisy environments, the improvement for training on the fly with GFCC based SR is limited to low SNRs.
7. Since the noise used to create the adapted reference model is the same noise that contaminated the input recognition signal, the performance of training on the fly based on noise extraction outperforms the one based on estimated emulated noise.

10.3 Recommendations for Future Work

This section briefly gives some suggestions for future work which could be used to extend the work presented in this research:

1. Examine the performance of using training on the fly approach with different state of the art modelling techniques for improving the accuracy of speaker recognition under noisy conditions, for example, i-vector baseline. In addition, further investigation of this approach with universal background model as mentioned before.
2. Improve the noise profile estimation to obtain the best estimation for noise contaminated in recognition signal, which should lead to improved performance of training on the fly.

3. The experiments in this research were conducted on artificial noisy speech samples which were already prepared for this purpose. For future work, speaker recognition using training on the fly should be applied to realistic noisy speech data collected directly from noisy environments (such as recorded speech samples in a cafeteria crowded with many people).
4. More investigation will be undertaken on the GFCC to improve their performance when the languages of test speech samples are different from enrolment speech samples. In this case, it may be possible to deal with speech samples in each language separately from other languages, to find the greatest impact of language mismatch on GFCC-baseline. This can be done using various speech samples recorded in different languages in the SALU-AC database.
5. Most of the speech enhancement algorithms used in this research are focused on improving speech quality but failed in improving speech intelligibility. Future work should focus on using approaches that deal with enhancing both quality and intelligibility of speech signals. For these purposes, Singular Spectral Analysis (SSA) will be employed.

APPENDICES

Appendix I
DET Graphs for Chapter 5

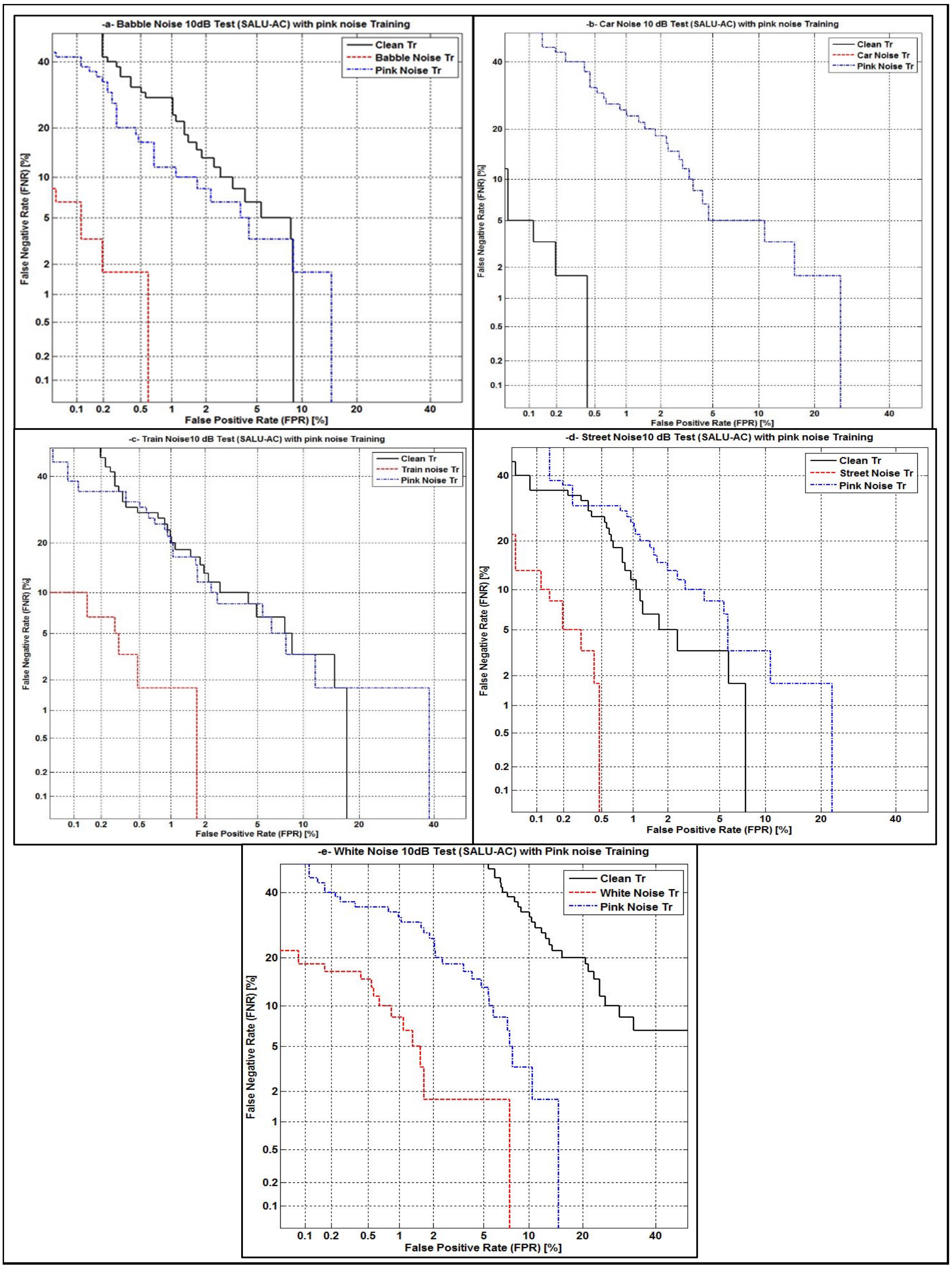


Figure I.1:DET graphs of noisy training for 10 dB SNR test samples

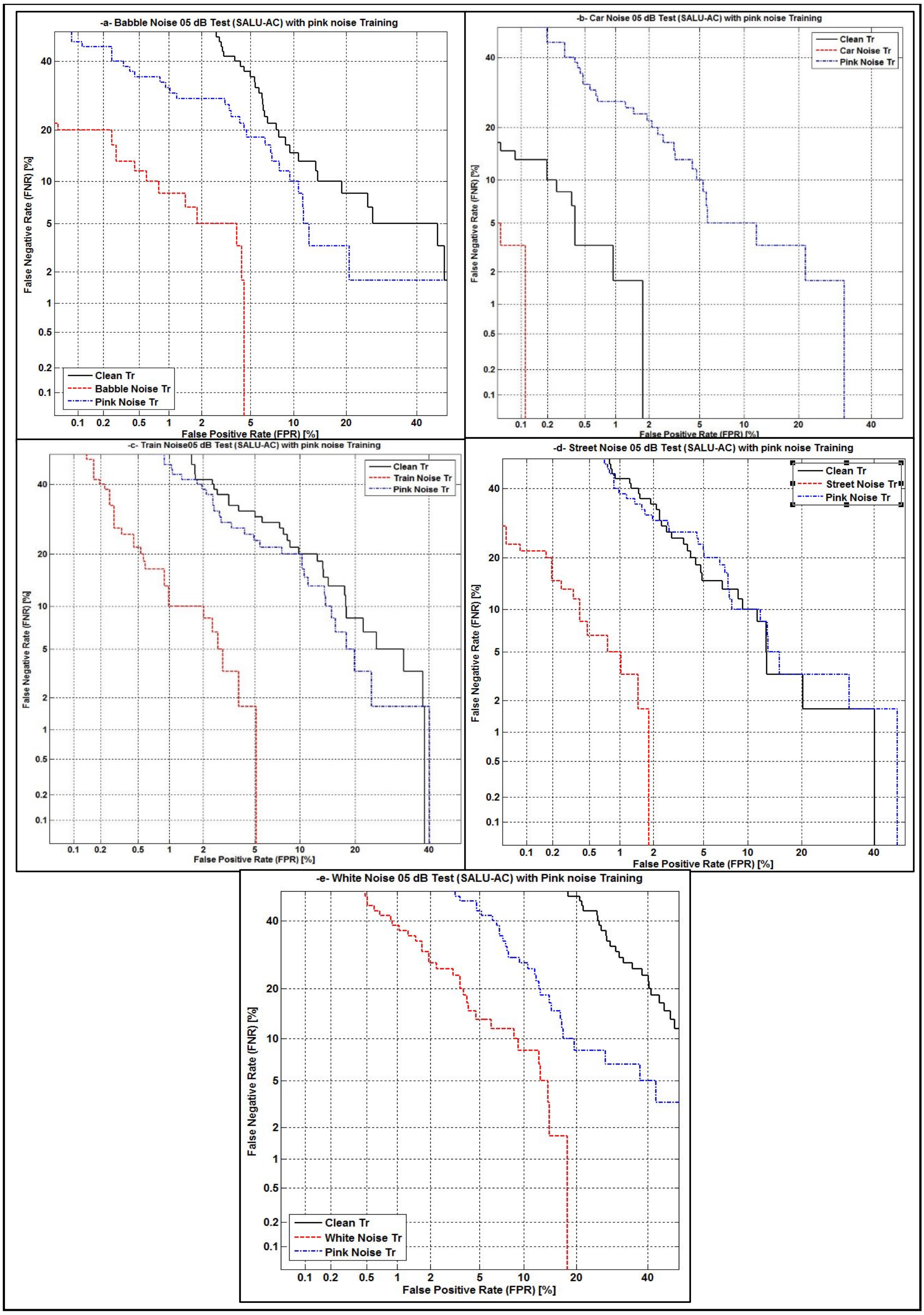


Figure I.2: DET graphs of noisy training for 05 dB SNR test samples

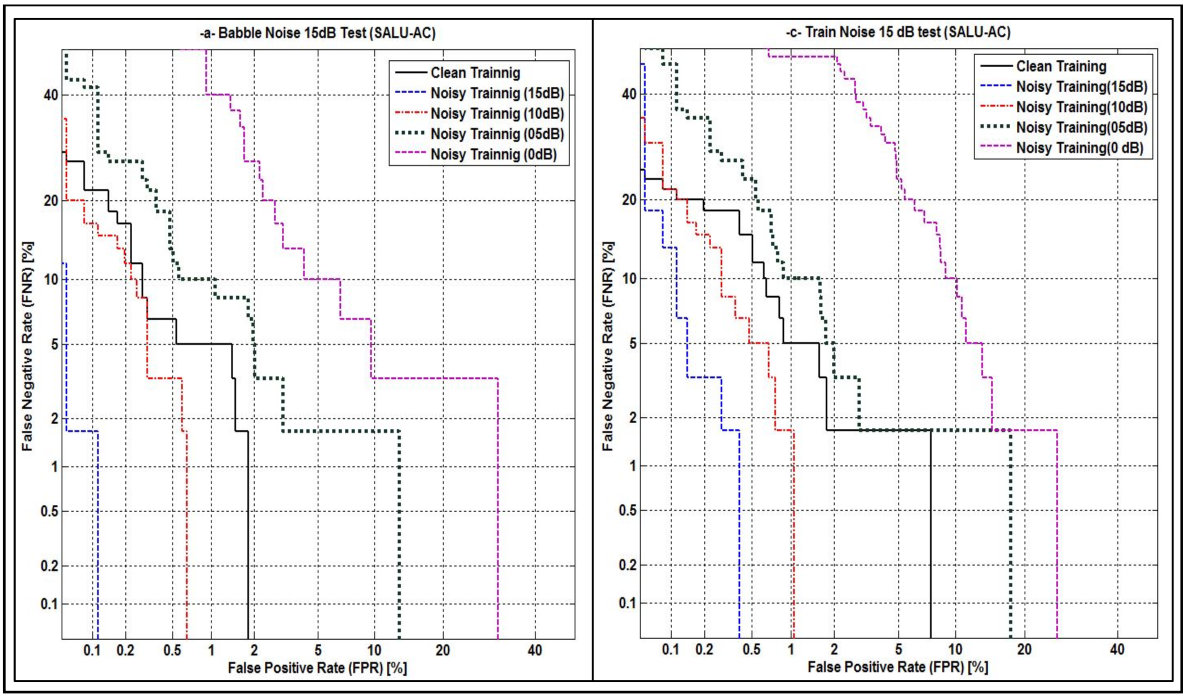


Figure I.3: DET Graphs for Clean and Different SNRs (in dB) Training (at 15 dB Test)

Appendix II
DET Graphs for Chapter 6

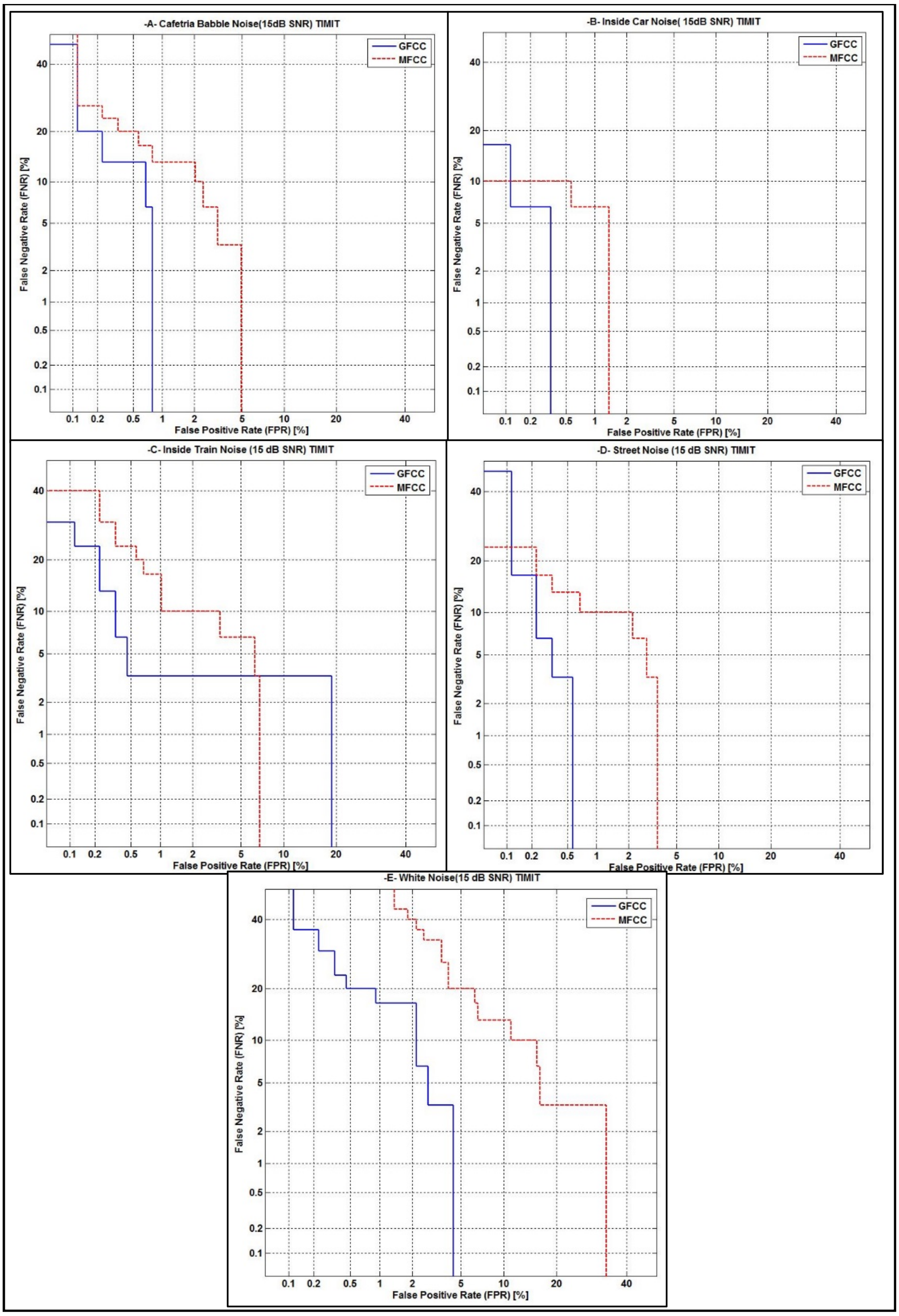


Figure II.1: DET Graphs for GFCC and MFCC baseline SR at 15 dB using TIMIT

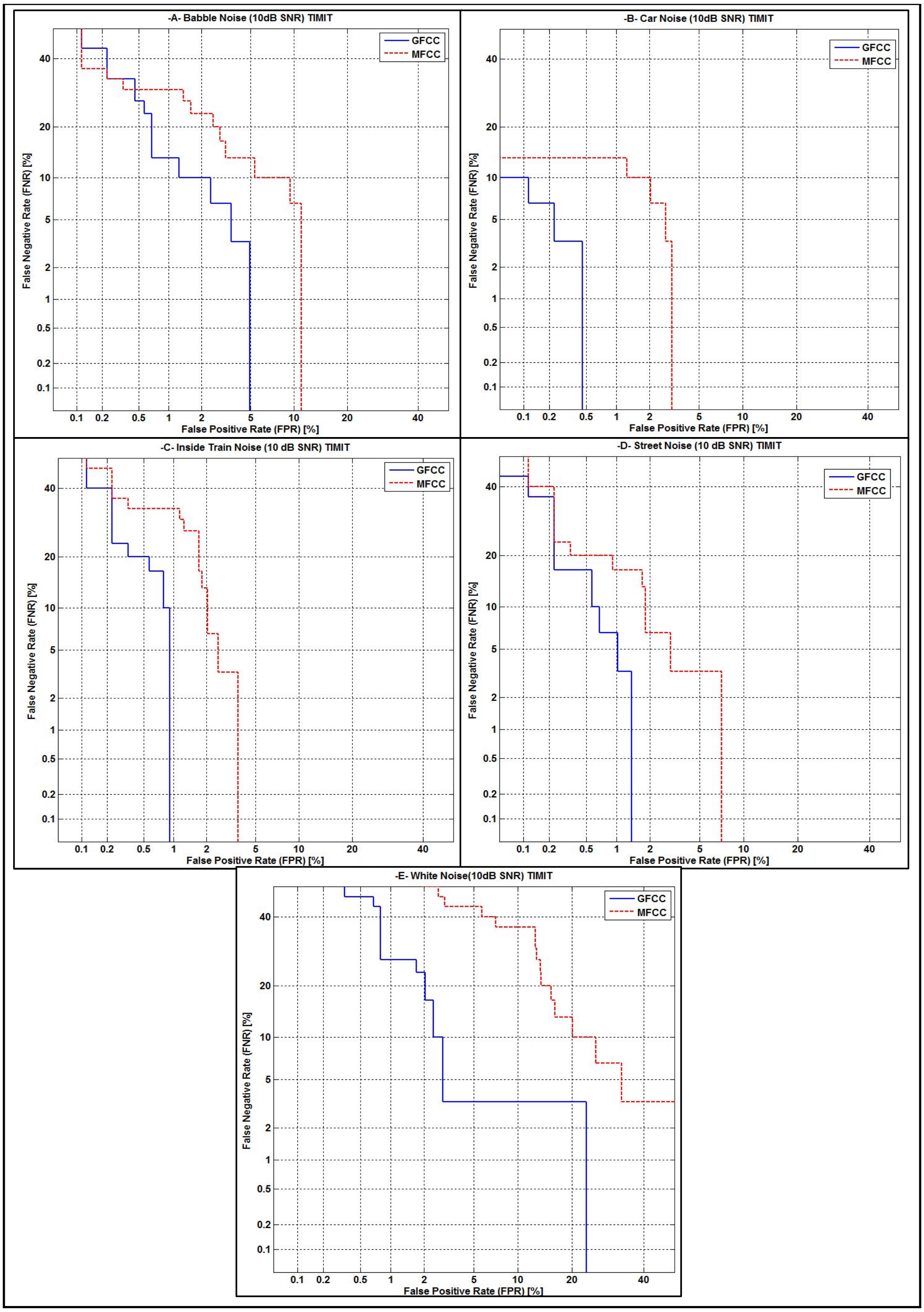


Figure II.2: DET Graphs for GFCC and MFCC baseline SR at 10 dB using TIMIT

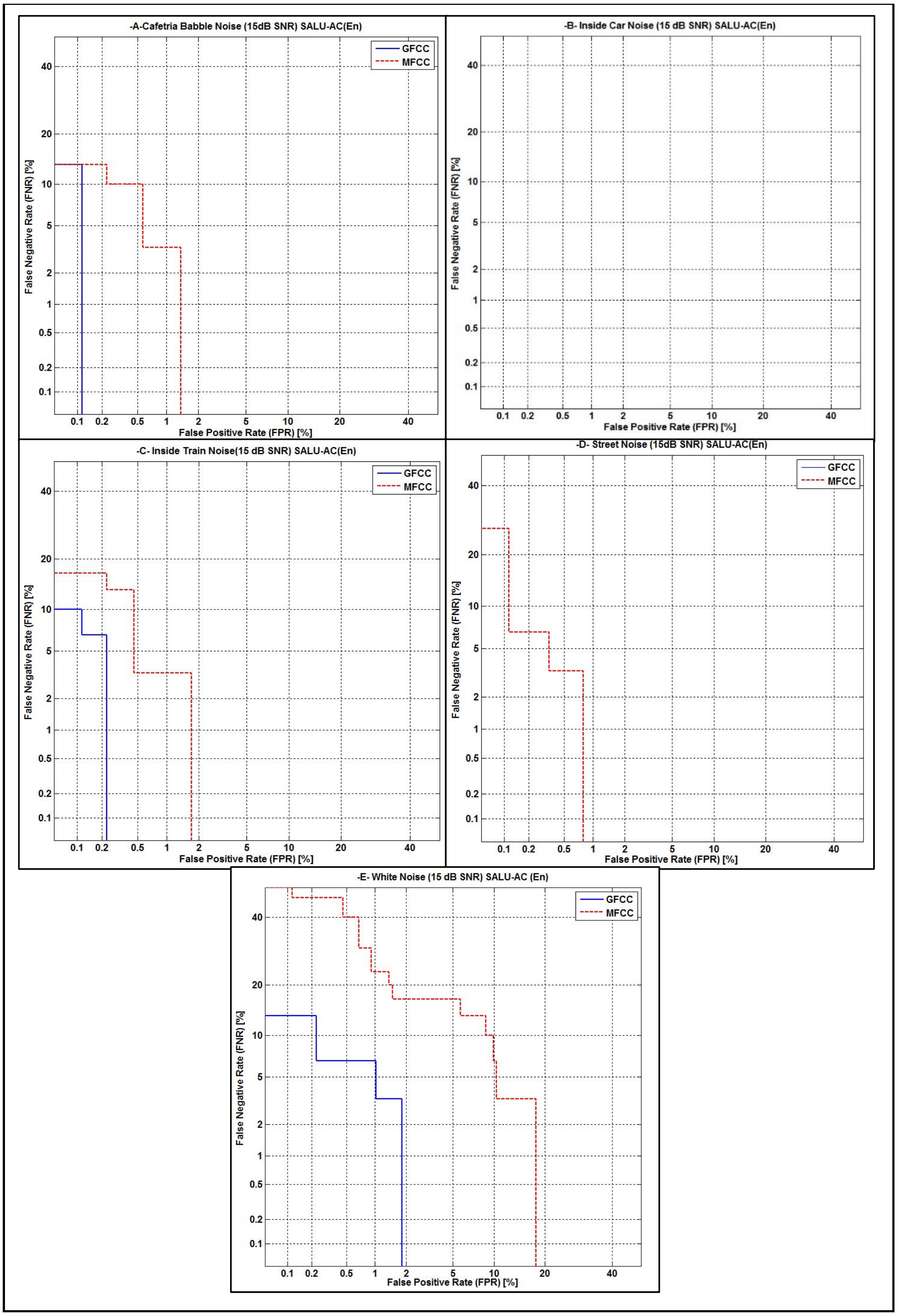


Figure II.3: DET Graphs for GFCC and MFCC baseline SR at 15 dB using SALU-AC (First Set)

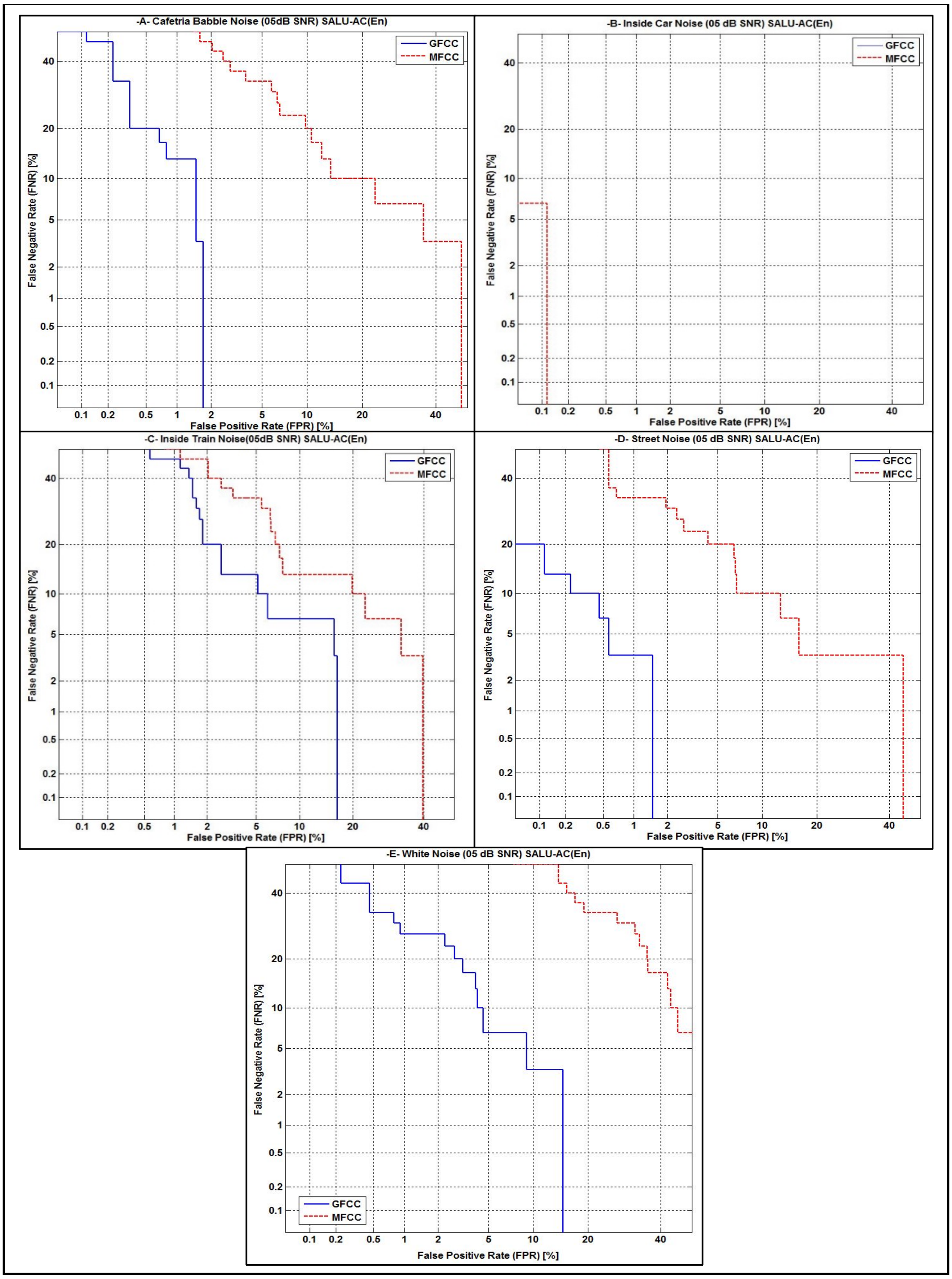


Figure II.4: DET Graphs for GFCC and MFCC baseline SR at 05 dB using SALU-AC (First Set)

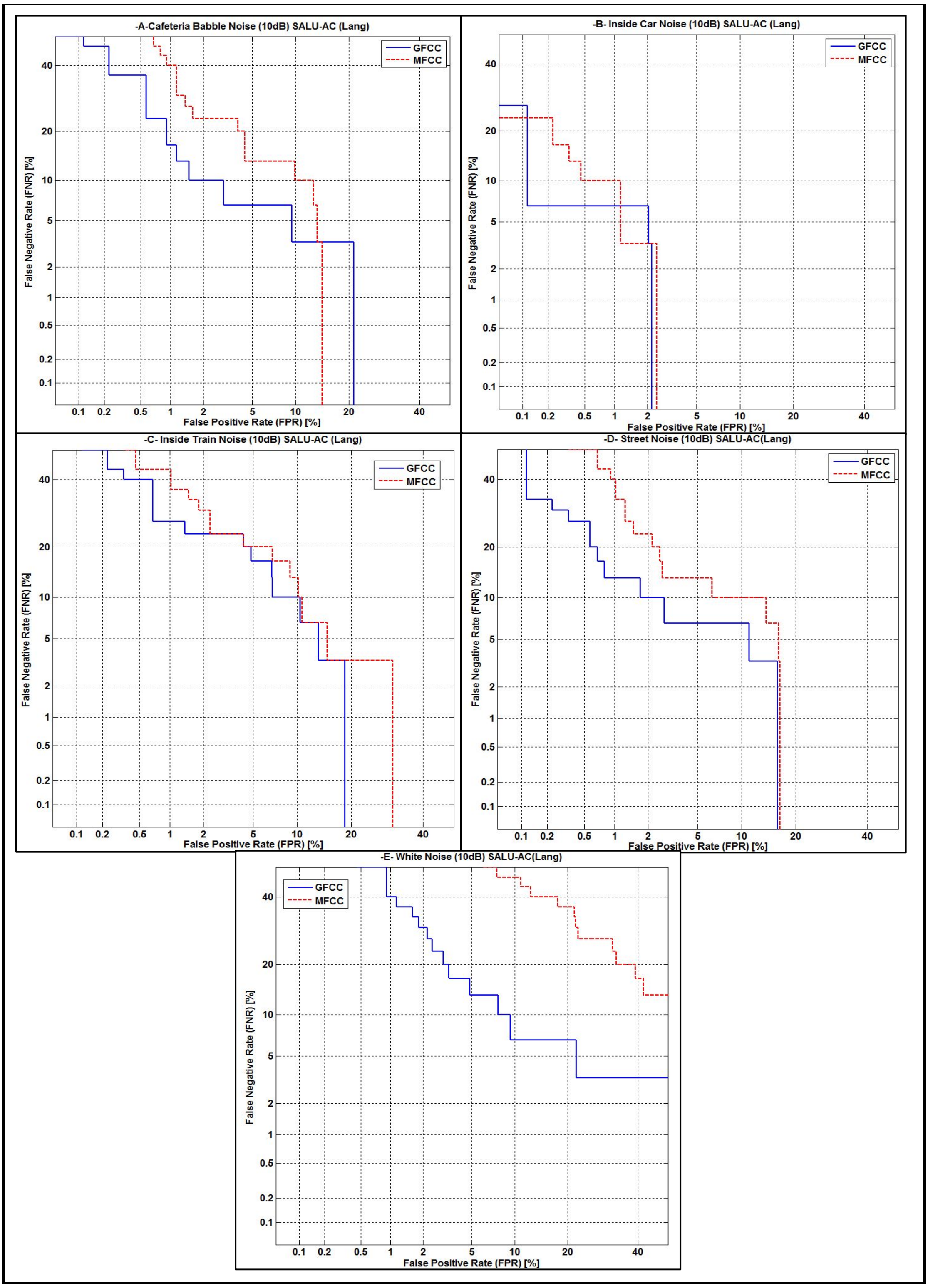


Figure II.5:DET Graphs for GFCC and MFCC baseline SR at 10 dB using SALU-AC (Second Set)

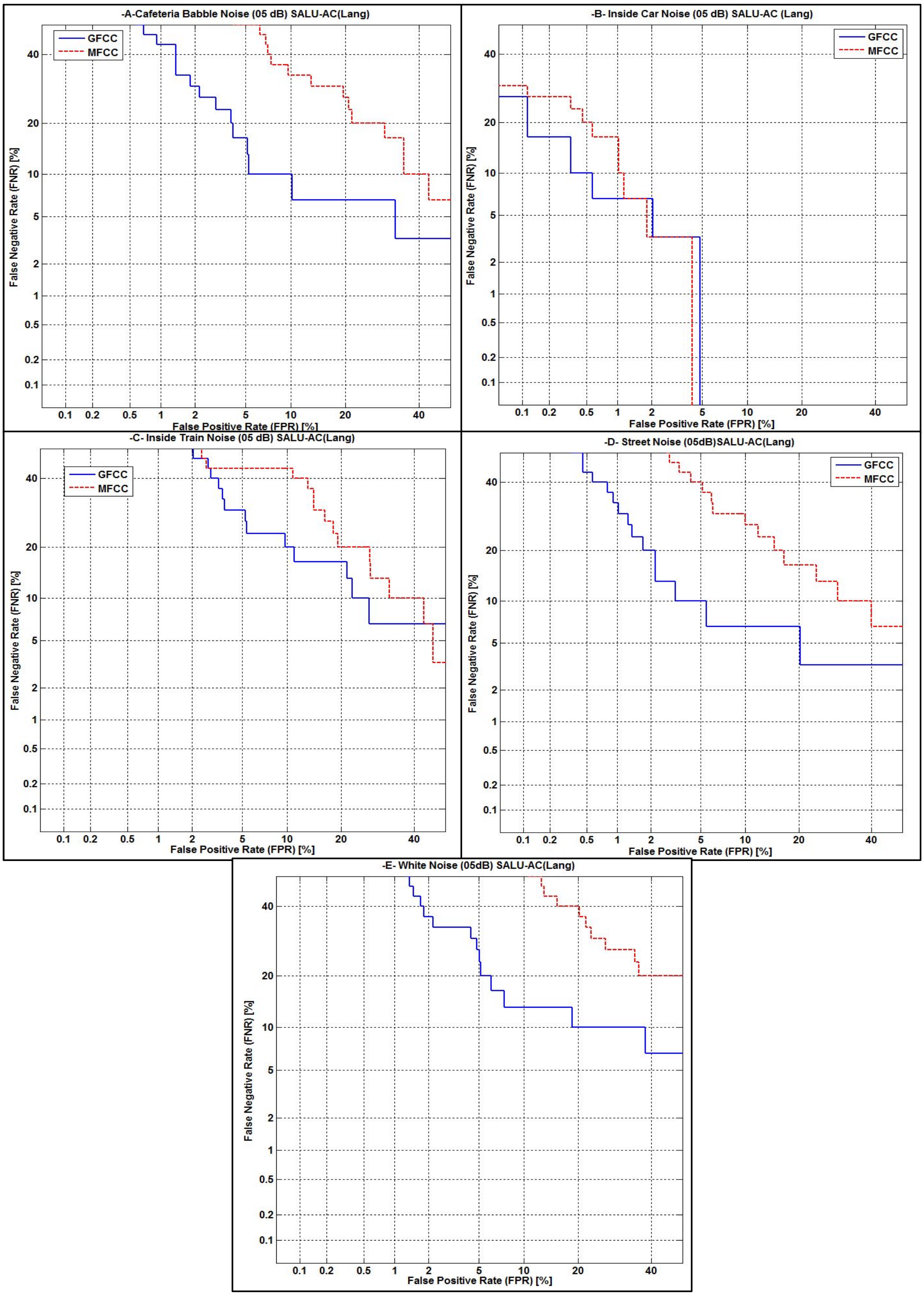


Figure II.6:DET Graphs for GFCC and MFCC baseline SR at 05 dB using SALU-AC (Second Set)

Appendix III
DET Graphs for Chapter 7

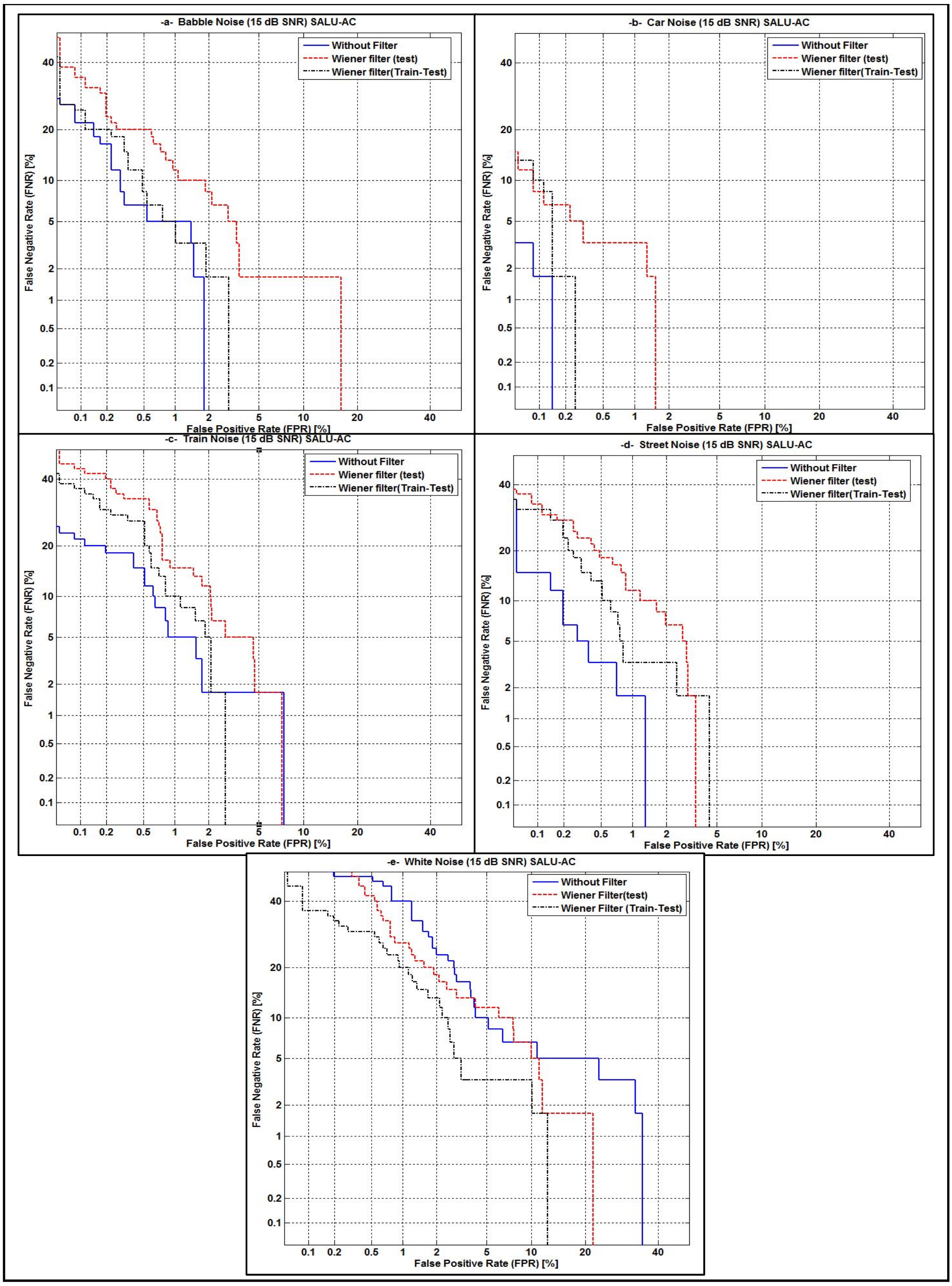


Figure III.1:DET graphs for SR with and Without Wiener filter at 15 dB

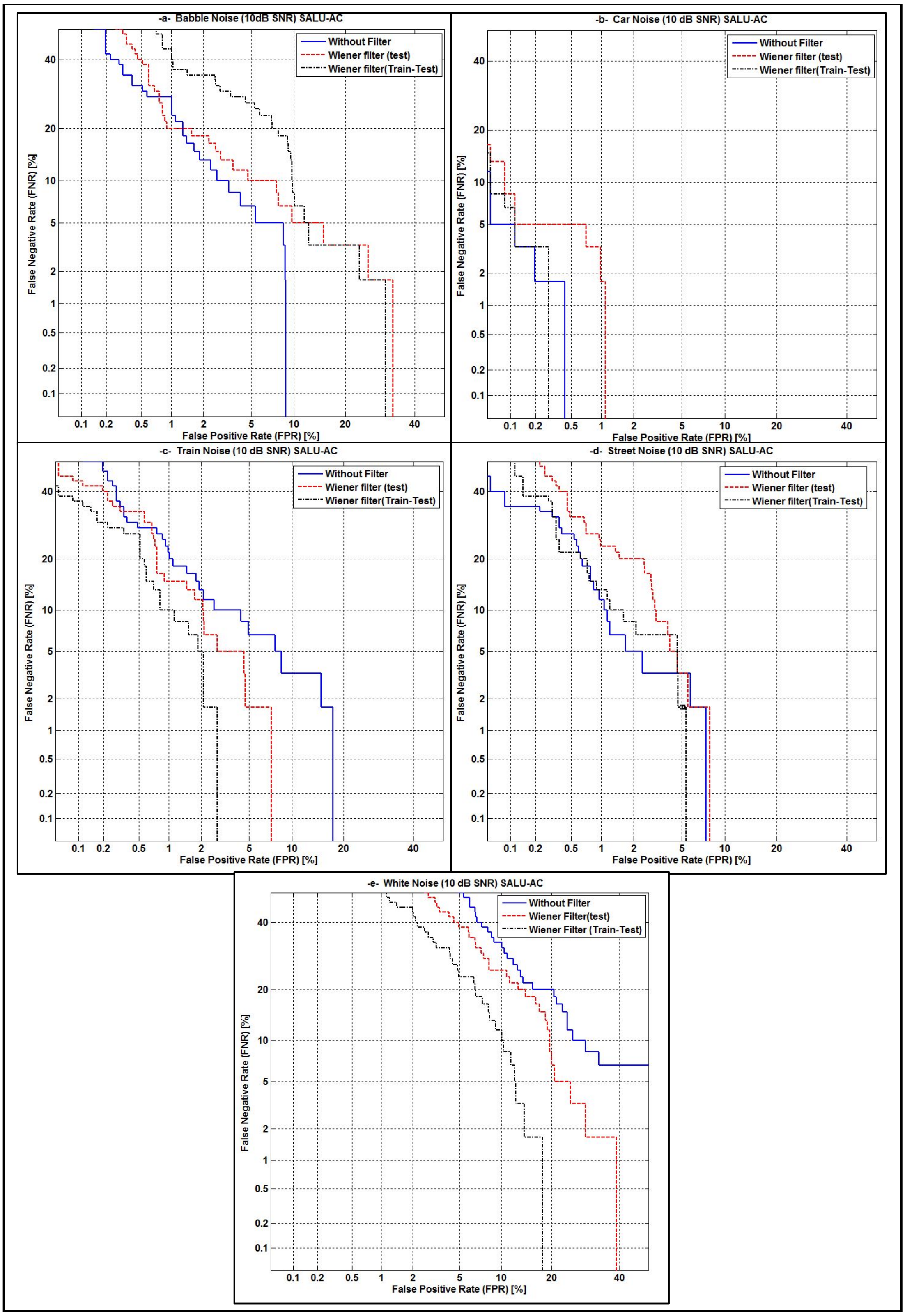


Figure III.2:DET graphs for SR with and Without Wiener filter at 10 dB

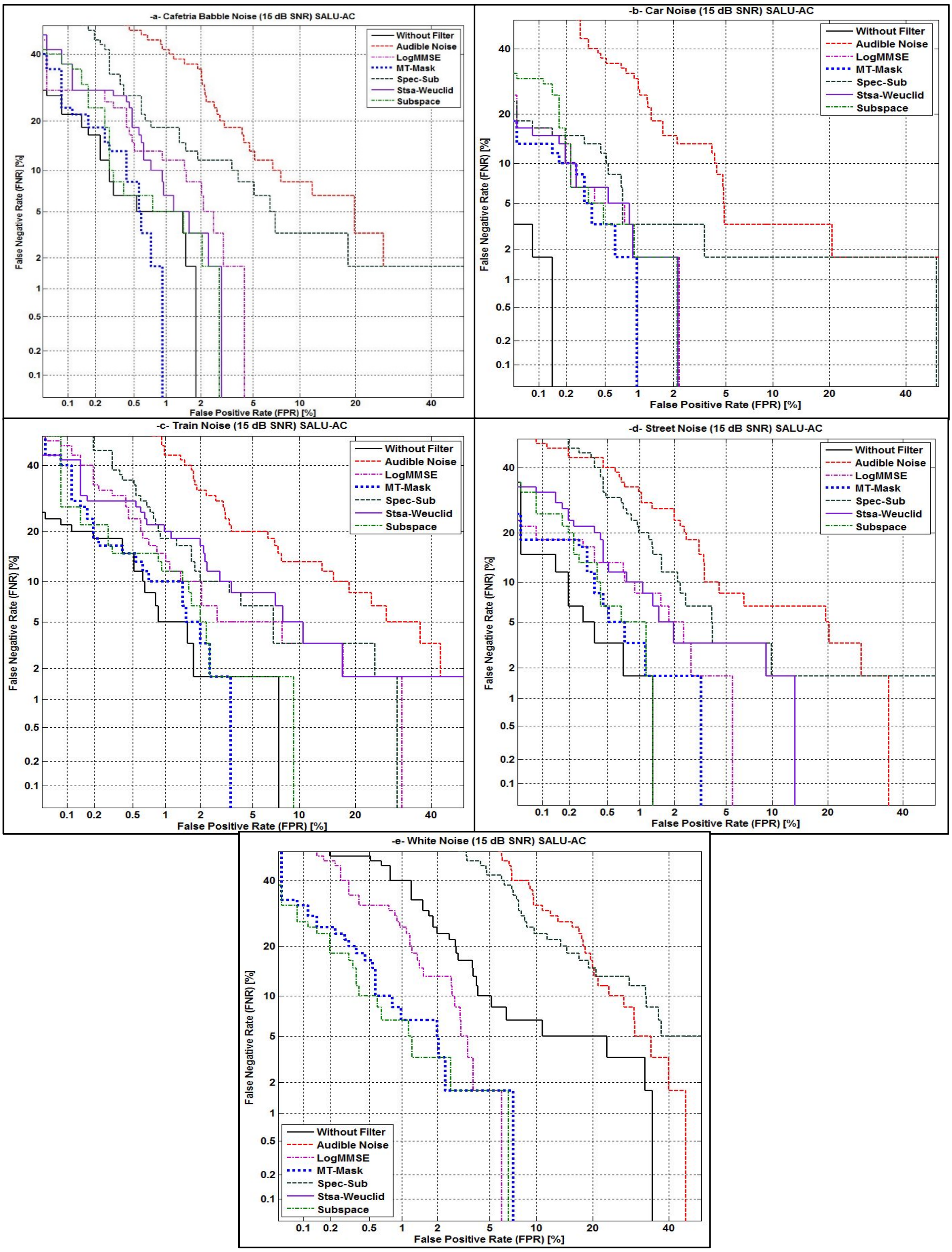


Figure III.3:DET graph with /without speech enhancement approaches at 15 dB SNR

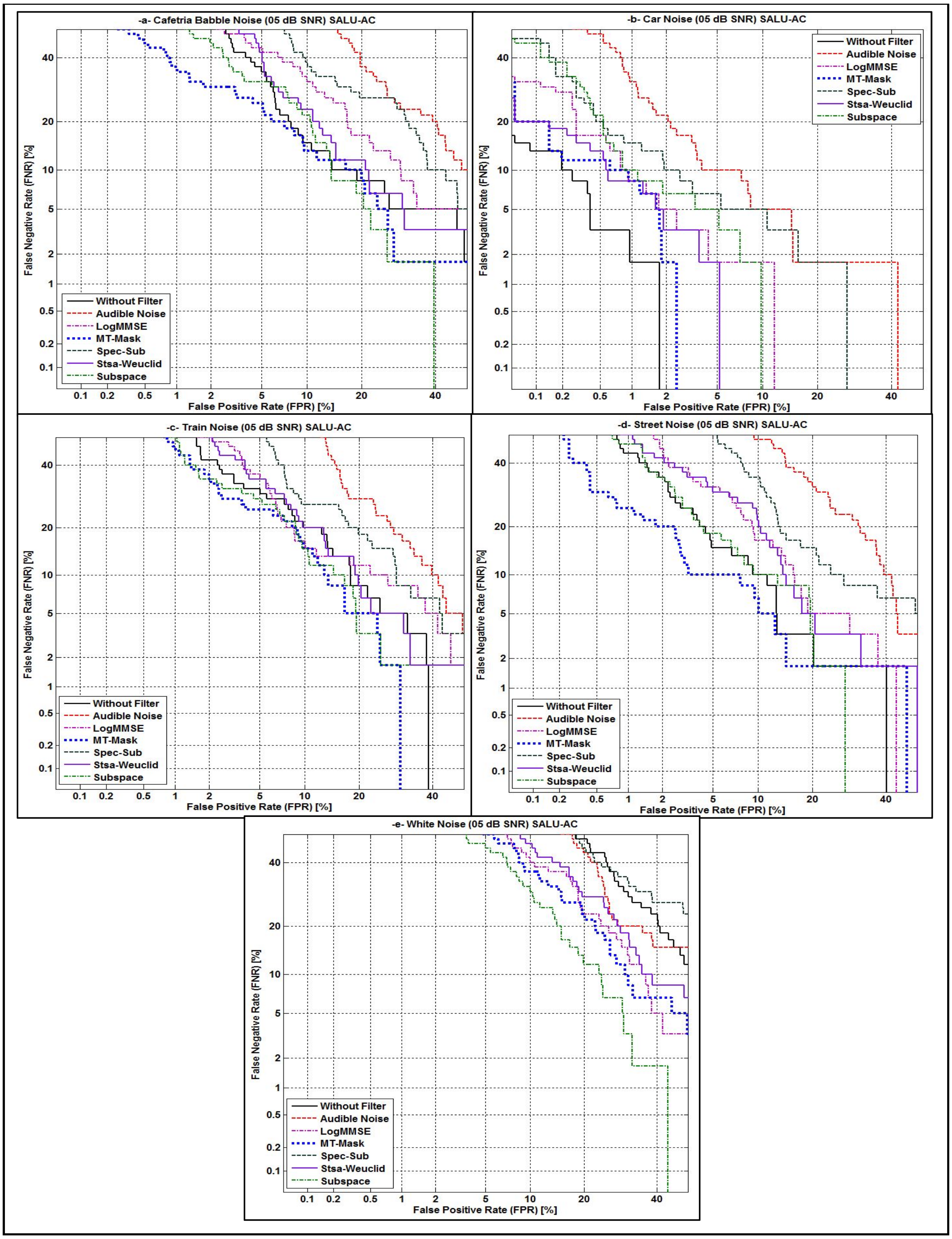


Figure III.4: DET graph with /without speech enhancement approaches at 05 dB SNR

Appendix IV
DET Graphs for chapter 8 and chapter 9
(Training on the fly)

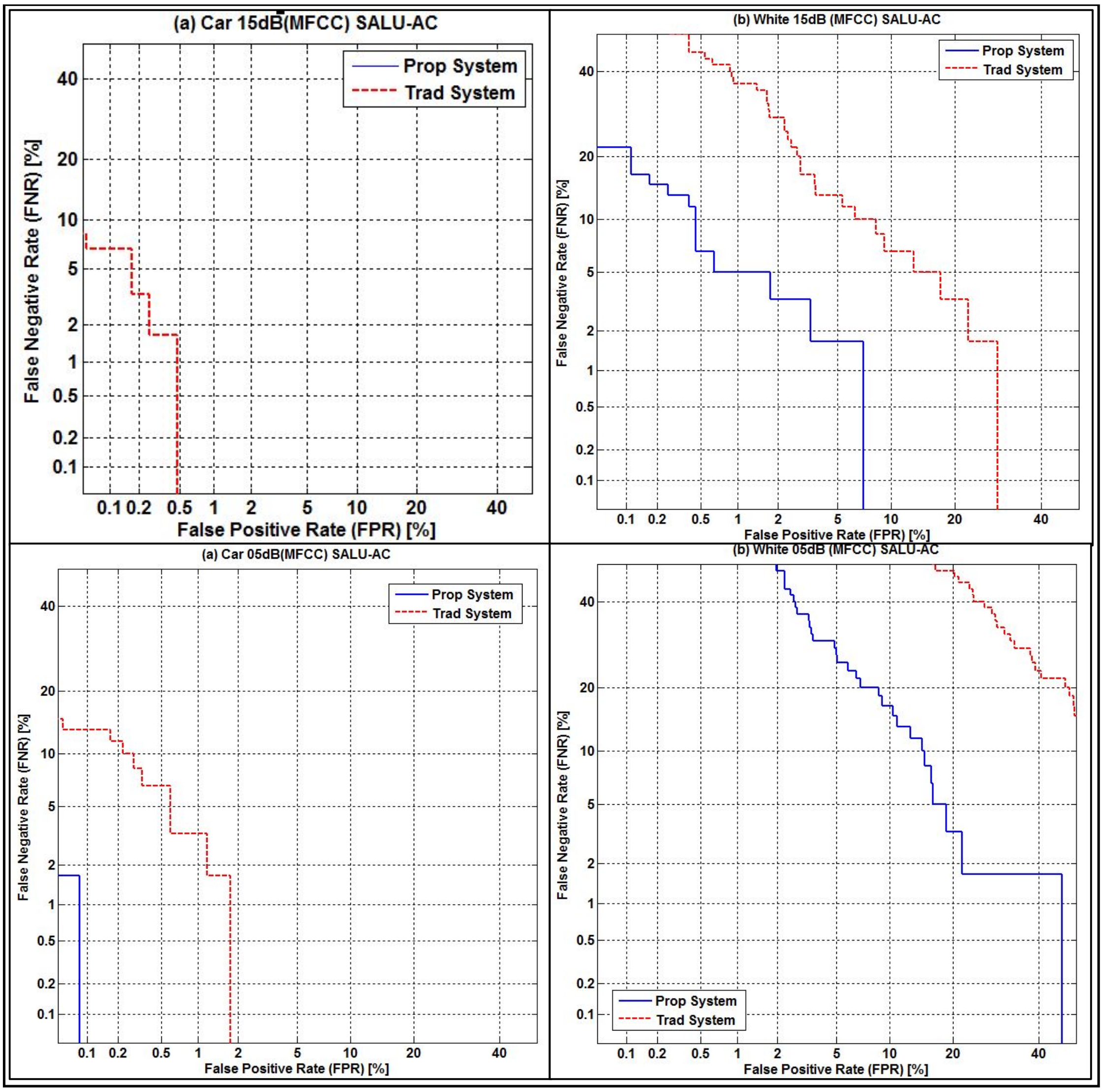


Figure IV.1:DET Graphs for Training on the fly (First technique) for Car, and White noise at 15db, 05 dB SNR using SALU-AC (MFCC baseline)

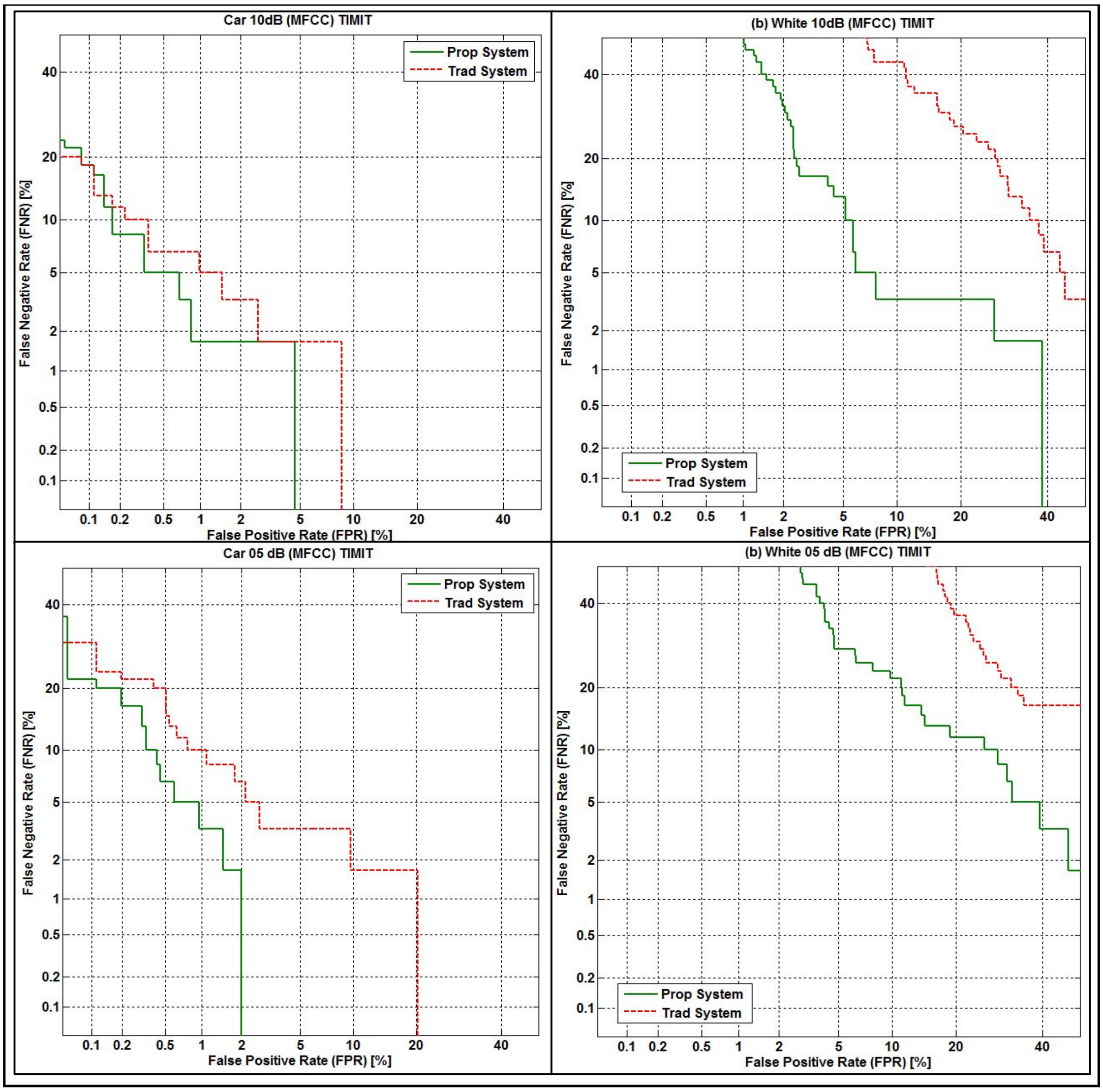


Figure IV.2: DET Graphs for Training on the fly(First technique) for Car, and White noise at 10db, 05 dB SNR using TIMIT (MFCC baseline)

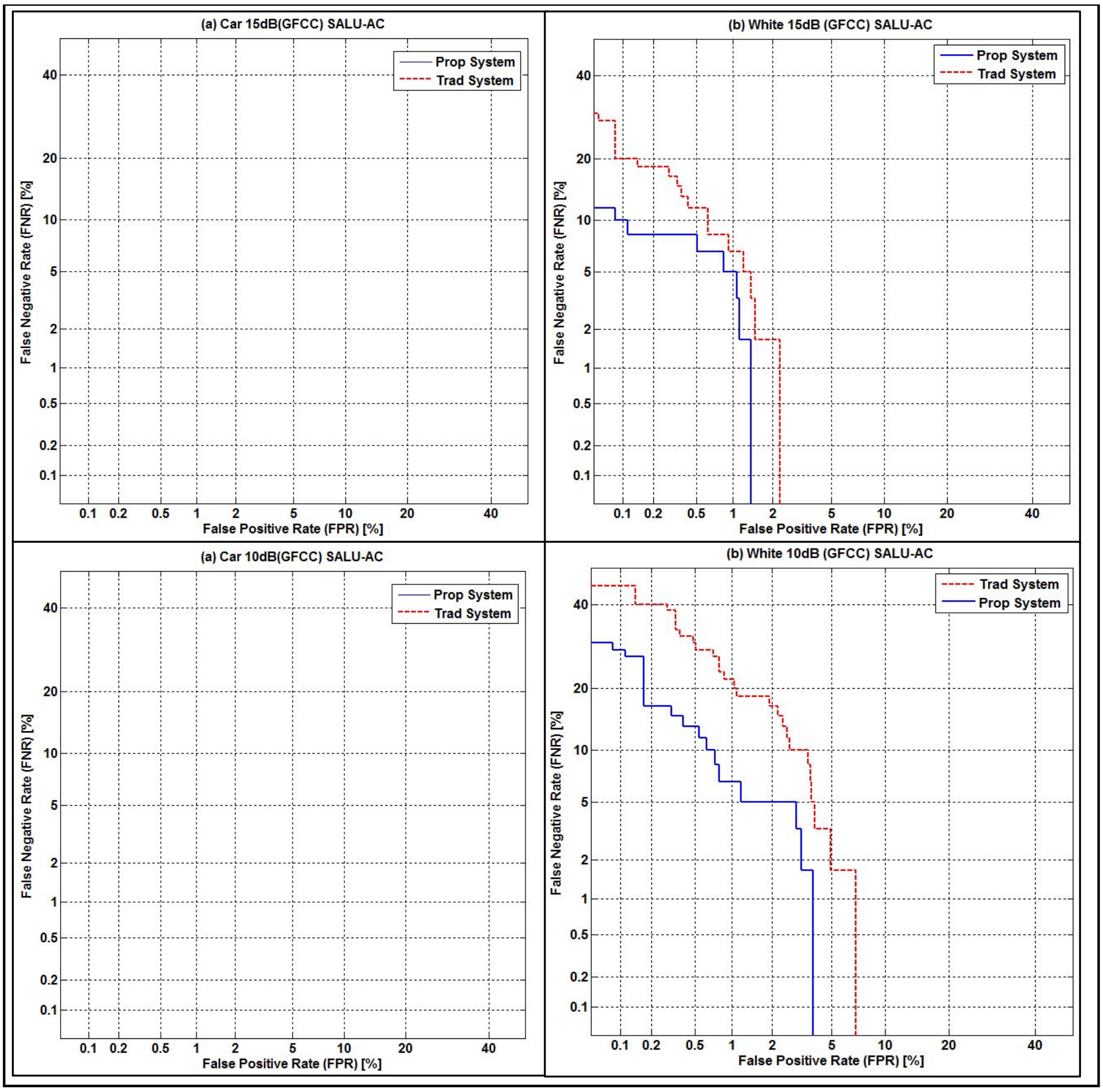


Figure IV.3:DET Graphs for Training on the fly (First technique) for Car, and White noise at 15db, 05 dB SNR using SALU-AC (GFCC baseline)

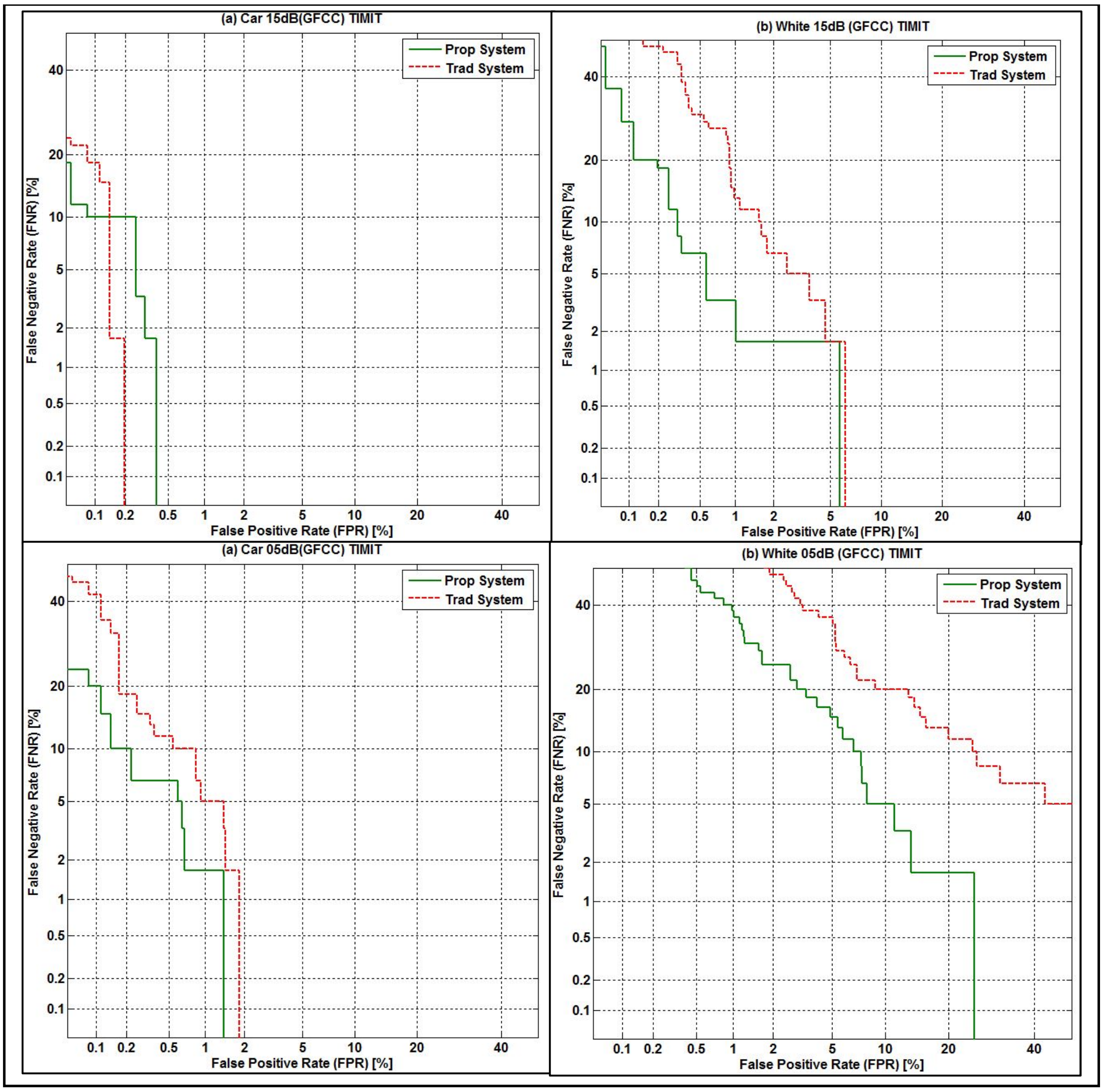


Figure IV.4: DET Graphs for Training on the fly (First technique) for Car, and White noise at 15db, 05 dB SNR using TIMIT (GFCC baseline)

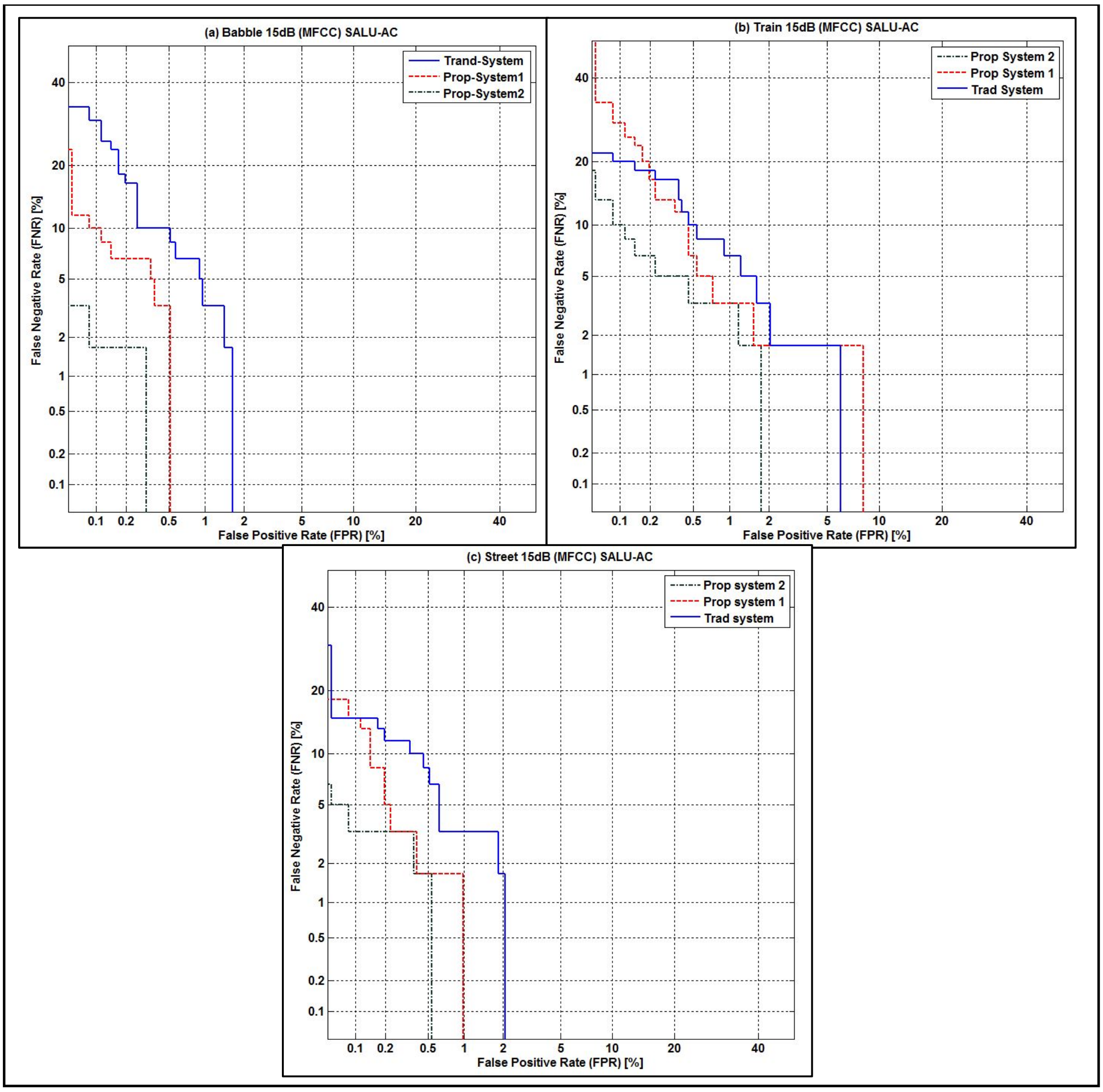


Figure IV.5:DET Graphs for Training on the fly (both techniques) for Babble, Train, and Street noise at 15dB SNR using SALU-AC (MFCC baseline)

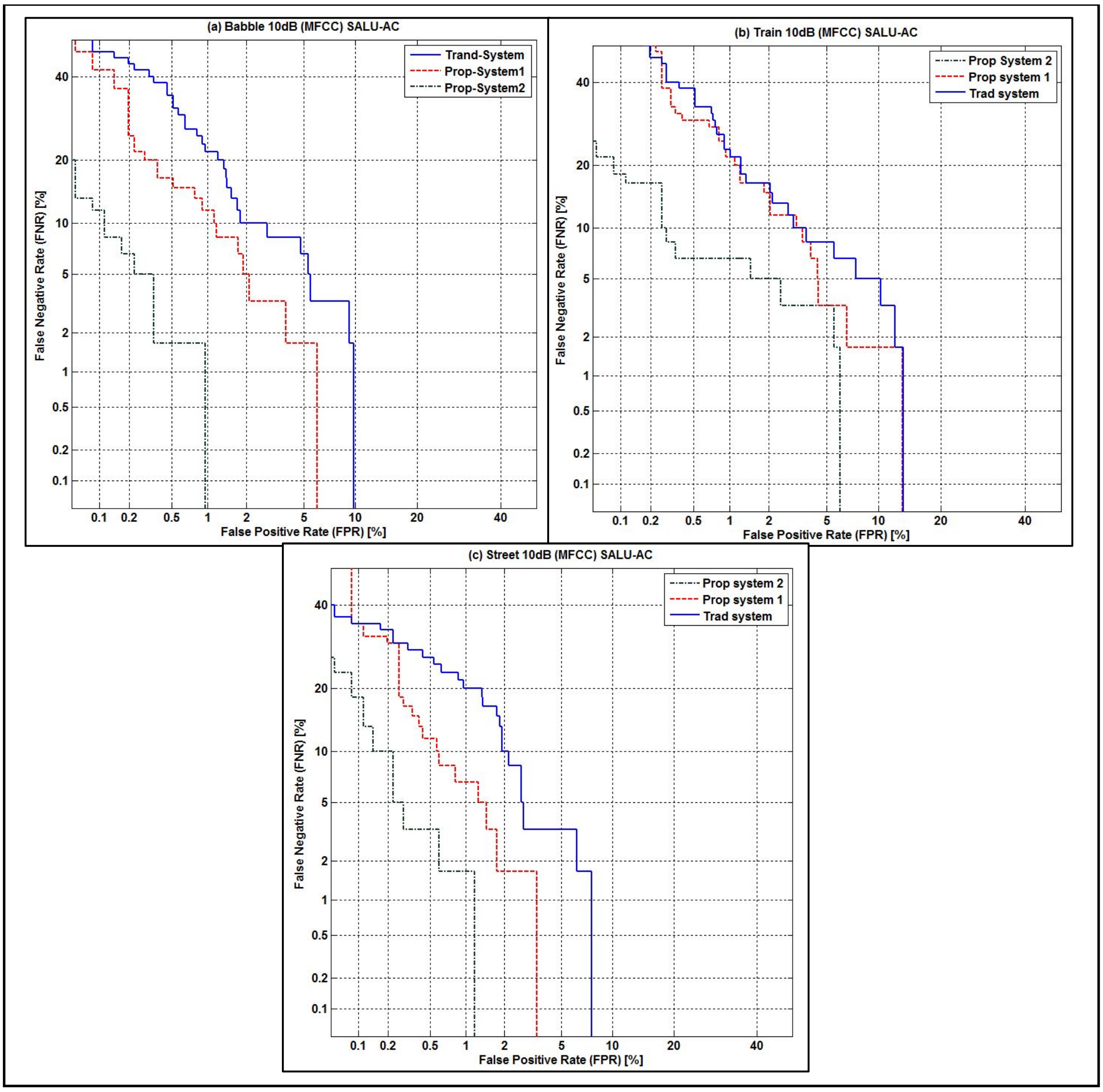


Figure IV.6:DET Graphs for Training on the fly (both techniques) for Babble, Train, and Street noise at 10 dB SNR using SALU-AC (MFCC baseline)

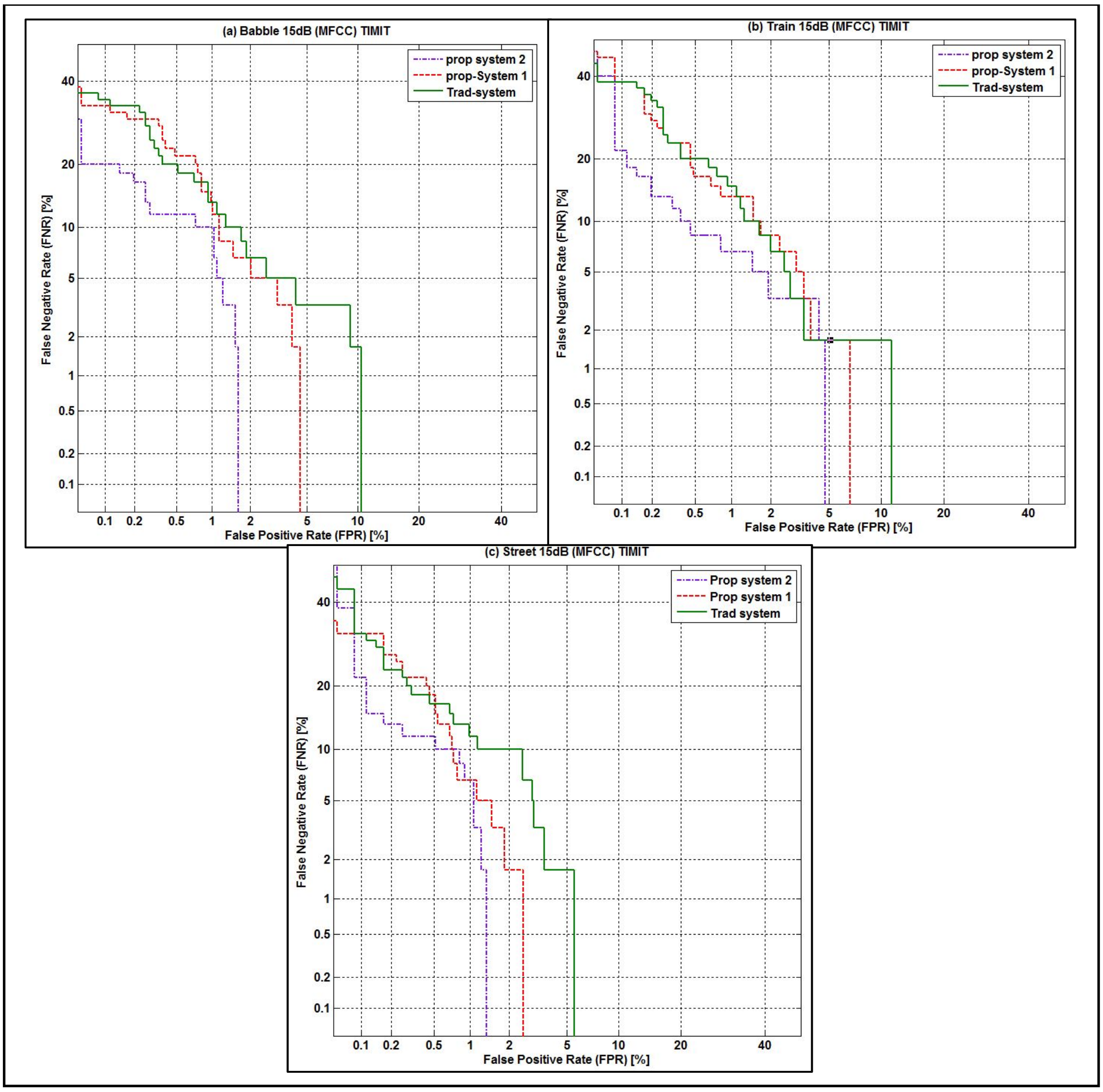
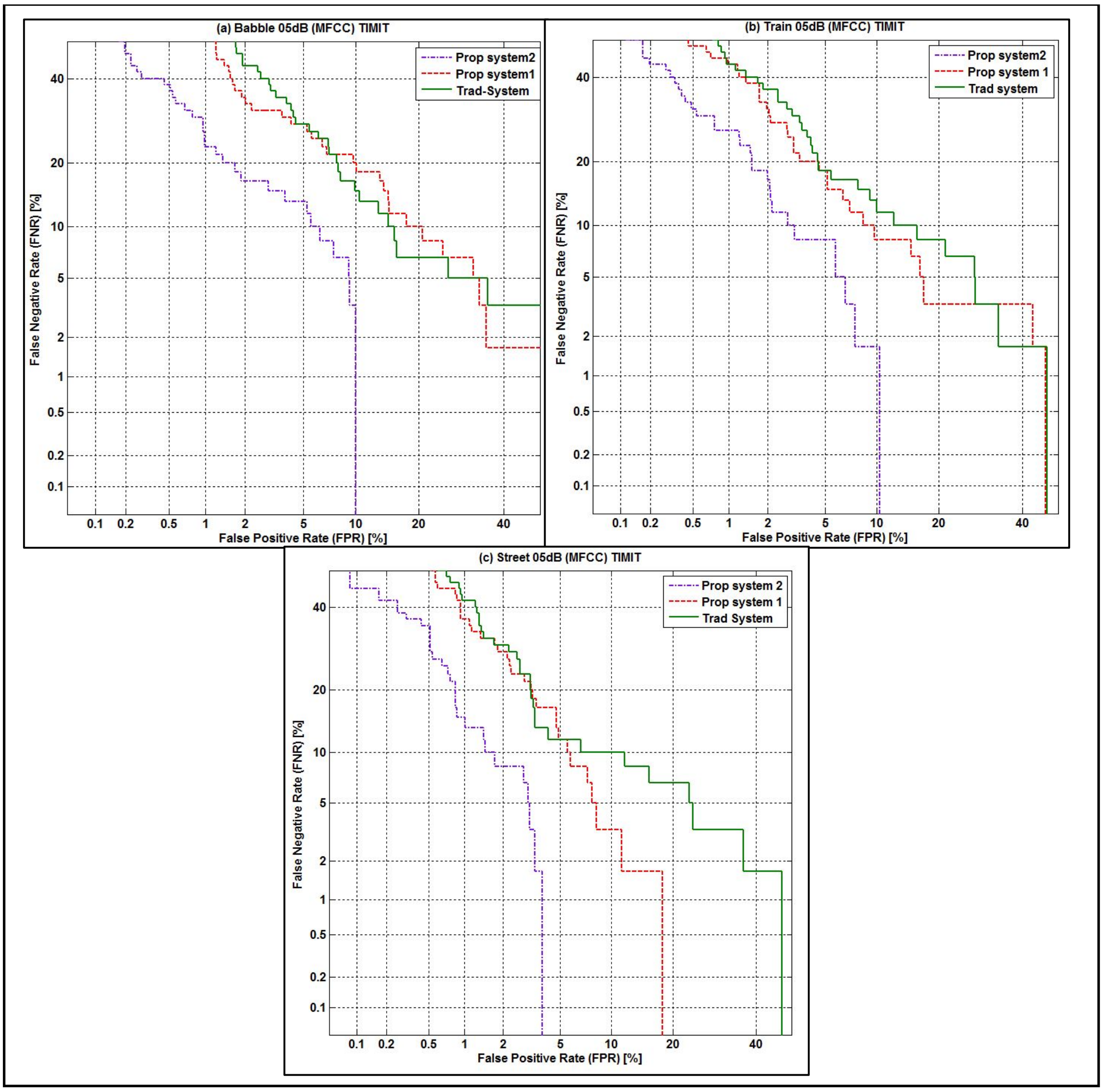


Figure IV.7:DET Graphs for Training on the fly (both techniques) for Babble, Train, and Street noise at 15 dB SNR using TIMIT (MFCC baseline)



(a) Babble 05dB (MFCC) TIMIT

(b) Train 05dB (MFCC) TIMIT

(c) Street 05dB (MFCC) TIMIT

Figure IV.8:DET Graphs for Training on the fly (both techniques) for Babble, Train, and Street noise at 05 dB SNR using TIMIT (MFCC baseline)

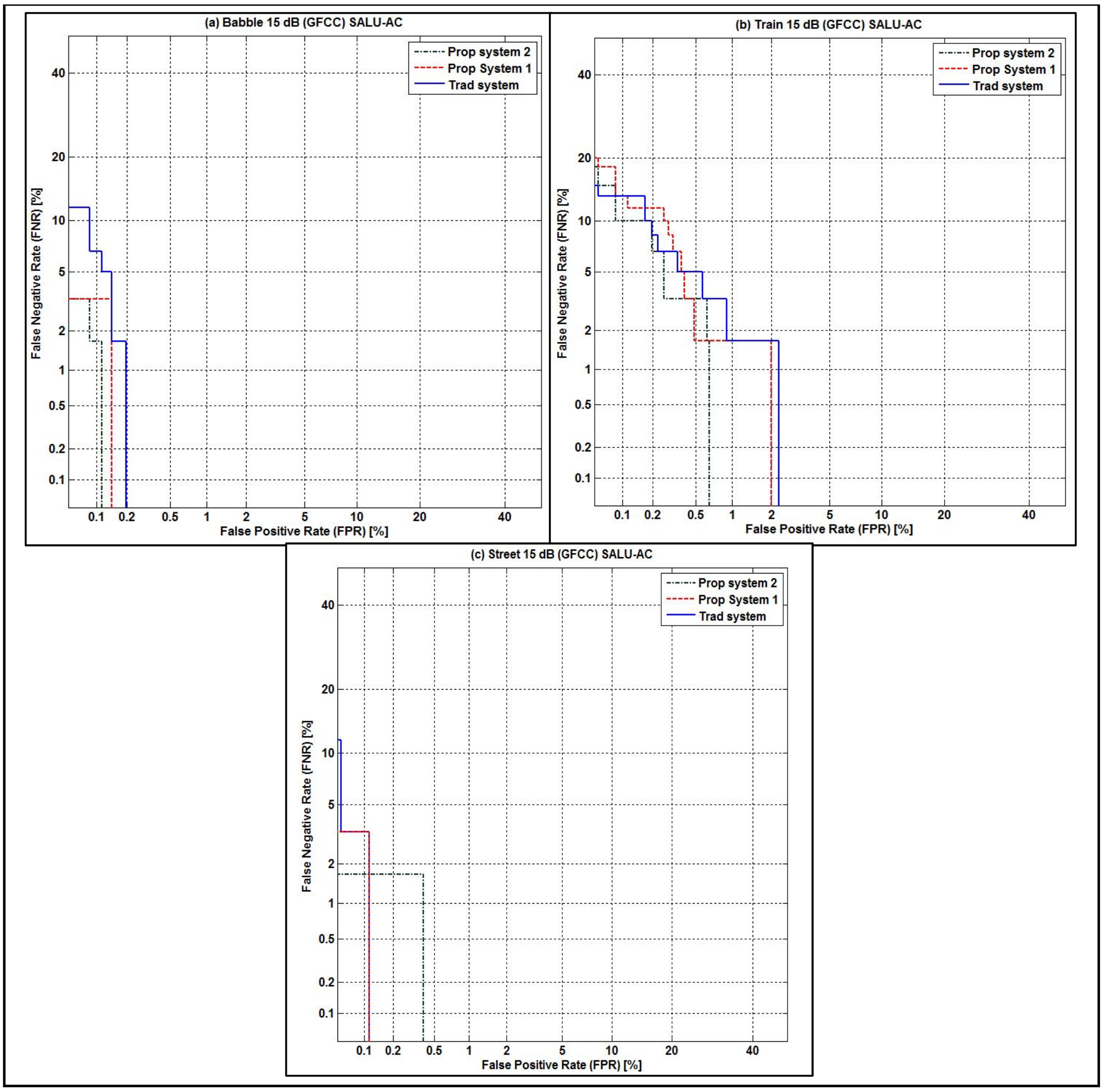


Figure IV.9 :DET Graphs for Training on the fly (both techniques) for Babble, Train, and Street noise at 15dB SNR using SALU-AC (GFCC baseline)

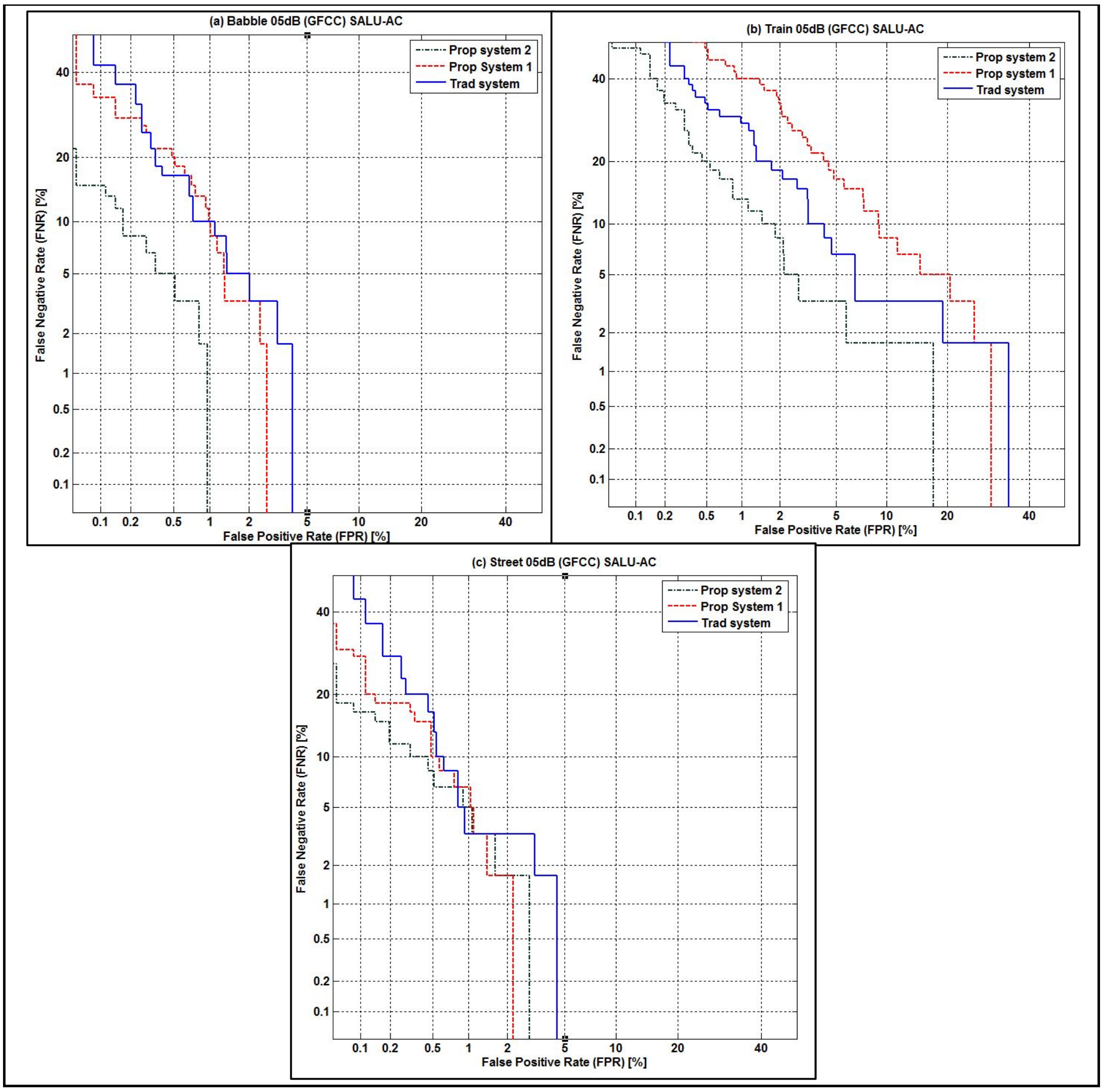


Figure IV.10:DET Graphs for Training on the fly (both techniques) for Babble, Train, and Street noise at 05dB SNR using SALU-AC (GFCC baseline)

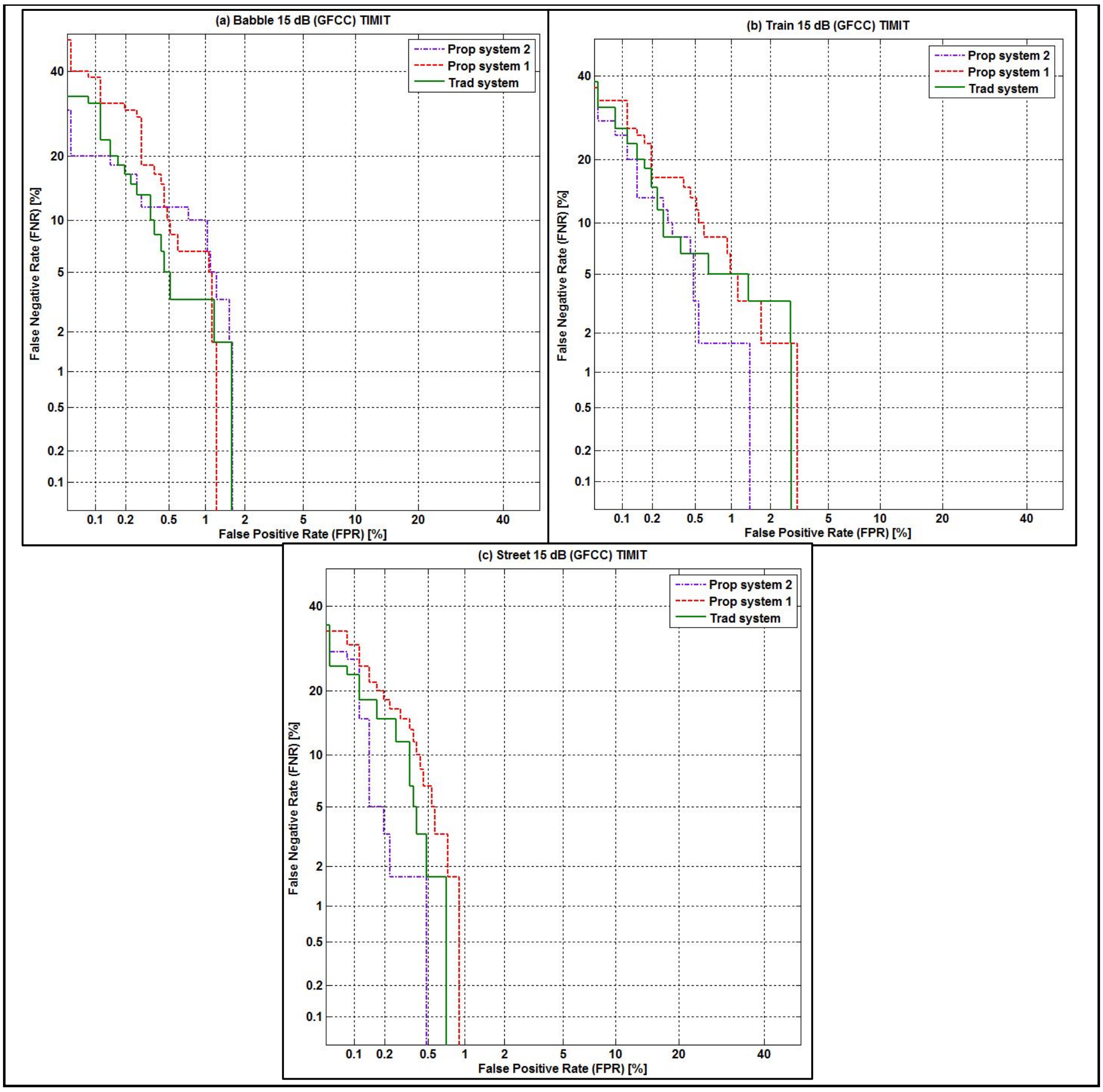


Figure IV.11: DET Graphs for Training on the fly (both techniques) for Babble, Train, and Street noise at 15dB SNR using TIMIT (GFCC baseline)

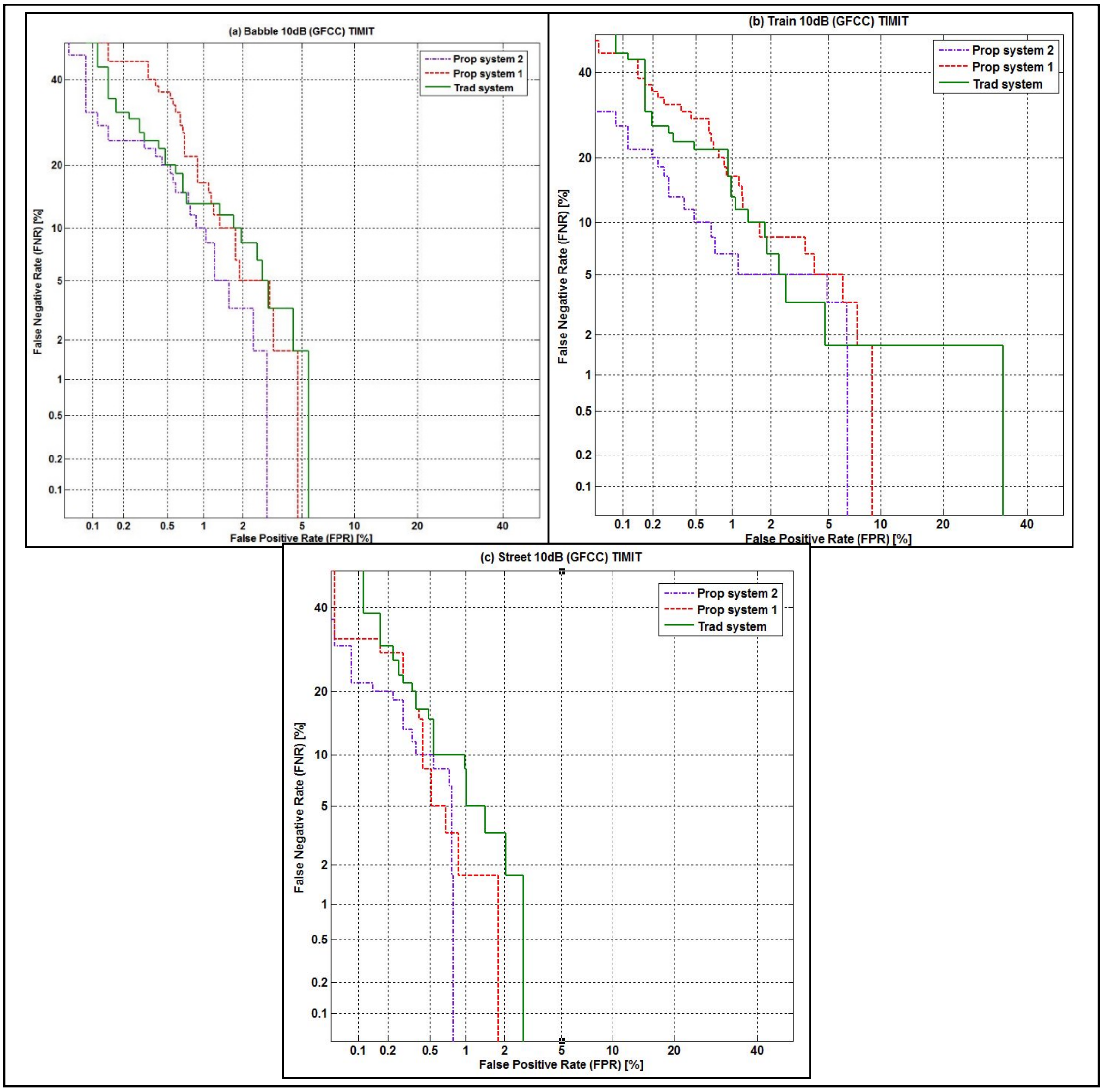
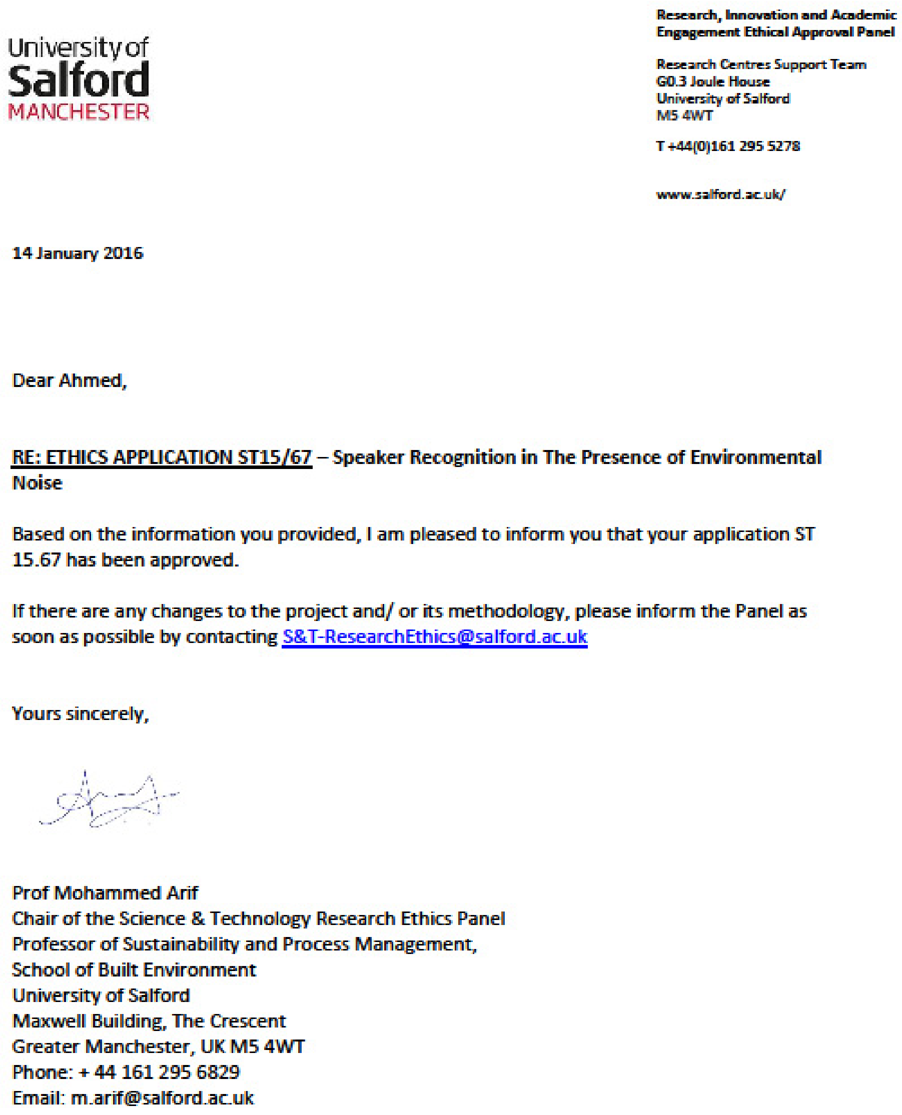


Figure IV.12: DET Graphs for Training on the fly (both techniques) for Babble, Train, and Street noise at 10 dB SNR using TIMIT (GFCC baseline)

Appendix V
Ethical Approval



University of Salford
MANCHESTER

Research, Innovation and Academic
Engagement Ethical Approval Panel

Research Centres Support Team
G0.3 Joule House
University of Salford
M5 4WT

T +44(0)161 295 5278

www.salford.ac.uk/

14 January 2016

Dear Ahmed,

RE: ETHICS APPLICATION ST15/67 – Speaker Recognition in The Presence of Environmental Noise

Based on the information you provided, I am pleased to inform you that your application ST 15.67 has been approved.

If there are any changes to the project and/ or its methodology, please inform the Panel as soon as possible by contacting S&T-ResearchEthics@salford.ac.uk

Yours sincerely,

Prof Mohammed Arif
Chair of the Science & Technology Research Ethics Panel
Professor of Sustainability and Process Management,
School of Built Environment
University of Salford
Maxwell Building, The Crescent
Greater Manchester, UK M5 4WT
Phone: + 44 161 295 6829
Email: m.arif@salford.ac.uk

REFERENCES

Ariyaeeinia, A.M. and Sivakumaran, P., 1997, September. Analysis and comparison of score normalisation methods for text-dependent speaker verification. In Eurospeech (Vol. 97, pp. 1379-1382).

Al-Noori, A., Li, F.F. and Duncan, P.J., 2016, May. Robustness of speaker recognition from noisy speech samples and mismatched languages. In Audio Engineering Society Convention 140. Audio Engineering Society.

Al-Noori, A., Li, F.F. and Duncan, P.J., 2017, Training 'on the fly' to improve the performance of Speaker Recognition in Noisy Environments. 2017 AES International Conference on Audio Forensics-Finding Signal in the Noise. Audio Engineering Society.

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

Auckenthaler, R., Carey, M. and Lloyd-Thomas, H., 2000. Score normalization for text-independent speaker verification systems. Digital Signal Processing, 10(1), pp.42-54.

Bao, H.C. and Juan, Z.C., 2012, June. The research of speaker recognition based on GMM and SVM. In System Science and Engineering (ICSSE), 2012 International Conference on (pp. 373-375). IEEE.

Beigi, H., 2009, July. Effects of time lapse on speaker recognition results. In Digital Signal Processing, 2009 16th International Conference on (pp. 1-6). IEEE.

Beigi, H., 2011. Fundamentals of speaker recognition. Springer Science & Business Media.

Beigi, H., 2012. Speaker Recognition: Advancements and Challenges. INTECH Open Access Publisher.

Beritelli, F., 2008, December. Effect of background noise on the snr estimation of biometric parameters in forensic speaker recognition. In Signal Processing and Communication Systems, 2008. ICSPCS 2008. 2nd International Conference on (pp. 1-5). IEEE.

Besacier, L. and Bonastre, J.F., 1998, May. Frame pruning for speaker recognition. In Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on (Vol. 2, pp. 765-768). IEEE.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on acoustics, speech, and signal processing, 27(2), pp.113-120.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), pp.121-167.

Burgos, W., 2014. Gammatone and MFCC Features in Speaker Recognition (Doctoral dissertation, Florida Institute of Technology).

Campbell, W.M., Assaleh, K.T. and Broun, C.C., 2002. Speaker recognition with polynomial classifiers. IEEE Transactions on Speech and Audio Processing, 10(4), pp.205-212.

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

Campbell, W.M., Sturim, D.E. and Reynolds, D.A., 2006b. Support vector machines using GMM supervectors for speaker verification. IEEE signal processing letters, 13(5), pp.308-311.

Campbell, W.M., Sturim, D.E., Reynolds, D.A. and Solomonoff, A., 2006, May. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on (Vol. 1, pp. I-I). IEEE.

Cerva, P., Silovsky, J. and Zdansky, J., 2009. Comparison of generative and discriminative approaches for speaker recognition with limited data. Radioengineering.

Chan, W.N., Zheng, N. and Lee, T., 2007. Discrimination power of vocal source and vocal tract related features for speaker segmentation. IEEE Transactions on Audio, Speech, and Language Processing, 15(6), pp.1884-1892.

Chatzis, V., Bors, A.G. and Pitas, I., 1999. Multimodal decision-level fusion for person authentication. IEEE transactions on systems, man, and cybernetics-part a: systems and humans, 29(6), pp.674-680.

Chauhan, T., Soni, H. and Zafar, S., 2013. A review of automatic speaker recognition system.

Chen, J., Benesty, J., Huang, Y. and Doclo, S., 2006. New insights into the noise reduction Wiener filter. IEEE Transactions on audio, speech, and language processing, 14(4), pp.1218-1234.

Cohen, I., 2002. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal processing letters, 9(4), pp.113-116.

Combrinck, H.P. and Botha, E.C., 1996. On the mel-scaled cepstrum. department of Electrical and Electronic Engineering, University of Pretoria.

Das, P. and Bhattacharjee, U., 2014, July. Robust speaker verification using GFCC and joint factor analysis. In Computing, Communication and Networking Technologies (ICCCNT), 2014 International Conference on (pp. 1-4). IEEE.

Davis, S. and 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), pp.357-366.

de Araújo, J.M., de Menezes, J.M.P., Moura de Albuquerque, A.A., da Mota Almeida, O. and Ugulino de Araújo, F.M., 2013. Assessment and certification of neonatal incubator sensors through an inferential neural network. Sensors, 13(11), pp.15613-15632.

Dehak, N., 2009. Discriminative and generative approaches for long-and short-term speaker characteristics modeling: application to speaker verification (Doctoral dissertation, École de technologie supérieure).

Dehak, N., Dehak, R., Glass, J.R., Reynolds, D.A. and Kenny, P., 2010, June. Cosine similarity scoring without score normalization techniques. In Odyssey (p. 15).

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

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), pp.1-38.

Dendrinos, M., Bakamidis, S. and Carayannis, G., 1991. Speech enhancement from noise: A regenerative approach. Speech Communication, 10(1), pp.45-57.

Doddington, G.R., 2001, September. Speaker recognition based on idiolectal differences between speakers. In Interspeech (pp. 2521-2524).

Drygajlo, A. and El-Maliki, M., 1998, May. Speaker verification in noisy environments with combined spectral subtraction and missing feature theory. In Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on (Vol. 1, pp. 121-124). IEEE.

Ephraim, Y. and Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(6), pp.1109-1121.

Ephraim, Y. and Van Trees, H.L., 1995. A signal subspace approach for speech enhancement. IEEE Transactions on speech and audio processing, 3(4), pp.251-266.

ETSI, 2007. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. ETSI es, 202(050), p.v1.

Ezzaidi, H. and Rouat, J., 2004, May. Pitch and MFCC dependent GMM models for speaker identification systems. In Electrical and Computer Engineering, 2004. Canadian Conference on (Vol. 1, pp. 43-46). IEEE.

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

Furui, S., 2005. 50 years of progress in speech and speaker recognition. SPECOM 2005, Patras, pp.1-9.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G. and Pallett, D.S., 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon technical report n, 93.

Glembek, O., Burget, L., Matějka, P., Karafiát, M. and Kenny, P., 2011, May. Simplification and optimization of i-vector extraction. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (pp. 4516-4519). IEEE.

González-Rodríguez, J., Ortega-García, J., Martín, C. and Hernández, L., 1996, October. Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays. In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on (Vol. 3, pp. 1333-1336). IEEE.

Gruber, M.H., 1997. Statistical digital signal processing and modeling. Taylor & Francis Group.

Gustafsson, H., Nordholm, S.E. and Claesson, I., 2001. Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Transactions on Speech and Audio Processing, 9(8), pp.799-807.

Hasan, M.R., Jamil, M. and Rahman, M.G.R.M.S., 2004. Speaker identification using mel frequency cepstral coefficients. variations, 1(4).

Heck, L.P., Konig, Y., Sönmez, M.K. and Weintraub, M., 2000. Robustness to telephone handset distortion in speaker recognition by discriminative feature design. Speech Communication, 31(2), pp.181-192.

Hermansky, H. and Morgan, N., 1994. RASTA processing of speech. IEEE transactions on speech and audio processing, 2(4), pp.578-589.

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

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. and Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), pp.82-97.

Holmes, W., 2001. Speech synthesis and recognition. CRC press.

Hu, Y. and Loizou, P.C., 2003. A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Transactions on Speech and Audio Processing, 11(4), pp.334-341.

Hu, Y. and Loizou, P.C., 2004. Incorporating a psychoacoustical model in frequency domain speech enhancement. IEEE signal processing letters, 11(2), pp.270-273.

Jabloun, F. and Champagne, B., 2003. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. IEEE Transactions on Speech and Audio Processing, 11(6), pp.700-708.

Jayamaha, R.M.M., Senadheera, M.R., Gamage, T.N.C., Weerasekara, K.P.B., Dissanayaka, G.A. and Kodagoda, G.N., 2008, December. Voizlock-human voice authentication system using hidden markov model. In Information and Automation for Sustainability, 2008. ICIAFS 2008. 4th International Conference on (pp. 330-335). IEEE.

Jayanna, H.S. and Prasanna, S.M., 2009. Analysis, feature extraction, modeling and testing techniques for speaker recognition. IETE Technical Review, 26(3), pp.181-190.

Jeevakumar, K., 1993. Joint estimation of vocal tract and source parameters of a speech production model (Doctoral dissertation, Dublin City University).

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

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

Kheder, W.B., Matrouf, D., Bousquet, P.M., Bonastre, J.F. and Ajili, M., 2017. Fast i-vector denoising using MAP estimation and a noise distributions database for robust speaker recognition. Computer Speech & Language, 45, pp.104-122.

Kheder, W.B., Matrouf, D., Bousquet, P.M., Bonastre, J.F. and Ajili, M., 2014, October. Robust speaker recognition using map estimation of additive noise in i-vectors space. In International Conference on Statistical Language and Speech Processing (pp. 97-107). Springer, Cham.

Khoury, E., El Shafey, L. and Marcel, S., 2014. An open source toolbox for speaker recognition based on Bob. In Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (No. EPFL-CONF-196984).

Kim, H.G., Moreau, N. and Sikora, T., 2006. MPEG-7 audio and beyond: Audio content indexing and retrieval. John Wiley & Sons.

Kim, K. and Kim, M.Y., 2010. Robust speaker recognition against background noise in an enhanced multi-condition domain. IEEE Transactions on Consumer Electronics, 56(3).

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

Lei, Y., Scheffer, N., Ferrer, L. and McLaren, M., 2014, May. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on (pp. 1695-1699). IEEE.

Li, Q. and Huang, Y., 2010, March. Robust speaker identification using an auditory-based feature. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (pp. 4514-4517). IEEE.

Li, Q. and Huang, Y., 2011. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE transactions on audio, speech, and language processing, 19(6), pp.1791-1801.

Li, Q., 2009, October. An auditory-based transfrom for audio signal processing. In Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on (pp. 181-184). IEEE.

Lim, J. and Oppenheim, A., 1978. All-pole modeling of degraded speech. IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(3), pp.197-210.

Lim, J.S. and Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proceedings of the IEEE, 67(12), pp.1586-1604.

Lim, Shin-Cheol, et al. "Hard-mask missing feature theory for robust speaker recognition." IEEE Transactions on Consumer Electronics 57.3 (2011).

Lin, L. and Wang, S., 2006, May. A Kernel method for speaker recognition with little data. In Signal Processing, 2006 8th International Conference on (Vol. 1). IEEE.

Loizou, P.C., 2005. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. IEEE Transactions on Speech and Audio Processing, 13(5), pp.857-869.

Loizou, P.C., 2013. Speech enhancement: theory and practice. CRC press.

MacQueen, J., 1967, June. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Mandasari, M.I., McLaren, M. and van Leeuwen, D.A., 2012, March. The effect of noise on modern automatic speaker recognition systems. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (pp. 4249-4252). IEEE.

Martin, A. and Przybocki, M., 2000. The NIST 1999 speaker recognition evaluation—An overview. Digital signal processing, 10(1-3), pp.1-18.

Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M., 1997. The DET curve in assessment of detection task performance. NATIONAL INST OF STANDARDS AND TECHNOLOGY GAITHERSBURG MD.

Martin, A.F. and Greenberg, C.S., 2010. The NIST 2010 speaker recognition evaluation. In Eleventh Annual Conference of the International Speech Communication Association.

Mary, L. and Yegnanarayana, B., 2008. Extraction and representation of prosodic features for language and speaker recognition. Speech communication, 50(10), pp.782-796.

Matsui, T. and Furui, S., 1994. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's. IEEE transactions on speech and audio processing, 2(3), pp.456-459.

Matsui, T., Kanno, T. and Furui, S., 1996. Speaker recognition using HMM composition in noisy environments. Computer Speech & Language, 10(2), pp.107-116.

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

McLaren, M., Vogt, R. and Sridharan, S., 2007. SVM speaker verification using session variability modelling and GMM supervectors. Advances in Biometrics, pp.1077-1084.

Meriem, F., Farid, H., Messaoud, B. and Abderrahmene, A., 2017. New Front End Based on Multitaper and Gammatone Filters for Robust Speaker Verification. In Recent Advances in Electrical Engineering and Control Applications (pp. 344-354). Springer International Publishing.

Ming, J., Hazen, T.J. and Glass, J.R., 2006, June. A comparative study of methods for handheld speaker verification in realistic noisy conditions. In Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The (pp. 1-8). IEEE.

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

Moinuddin, M. and Kanthi, A.N., 2014. Speaker Identification based on GFCC using GMM.

Muda, L., Begam, M. and Elamvazuthi, I., 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083.

Nemati, S. and Basiri, M.E., 2011. Text-independent speaker verification using ant colony optimization-based selected features. Expert Systems with Applications, 38(1), pp.620-630.

NENGHENG, Z. 2005. Speaker Recognition using Complementary Information from Vocal Source and Vocal Tract. The Chinese University of Hong Kong.

NIST 1998. A Universal Transcription Format (UTF) Annotation Specification for Evaluation of Spoken Language Technology Corpora.

Ortega-García, J. and González-Rodríguez, J., 1996, October. Overview of speech enhancement techniques for automatic speaker recognition. In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on (Vol. 2, pp. 929-932). IEEE.

Pati, D. and Prasanna, S.M., 2010. Speaker recognition from excitation source perspective. IETE Technical Review, 27(2), pp.138-157.

Patterson, R.D., Holdsworth, J. and Allerhand, M., 1992. Auditory models as preprocessors for speech recognition. The Auditory Processing of Speech: from Auditory Periphery to Words, pp.67-89.

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

Pillay, S.G., 2011. Voice Biometrics Under Mismatched Noise Conditions (Doctoral dissertation).

Pinheiro, H.N., Ren, T.I., Cavalcanti, G.D., Jyh, T.I. and Sijbers, J., 2013, October. Type-2 fuzzy GMM-UBM for text-independent speaker verification. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on (pp. 4328-4331). IEEE.

Przybocki, M. and Martin, A., 2004. Speaker Recognition Evaluation Chronicles. In Proc. Odyssey 2004, The Speaker and Language Recognition Workshop.

Pullella, D., Kuhne, M. and Togneri, R., 2008, March. Robust speaker identification using combined feature selection and missing data recognition. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on (pp. 4833-4836). IEEE.

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

Rao, K.S. and Sarkar, S., 2014. Robust speaker recognition in noisy environments. Springer.

Revathi, A. and Venkataramani, Y., 2009, November. Source and system features for text independent speaker identification using iterative clustering approach. In Signal and Image Processing Applications (ICSIPA), 2009 IEEE International Conference on (pp. 1-5). IEEE.

Reynolds, D.A. and Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE transactions on speech and audio processing, 3(1), pp.72-83.

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

Reynolds, D.A., 1997, September. Comparison of background normalization methods for text-independent speaker verification. In Eurospeech.

Reynolds, D.A., Quatieri, T.F. and Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital signal processing, 10(1-3), pp.19-41.

Richardson, F., Reynolds, D. and Dehak, N., 2015. Deep neural network approaches to speaker and language recognition. IEEE Signal Processing Letters, 22(10), pp.1671-1675.

Sadjadi, S.O., Slaney, M. and Heck, L., 2013. Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research. Speech and Language Processing Technical Committee Newsletter, 1(4).

Scalart, P., 1996, May. Speech enhancement based on a priori signal to noise estimation. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on (Vol. 2, pp. 629-632). IEEE.

Seltzer, M.L., Raj, B. and Stern, R.M., 2004. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Communication, 43(4), pp.379-393.

Senoussaoui, M., Kenny, P., Dehak, N. and Dumouchel, P., 2010, June. An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech. In Odyssey (p. 6).

Shao, Y. and Wang, D., 2008, March. Robust speaker identification using auditory features and computational auditory scene analysis. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on (pp. 1589-1592). IEEE.

Shao, Y., Srinivasan, S. and Wang, D., 2007, April. Incorporating auditory feature uncertainties in robust speaker identification. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on (Vol. 4, pp. IV-277). IEEE.

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

Sinith, M.S., Salim, A., Sankar, K.G., Narayanan, K.S. and Soman, V., 2010, November. A novel method for text-independent speaker identification using mfcc and gmm. In Audio Language and Image Processing (ICALIP), 2010 International Conference on (pp. 292-296). IEEE.

Soong, F.K., Rosenberg, A.E., Juang, B.H. and Rabiner, L.R., 1987. Report: A vector quantization approach to speaker recognition. AT&T technical journal, 66(2), pp.14-26.

Stevens, K.N., Williams, C.E., Carbonell, J.R. and Woods, B., 1968. Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material. The Journal of the Acoustical Society of America, 44(6), pp.1596-1607.

Sukhwal, A. and Kumar, M., 2015, October. Comparative study of different classifiers based speaker recognition system using modified MFCC for noisy environment. In Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on(pp. 976-980). IEEE.

Tsoukalas, D.E., Mourjopoulos, J.N. and Kokkinakis, G., 1997. Speech enhancement based on audible noise suppression. IEEE Transactions on Speech and Audio Processing, 5(6), pp.497-514.

Tufekci, Z. and Gurbuz, S., 2005, March. Noise robust speaker verification using mel-frequency discrete wavelet coefficients and parallel model compensation. In Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on (Vol. 1, pp. I-657). IEEE.

Vizinho, A., Green, P.D., Cooke, M. and Josifovski, L., 1999, September. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study. In Eurospeech (Vol. 99, pp. 2407-2410).

Wahab, A., Ng, G.S. and Dickiyanto, R., 2005. Speaker authentication system using soft computing approaches. Neurocomputing, 68, pp.13-37.

Wan, V. and Renals, S., 2005. Speaker verification using sequence discriminant support vector machines. IEEE transactions on speech and audio processing, 13(2), pp.203-210.

Wang, C., Miao, Z. and Meng, X., 2008, May. Differential mfcc and vector quantization used for real-time speaker recognition system. In Image and Signal Processing, 2008. CISP'08. Congress on (Vol. 5, pp. 319-323). IEEE.

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

Wang, J.C., Wang, C.Y., Chin, Y.H., Liu, Y.T., Chen, E.T. and Chang, P.C., 2017. Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition. *Multimedia Tools and Applications*, *76*(3), pp.4055-4068.

Wang, N., Ching, P.C., Zheng, N. and Lee, T., 2011. Robust speaker recognition using denoised vocal source and vocal tract features. IEEE transactions on audio, speech, and language processing, 19(1), pp.196-205.

Wang, Y., Li, B., Jiang, X., Liu, F. & Wang, L. Speaker recognition based on dynamic MFCC parameters. Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on, 11-12 April 2009 2009. 406-409.

Weiss, M.R., Aschkenasy, E. and Parsons, T.W., 1975. Study and development of the INTEL technique for improving speech intelligibility. NICOLET SCIENTIFIC CORP NORTHVALE NJ.

Weng, Z., Li, L. and Guo, D., 2010, July. Speaker recognition using weighted dynamic MFCC based on GMM. In Anti-Counterfeiting Security and Identification in Communication (ASID), 2010 International Conference on (pp. 285-288). IEEE.

Wong, L.P. and Russell, M., 2001. Text-dependent speaker verification under noisy conditions using parallel model combination. In Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on (Vol. 1, pp. 457-460). IEEE.

Yoshida, K., Takagi, K. and Ozeki, K., 2004, October. Improved model training and automatic weight adjustment for multi-SNR multi-band speaker identification system. In INTERSPEECH.

You, C.H., Lee, K.A. and Li, H., 2009. An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. IEEE Signal processing letters, 16(1), pp.49-52.

Yujin, Y., Peihua, Z. and Qun, Z., 2010, October. Research of speaker recognition based on combination of LPCC and MFCC. In Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on (Vol. 3, pp. 765-767). IEEE.

Yutai, W., Bo, L., Xiaoqing, J., Feng, L. and Lihao, W., 2009, April. Speaker recognition based on dynamic MFCC parameters. In Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on (pp. 406-409). IEEE.

Zajíc, Z., Vaněk, J., Machlica, L. and Padrta, A., 2007. A cohort methods for score normalization in speaker verification system, acceleration of on-line cohort methods.

Zhao, X. and Wang, D., 2013, May. Analyzing noise robustness of MFCC and GFCC features in speaker identification. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7204-7208). IEEE.

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

Zhao, X., Wang, Y. and Wang, D., 2014. Robust speaker identification in noisy and reverberant conditions. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22(4), pp.836-845.

zohra Chelali, F., Djeradi, A. and Djeradi, R., 2011. Speaker identification system based on PLP coefficients and artificial neural network. environments, 1, p.2.