SPEAKER VERIFICATION USING SCORE LEVEL FUSION OF MFCC AND GFCC

A Dissertation submitted in fulfillment of the requirements for the Degree of

MASTER OF ENGINEERING

in

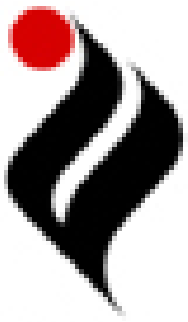
Electronic Instrumentation & Control Engineering

Submitted by

Preet Kiran Kaur
801451019

Under the Guidance of

Dr. Saurabh Bhardwaj
Assistant Professor, EIED



2016

Electrical and Instrumentation Engineering Department
Thapar University, Patiala
(Declared as Deemed-to-be-University u/s 3 of the UGC Act., 1956)
Post Bag No. 32, Patiala – 147004
Punjab (India)

DECLARATION

I hereby certify that the work which is presented in dissertation entitled, "**Speaker Verification Using Score Level Fusion of MFCC and GFCC**" in partial fulfilment of the requirements for the award of the degree of Master of Engineering in Electronics (Instrumentation & Control), submitted to Electrical & Instrumentation Engineering Department of Thapar University, Patiala is as authentic record of my own work carried under the supervision of **Dr. Saurabh Bhardwaj**, Assistant Professor, Electrical and Instrumentation Engineering Department, Thapar University, Patiala, Punjab. It refers others researcher's work which are duly listed in the reference section. The matter contained in this dissertation has not been submitted, neither in part nor in full to any other degree to any other university or institute except as reported in text and references.

Preetkiran

Place: Patiala

Date: 15-07-2016

(Preet Kiran Kaur)

Roll No. 801451019

It is certified that the above statement made by the student is correct to the best of my knowledge and belief.

(Dr. Saurabh Bhardwaj)
Assistant Professor
Electrical & Instrumentation Engineering Department
Thapar University, Patiala

Countersigned by:

(Dr. Ravinder Aggarwal)
Head
Electrical & Instrumentation Engineering Department
Thapar University, Patiala

(Dr. S.S Bhatia)
Dean (Academic Affairs)
Thapar University, Patiala

13/8/2016

DECLARATION

I hereby certify that the work which is presented in dissertation entitled, "**Speaker Verification Using Score Level Fusion of MFCC and GFCC**" in partial fulfilment of the requirements for the award of the degree of Master of Engineering in Electronics (Instrumentation & Control), submitted to Electrical & Instrumentation Engineering Department of Thapar University, Patiala is as authentic record of my own work carried under the supervision of **Dr. Saurabh Bhardwaj**, Assistant Professor, Electrical and Instrumentation Engineering Department, Thapar University, Patiala, Punjab. It refers others researcher's work which are duly listed in the reference section. The matter contained in this dissertation has not been submitted, neither in part nor in full to any other degree to any other university or institute except as reported in text and references.

Place: **(Preet Kiran Kaur)**

Date: **Roll No. 801451019**

It is certified that the above statement made by the student is correct to the best of my knowledge and belief.

(Dr. Saurabh Bhardwaj)
Assistant Professor
Electrical & Instrumentation Engineering Department
Thapar University, Patiala

Countersigned by:

(Dr. Ravinder Aggarwal)
Head
Electrical & Instrumentation Engineering Department
Thapar University, Patiala

(Dr. S.S Bhatia)
Dean (Academic Affairs)
Thapar University, Patiala

ACKNOWLEDGEMENT

First of all, I would like to express my deep gratitude towards my advisor and guide **Dr. Saurabh Bhardwaj, Assistant Professor**, Electrical and Instrumentation Engineering Department (EIED), Thapar University, Patiala for his guidance and support. His dedication and keen interest above all his overwhelming attitude to help his students had been solely and mainly responsible for completing my thesis work. His timely advice, meticulous scrutiny, crystal clear concepts and depths of knowledge has helped me to a very great extent to accomplish this work. I consider it a great opportunity to work with such an influential personality.

I am extremely thankful to our **Head of the Department, Dr. Ravinder Agarwal** as well as **PG Coordinator, Mr. Nirbhowjap Singh, Assistant Professor**, Electrical and Instrumentation Engineering Department. I would also like to thank my friends who devoted their valuable time and helped me in all possible ways towards successful completion of this work. I thank all those who have contributed directly or indirectly to this work. Lastly, I would like to thank my parents and brother for their years of unyielding love and encouragement. They have always wanted the best for me and I admire their determination and sacrifice.

Date

Place: Patiala

Preet Kiran Kaur

M.E.(EICE) 2nd year

801451019

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASR – Automatic Speaker Recognition

LPCC- Linear Predictive Cepstral Coefficients

HMM-Hidden Markov Model

VQ-Vector Quantization

GMM-Gaussian Mixture Model

UBM-universal Background Model

GMMIN- Gaussian Mixture Model Integrated

EM- Expectation Maximization

MFCC-Mel Frequency Cepstral Coefficient

LFCC-Linear Frequency Cepstral Coefficient

PLP- Perceptual Linear Prediction

GF- Gammatone Features

CASA- Computational Auditory Scene Analysis

SID- Speaker Identification

FAR- False Acceptance Rate

FRR- False Rejection Rate

GAR- Genuine Acceptance Rate

EER- Equal Error Rate

ROC- Receiver Operating Curve

FFT- Fast Fourier Transform

DCT- Discrete Cosine Transform

GFCC- Gammatone Frequency Cepstral Coefficients

ERB- Equivalent Rectangular Bandwidth

MAP- Maximum A-Posteriori

KNN- k nearest neighbours

ABSTRACT

The feature analysis component of an Automated Speaker Verification (ASV) system plays a crucial role in the overall performance of the system. There are many feature extraction techniques available, but ultimately we want to maximize the performance of these systems. Current state-of-the-art ASV system performs quite well in a controlled environment where the speech signal is noise free. The objective of this thesis investigates the results that can be obtained when you combine Mel-Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC) at score level.

The MFCC and GFCC feature components combined are suggested to improve the reliability of a speaker verification system. The MFCC are typically the "*de facto"* standard for speaker recognition systems because of their high accuracy and low complexity; however they are not very robust at the presence of additive noise. The GFCC features in recent studies have shown very good robustness against noise and acoustic change. The main idea is to integrate MFCC & GFCC features to improve the overall Speaker Verification system performance in low signal to noise ratio (SNR) conditions. Also, we employed a simpler Gaussian membership function (GMF) based matching process. Last, we use Gaussian Mixture Model (GMM) and K-nearest neighbour (KNN) to measure the similarity in the verification stage. The experiments are conducted on the text independent VoxForge database, where the test utterances are mixed with noises at various SNR levels (-5, 5, 10 and 20 dB). Experimental results verify the validity of our proposed approaches in personal authentication.

CHAPTER 1

INTRODUCTION

1.1 Overview

In biometric recognition systems various signals and measurements have been projected. Among them fingerprint, voice and face are the most popular, each of them have pros and cons depending on their accuracy and deployment. Voice outperforms all other biometrics because of two main factors. Firstly, speech is a natural and universal way of communication therefore providing speech sample is not considered as an officious or intrusive step. Secondly, for taking the sample no high cost or special device is needed, readily available devices such as sound card and microphone can be used.

Humans could distinguish the speakers only by hearing the voice of the speaker. This is because the human auditory system extracts the necessary information from the voice of the speaker in a number of environmental conditions. So a speaker recognition system should be developed in a way that it is able to verify the speaker from its speech in a meaningful way. In the present work, features extraction from the voice signal is done by MFCC technique which is the most popular and conventional method. These spectral based features are extracted by applying Fast Fourier Transform, and then converted into robust, de-correlated, flexible and compressed form of cepstral coefficients. MFCC outperforms most of the feature extraction techniques only under clean conditions but do not perform well under noisy and mismatched conditions. Also it is very difficult to remove the noise from those portions where it overlies signal spectrum. All the frequency bands of speech are corrupted by the noise because discrete cosine transform is applied over all the frequency bands and the coefficients thus derived are also corrupted.

Practically the speaker verification system comes across various challenges such as additive noise, channel variations and room reverberation. Therefore the robustness of speaker verification system is a critical need for real world applications. Also, the ability of the humans to recognise the speaker in noisy environmental conditions has motivated the development of robust speaker recognition system from the new perspective i.e. computational auditory scene analysis. GFCC are extracted from gammtone filter bank based on cochlear filtering and are robust in noisy and mismatched conditions due to the non linear frequency distribution characteristics.

Unimodal biometric system have various problems like spoof attacks, noisy sensor data, unacceptable error rates and non-universality of chosen trait. Therefore, a multimodal system is

formed by combining two systems to overcome their problems. In the present work, a robust multi-model system is developed by fusing both MFCC and GFCC based models to get the advantages of both models. The two systems are combined using score level fusion technique.

1.2 Human Auditory Perception

Humans could recognize and distinguish a speaker by his/her voice, it's because human auditory system extracts the necessary information which differentiate the speakers even in a number of environmental conditions. There are three major areas loudness, pitch and timbre based on which human auditory system perceives and categorize the audio signals. As a person grows older their auditory system also goes through certain changes. It is very well known that young people's ability to hear high pitched sounds is very much better, but with every 10 years the frequency range of hearing decreases.

1.2.1 Pitch

Here, Pitch is defined as a perceived quantity which relates to the fundamental frequency of vibration produced by the vocal cords with respect to time. Typically it is the degree of highness or lowness of a tone. The value of the pitch is perceived value, not a particular value is recognised by humans. The frequency of vibration generated by the vocal cords is a function of its shape, tension and characteristics of air flow and the configuration of vocal tract at the time of utterance. Pitch is actually an average value [1]. Since, pitch is very immanent across different speakers and is dependent on one's age, therefore it is very hard to analyze. Pitch level and its variations can represent different areas and also can be segmented.

Generally, one do not utilise the whole pitch range while carrying out the conversation, and also understands that by changing the impression of voice he/she can express emotions (such as excitement or sadness) and subtle variations. For example, we may relate it to a situation when the message sent by us conveyed a completely different message to another person, as it contained paralinguistic qualities (such as pitch variations, sonority variations and facial expressions etc) which were hard to include in a text message.When a person tries to distinguish the speakers on the bases of the vocal characteristics, then he/she tries to realize the pitch of the individual for reference. The first thing the imposter do to mimic someone's vocal characteristics is the modification of the pitch level to match the pitch of the target speaker. Therefore as a speaker recognition feature, it is not reliable to distinguish a speaker on the basis of the average value of pitch. Moreover, the cepstral features the most popular ones in speaker recognition, do not preserve much of the tonal information of the audio.

The figures below show the paralinguistic variation of pitch by the utterance of verb [try] with 2 – different ways of expressions. This pitch variation is of great importance for speaker recognition as these variations contains a lot of speaker dependent content of expressing.

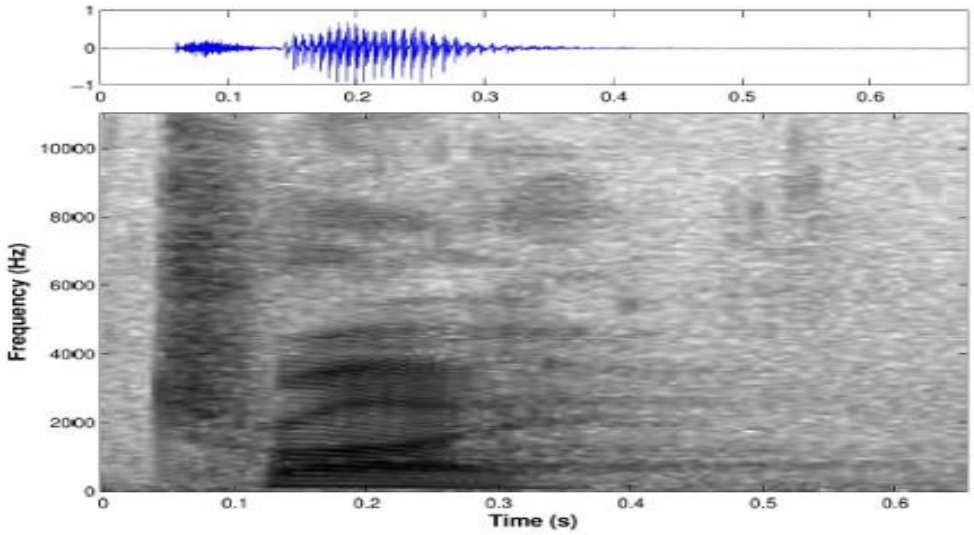


Fig 1.1 Using [try] as short and assertive expression [1]

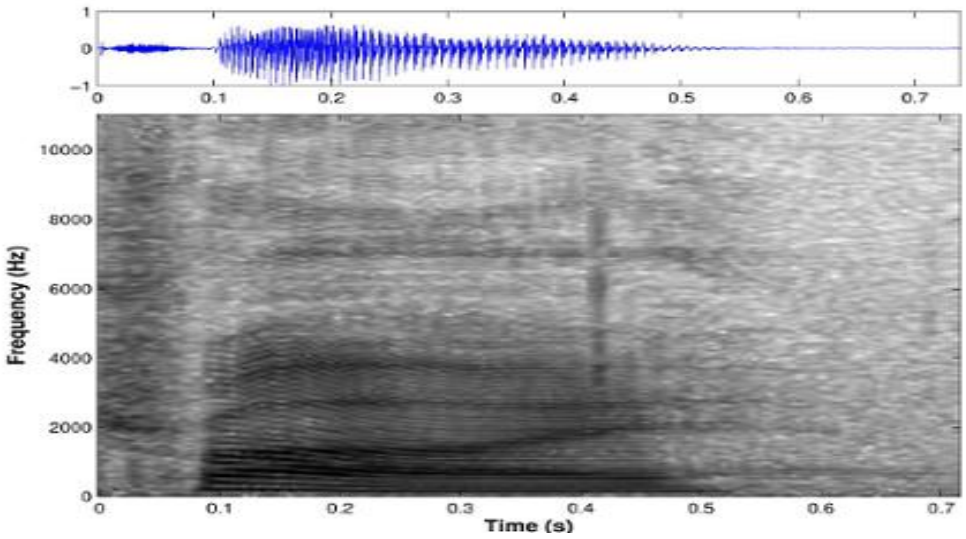


Fig 1.2 Using [try] as assertive with a strong interrogative expression (longer with a higher pitch level)

1.2.2 Loudness

Loudness is another perceived quantity which is related to the intensity of vibrations produced by the vocal cord with respect to time at a particular value of pitch. The perceived loudness also depends on the frequency, which implies its dependence on pitch as well [1]. It can also be defined as a psychological term which describes the measure of an auditory sensation and is dependent on the intensity and frequency of sound. It also depends on psychological factors such

as fatigue, alertness and attention. The loudness decreases from the beginning to the end of an utterance, like pitch.

For the quantitative analysis of loudness, let us define the term intensity. Intensity is a power measure of a sound wave and is equal to power per unit area (units= $\frac{W}{m^2}$). Also the sensitivity of human ear differs with the frequency. The relationship of intensity with the pressure differential is given as:

$$I = \frac{p^2}{£} \tag{1.1}$$

, here £ is the specific acoustic impedance of sound and p is the differential pressure. The value of intensity at 1000Hz frequency is sufficient for the human ear for hearing any tone.

1.2.3 Timbre

Timbre defines the quality and character of musical sounds which is different from their pitch and intensity. It is related to the harmonic content and the dynamic characteristics of the audio. The harmonic content in humans is associated with the location of the formants and its characteristics. The complex modulations (the dynamic characteristics) could be the amplitude or frequency modulations of the signal. Timbre reflects the speaker characteristics so it's a best feature among the three for speaker recognition. So while performing speaker recognition it's important to distinguish the timbres of different speakers.

1.3 Speaker Recognition

If given a voice sample, the process of speaker recognition is the identification of the source of speech based on the clues or information extracted from the speech sample. It can also be called as Voice recognition [2].The speaker is recognized due to the distinct qualities of the speaker voices, which needs to be preserved in the speech signal. In general the ASR system aims to model the vocal tract characteristics; the model is a statistical model or mathematical model.

The task of Speaker Recognition is classified into speaker verification and identification, as shown in Fig [1.3].

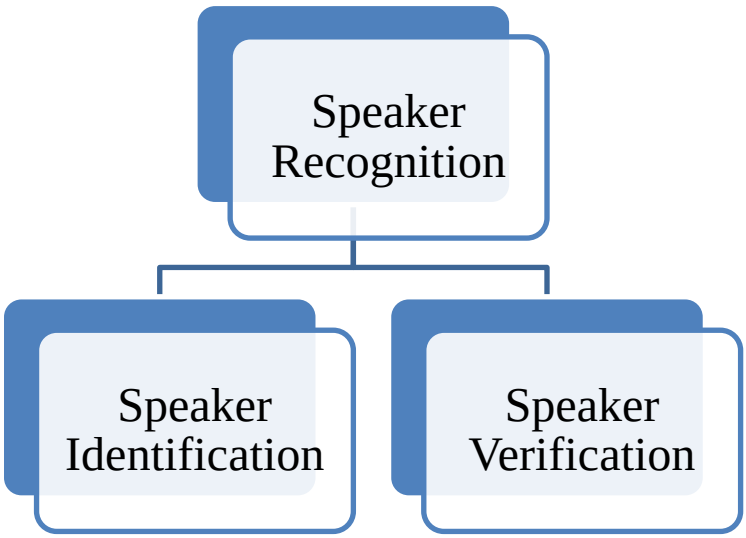


Figure 1.3 Types of Speaker Recognition

Speaker Recognition is different from speech recognition but these usually create confusion because both of them use speech signals. In speaker recognition the goal is to find 'who' is speaking where as in speech recognition the main concern is to know 'what' was spoken i.e. the speech contents [2].

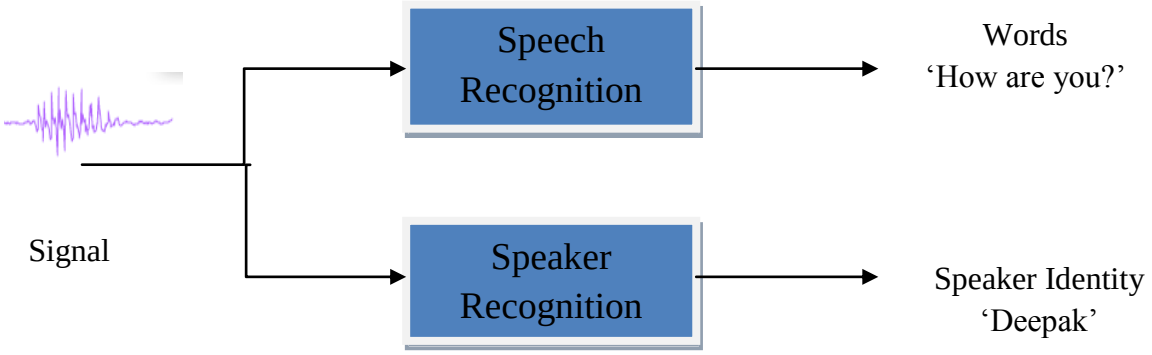


Fig 1.4 Speech Vs Speaker Recognition

1.3.1 Types of Speaker Recognition

Speaker Identification

Speaker Identification is a task to determine who is speaking from the set of registered speakers. This process is similar to the process when the police officer compares the sketch of a suspect with the photos of the criminals in record to find the best match.

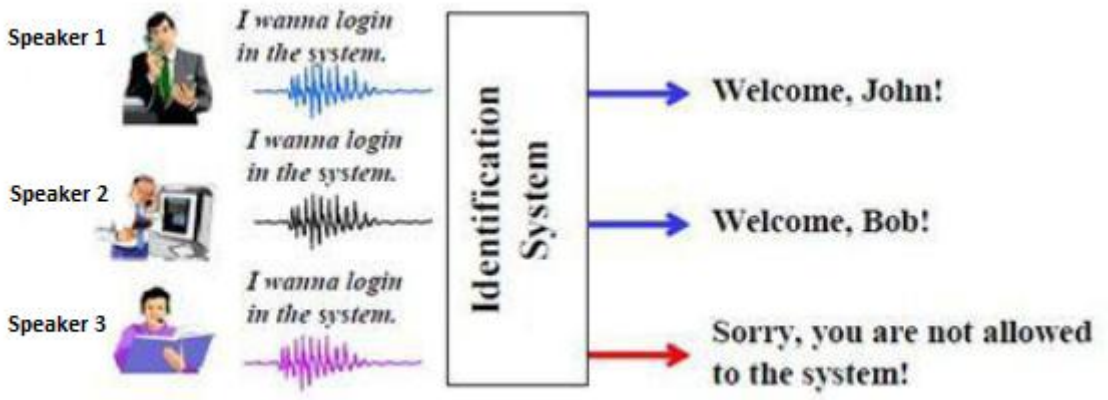


Fig 1.5 Example of Speaker Identification

Speaker identification is further divided into: closed-set and open-set. In closed-set identification there is 1:N match, the test speaker's voice sample is compared with all other N templates and the system comes up with the template that best matches the speech sample [3]. There is no rejection scheme in this process as there is always a speaker in the database that is closest to the test speaker sample. In a practical case, where the test sample is the voice of a 6-year old child and the speaker model database consists of all adults, still there will be a model from the adults that best matches the child sample. Closed-set identification has its main application in departmental organisation where there is a known number of group members, their speaker models are stored in a database and thus the identification process is internal as there is no outsider. Open set identification occurs in two steps. In first step closed set identification is done and the resulting identity obtained from the first step is used further. The second step is speaker verification where the test sample is again verified with the target speaker (from the first step), and thus the ID is accepted or rejected. Although, open-set identification is complex but it gives true results whether a person belongs to the group (set of known individuals in the database) or do not [4].

Speaker Verification

Speaker Verification is the task of confirming a speaker's claimed identity using the features extracted from their speech sample. This process is also called Speaker Authentication. The ID provided by the speaker is used to retrieve its model from the database, which is called as the target speaker model. Futher the comparison is made between the speech sample of the speaker who provided the ID and the target speaker model, then the decision is given in the form of accepted or rejected. It is a 1:1 matching process in which one users' speech is matched to only one template or claimed speaker model, so even if the population grows, the computation

requirement for recognition remains constant. For example, presenting passport at border control where the agent compares the picture in the passport with your face is a verification process [1].

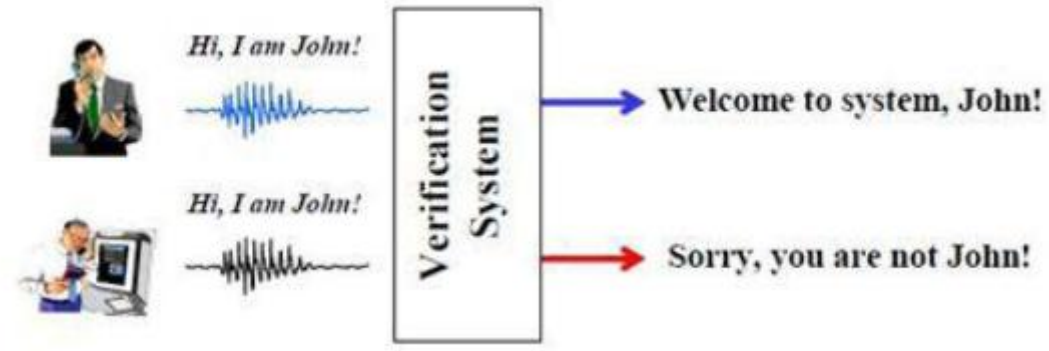


Fig 1.6 Example of Speaker Verification

1.3.2 Text Dependent Vs Text Independent

The process of speaker recognition can happen as: Text-dependent or Text independent.

In Text-dependent system the text used by the speaker is same during the training and testing phase. The prompts used are known to system, it is in the form of password, pins or a fixed phrase which could be same for all speakers or unique. To recognize a speaker in this system, the speaker characteristics and the prompted phrase both have to match. This improves the system performance by reducing the imposter risk using voice recordings. Text independent recognition makes sense only for speaker verification.

In Text-independent system there are no restrictions on the content of speech. The speech or text during the training and testing phase differ and in fact in case of forensics, the enrolment can happen without the user's awareness [1]. The text independent system relies only on speaker characteristics. It is most versatile and can be used in speaker verification and identification both. This system is commercially very attractive and convenient to users, as the users are free to speak anything. The main problem is that it requires longer enrolment of test utterances to achieve better performance.

1.4 Outline of the Thesis

Chapter 1 provides an overview of the thesis work. It gives a general framework of the speaker recognition system and also give the idea about the speaker specific characteristics which helps to distinguish speakers.

Chapter 2 describes the previous contributions and researches done by various researchers in the field of speaker verification.

Chapter 3 contains the layout of the Automatic Speaker Verification (ASV) system, which is the base of the work done. In this chapter the structure, the decision method and the performance measures of verification system are described.

Chapter 4 gives the detailed explanation of each step of the methodology. The details of the procedure of the feature extraction techniques: MFCC and GFCC are given, also the two classifier's (KNN and GMM-UBM) training and testing process is described. This chapter also includes the problem statement and the solution proposed to overcome those problems.

Chapter 5 contains the results obtained from each of the system described in methodology and comparisons between different methods are shown in form of plots and tables. Also the database used and the experimental setup is described in the starting of the chapter.

Chapter 6 concludes our study and project work. The future scope and the improvements that can be made in our study are also mentioned.

CHAPTER 2

LITERATURE REVIEW

Several advancements occurred in the Automatic Speaker Recognition (ASR) system over the last few decades. The development of the system for speaker recognition was first attempted in 1960's. S. Pruzansky, [5] performed the similarity measurement using filter bank array at Bell Labs. To determine the similarity between the two speech samples, their spectrograms were cross correlated. S. Pruzansky et.al. [6], in this paper improved the similarity measurement technique by the analysis of variances of some feature set. K.P. Li et.al [7], used adaptive system to distinguish the speakers. Improvements were achieved significantly by implementing linear discrimination analysis and it gave 90% accuracy. B.S. Atal et.al [8], tried using a new technique: formant analysis to make some advancements in speaker recognition system. Also they tried modelling the vocal tract by the use of mathematical filter.

G. R. Doddington [9], in his work used pitch energy, speech energy and the formant frequencies along with the pre-existing filter bank techniques in the verification system. The results so obtained have decreased the error rates of verification system by four times. W. Haberman et.al [10] introduced the first wholly automated speaker verification system. The whole system was built in texas instruments and was tested and procured by the MITRE Corporation and U.S. Air force. The system was tested several number of times over 200 speaker samples. For the spectral analysis digital filter bank was used and also a prompter was used to select the 4 word phrase randomly.

S. Furui [11] introduced a new technique for speaker verification. Cepstral coefficients are extracted using time frames and are combined with its 1^{st} and 2^{nd} polynomial to remove the distortions due to transmission systems. The final decision of accepting or rejecting the claimed identity was made using the overall distance rule. The results obtained has one percent error rate. J. M. Naik et.al [12] developed a technique for speaker verification over large telephone lines. Comparison is made between a new technique, template based matching and HMM. The speaker models used in this system came from multi-word sentences and results shows that the new technique improved the discrimination of true speaker from imposter. A. E. Rosenberg et.al [13] presented and evaluated a VQ-based approach for speaker recognition system. The LPCC feature vectors for each speaker are clustered using Vector quantization in VQ codebooks. Thus the feature data set is represented using VQ codebook which is a small set of numbers. Although this

technique is intrinsically used for text independent speech but could be extended to text dependent speech system.

S. Furui [14] in this survey told that in 1990s the focus was on increasing the robustness and to present speaker recognition as a workable or effective biometric technology. Comparable to VQ-based method, HMM method was found be more accurate when a large dataset is available. Tomoko Matsui et.al [15] proposed a method based on VQ in which speaker recognition is done using pitch and vocal tract information. This method combined cepstral coefficients and delta cepstral coefficients with the delta pitch frequencies and also two different codebooks: voiced as well as unvoiced codebooks are used for each speaker, giving more accurate results. To find the distance between the test vectors and VQ codebook a new Distortion-intersection Measure (DIM) was proposed.

The input speech signal for Automatic Speaker Recognition (ASR) mostly contains noise which further leads to signal distortion and loss of information. D. A. Reynolds [16] introduced a new speaker model: Gaussian mixture speaker model that characterizes speakers' voice by statistical modelling of the speech sound and gives high performance even for short test utterances. R. C. Rose et al. [17], [18] extended the use of probabilistic Gaussian mixture model classifier to noisy signals. They combined two separate models: speech model and background noise model in a statistical manner for a robust speaker identification system. Thus a scoring procedure to compute the likelihood of the speech corrupted with non Gaussian signals is developed. Two noise compensation techniques are used and compared: first, direct integration of the noise background model into the speech model classifier and second the use of noise pre-processing techniques before applying the classifier. The results show that the GMMIN model performs better than the GMM model.

D. A. Reynolds et al. [19], presented a procedure based on maximum likelihood for estimation of speech signal model parameters using Expectation-Maximization algorithm. They dealt with acoustic noise effects in the signal processing and developed robust statistical models. A mechanism is described for the integration of acoustic background model with the underlying signal model for noise compensations. The results are evaluated on text independent speech signals in noisy environment and the improvement in the performance is obtained with Gaussian mixture integrated background process.

The performance of recognition system is affected due to noise, channel variability and nonlinear distortions in the data. D.A. Reynolds [20], applied several features and channel compensation method to make the speaker recognition system more robust. All the other steps, processing and

classification are kept constant. Results shows that the difference in the performance between the features is less and most of the improvement in performance is due to the compensation techniques applied in conjunction with the features. He [21] evaluated GMM model on conversational and telephonic speech. Robustness techniques: frequency warping, long term mean removal and difference coefficients are used for the compensation of spectral variability introduced by the device channel.

D. A. Reynolds [22], presented GMM based robust verification and identification system for text-independent speech. The results evaluation is done using four different databases YOHO, TIMIT, Switchboard and NTIMIT having different levels of noise, variability and speech quality. The identification is based on maximum likelihood classifier and verification system is based on likelihood ratio hypothesis which also include the background speaker normalization. A new approach for the selection of background model is also proposed. The results show that GMM provide inexpensive and high recognition accuracy but the performance is affected due to transmission degradation such as noise and variability. So for improvement of performance under these conditions robustness techniques should be on classifier and front end analysis.

L. F. Lamel et al. [23], did the experimental study of the performance of speaker verification system based on the telephone speech. The database is in French consisted of 100 target speakers and 1000 imposters. The degradation in performance of the system due to model aging is neutralized by the use of adaptation techniques. The performance is assessed considering some factors: speaker model type, training data amount and recency, the speech style and the linguistic content. The evaluation is done using text-dependent and text-independent modes and Phone based model is compared with GMM. The minimum equal error rate came out to be 1% for text dependent signal allowing two trials per attempt and 1.5s of speech signal per trial.

D. A. Reynolds et al. [24], presented a GMM-UBM verification system fully based on the finest likelihood ratio test , for likelihood function simple and effective Gaussian Mixture Model (GMM) , for the representation of the competing alternative speakers a universal background model (UBM) , and to derive the hypothesized speaker models Bayesian adaptation technique is used. The performance of the verification system is highly improved by the use of handset detector and score normalization.

Claude Barras et al. [25], applied cepstral features and Gaussian mixture model (GMM) along with the feature and score normalization on the cellular data for speaker verification. Variance normalization, cepstral mean Subtraction (CMS), Z-norm, T-norm, feature warping and the cohort method were some of the normalization techniques used for evaluation. The NIST 2002

database was used for experimentation and the best results were seen by the use of the combination of feature warping and T-norm.

GMM is widely used for speaker recognition but these systems do not give good results in noisy and mismatched conditions. Yang Shao et al. [26], presented a robust recognition system in the existence of additive noise. The Gammatone features (GF) are extracted using Gammatone filters. The Gammatone filters are based on human cochlear filtering and consist of a bank of overlying bandpass filters [27] (by R. D. Patterson). These Gammatone features (GF) are used to derive Gammatone frequency cepstral coefficients (GFCC). Additionally, for the estimation of the auditory feature uncertainties, a binary T-F mask is used and likelihood scores are calculated using uncertainty decoding. The Computational Auditory Scene Analysis (CASA) generated binary T-F mask. This approach attained effective performance improvement over conventional recognition systems under noisy conditions.

Yang Shao et al. [28], further studied the auditory features based on CASA fpr improving the robustness of the speaker identification system. The auditory features are extracted by changing the feature dimensions and in addition with the static features dynamic coefficients are incorporated. The novel auditory feature based Speaker identification system (SID) is evaluated under five dissimilar noisy conditions. The 30-dimension GFCC with its delta coefficients gave more accuracy over others.

L. Hong et al. [29], reported that Unimodal systems which make use of single characteristic for identification are suffering from several problems: spoof attacks, noisy sensor data, error rate unacceptability, non universality of the chosen trait. By using the Multimodal biometric system which combine the identities or evidences from multiple source, some of the problems of the unimodal biometric system can be demolished.

A. K. Jain et al. [30], proposed the use of the soft biometrics such as age, gender, weight and height in combination with the primary biometric system. Only the use of soft biometric characteristics for identification is not reliable but by integrating it with traditional identifiers the accuracy of recognition highly increases. M. C. Cheung et al [31], presented the audio-visual fusion strategy for the ASR systems, in which speech information and image information is combined. Here, Lip movement is the visual information captured which is not affected by the background noise and on the other hand the lightening conditions do not affect speech information, thus the combination of two improves the reliability of the system.

Karthik Nandakumar et al. [32], proposed a process of combination of likelihood ratio test based match scores to form a multi-modal system. Likelihood ratio scores using GMM are calculated on the databases of face, fingerprint, speech and iris. Likelihood scores of face and fingerprint, fingerprint and iris, face and speech are fused and achieved high performance, giving optimal results as compared to other score fusion techniques. Sandipan Chakroborty et al. [33], presented the likelihood ratio score fusion of MFCC and IMFCC. The paper proposed the use of Gaussian filter for extracting MFCC and IMFCC instead of triangular filters and GMM is used as a classifier to obtain the likelihood ratio scores. The improvement in performance accuracy is achieved because of the combination and use of Gaussian shaped filters.

CHAPTER 3

AUTOMATIC SPEAKER VERIFICATION SYSTEM

3.1 Structure of Speaker Verification

The thesis work is based on the speaker verification system which is a subpart of speaker recognition as studied in previous chapter. Speaker verification system consists of two phases: training phase and testing phase [34].

3.1.1 Training Phase (Enrolment)

The system is developed or allowed to learn to recognize an individual in training phase. It is trained using a large set of speech samples from different speakers. The block diagram of the training phase of speaker verification systems is shown Fig [3.1].

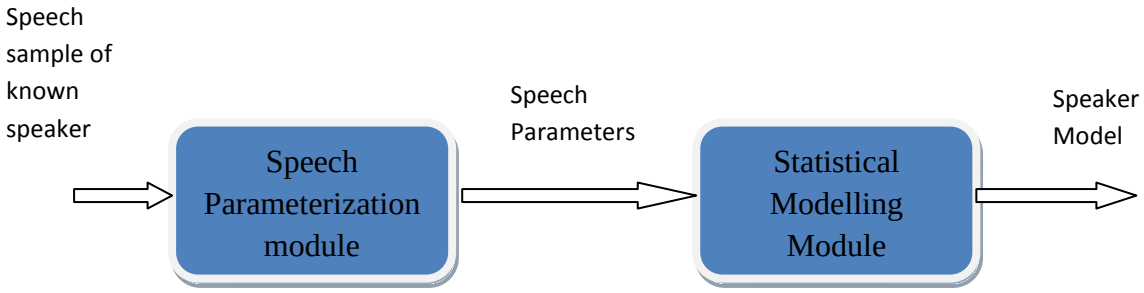


Fig 3.1: Block Diagram of the Training phase

Speech Parameterization Block: In this block the speech sample (continuous signal) is transformed into a set of feature vectors. The vectors thus obtained contain the information of the speaker in a compact and less redundant form, thus suitable for statistical modelling in next module.

Statistical Modelling Module: The feature vectors extracted above are used to create speaker models. Many modelling techniques are available and are used in speaker verification system. The selection of modelling technique depends largely on the speech type, the ease of enrolling and testing, storage and calculation considerations and the performance. Each speaker in the database is converted into a speaker model.

3.1.2 Testing Phase

In the testing phase we compare the test sample with the training sample (the sample claimed by the unknown speaker) and the aim is to verify if the test sample belongs to the claimed identity. The modular representation of testing phase of speaker verification system is shown in Fig [3.2].

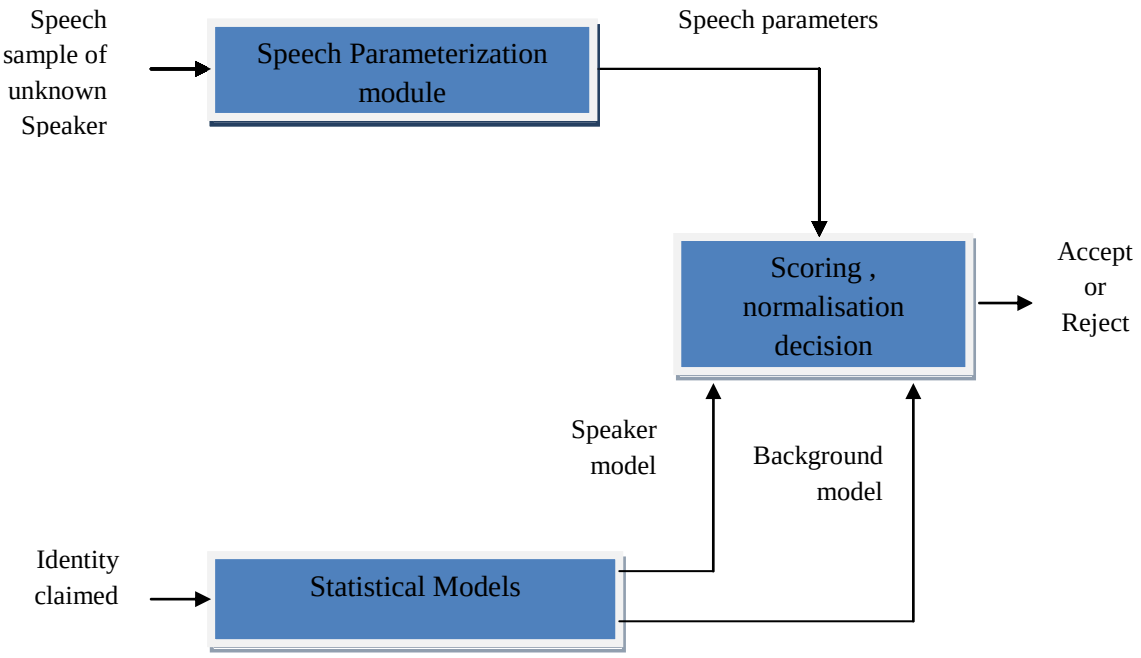


Fig 3.2 Block Diagram of Testing Phase

The input to the system is the speech sample from an unknown speaker and the identity claimed by the unknown speaker. Firstly, the features are extracted from the input speech sample using the same parameterization method as in the training phase. Then from the set of statistical models which were calculated during enrollment, the speaker model of the claimed identity and the background model is extracted [34]. The background model may consist of all other speakers except claimed speaker model or it's a single model developed using the speech samples from large population of speakers likely during recognition. Finally in the last module the input speech sample is compared with the speaker model of the claimed identity and also with the background model, and then the ratio of these two comparisons is taken and equated to a threshold. The decision is thus given as true or accepted when the result is above threshold and false or rejected when the result is below threshold.

3.2 Likelihood Ratio Detector

The modelling and classification in speaker verification system is wholly based on likelihood ratio detector. When given a speech segment as input, Z and claimed speaker or a hypothesized

speaker ‘S’ , the work of the speaker verification system is to determine if the speech ‘Z’ belongs to ‘S’ [24]. This task can also be termed as single speaker detection as Z is a speech from single speaker. If we do not have any prior information that Z belongs to a single speaker, then it becomes a multi-speaker detection task, but that is not a case in speaker verification system.

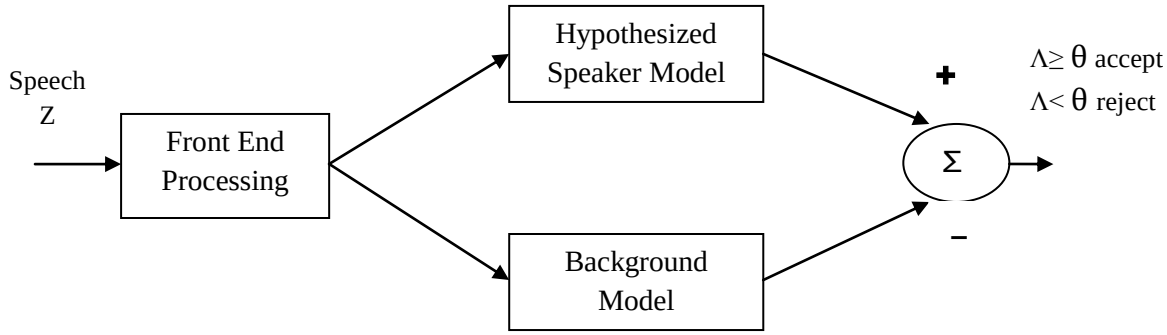


Fig 3.3 Likelihood Ratio Detector based Speaker Verification System

This task of detecting the speaker Z can be defined by the hypothesis test between:

H_1: Z belongs to a hypothesized speaker S

H_2 : Z does not belong to hypothesized speaker

The likelihood of the hypotheses H_i , i=1,2 evaluated over the speech segment Z is given by the probability density function $p(Z/ H_i)$. The likelihood ratio (LR) test to determine the decision between the two hypotheses is given by:

$$\frac{p(\frac{Z}{H_1})}{p(\frac{Z}{H_2})} \quad \begin{cases} \geq \theta & \text{accept } H_1 \\ < \theta & reject\ H_1 \end{cases} \tag{3.1}$$

Here, θ is the decision threshold to accept or reject the hypotheses H_1. The main goal in designing the speaker verification system is to specify techniques to calculate the values of the likelihoods $p(\frac{Z}{H_1})$ and $p(\frac{Z}{H_2})$ .The Fig [3.3] shows the basic blocks found in likelihood ratio based speaker verification system In front end processing the features containing speaker information are extracted from a sample of speech. The output of this block is in the form of feature vectors ‘X’ representing the test sample. These feature vector extracted is then used to calculate the likelihood of H_1, H_2 . For the mathematical calculations, γ_{hyp} is used to denote the modeller representation of H_1 that characterizes the claimed speaker S and $\gamma_{\overline{hyp}}$ is used to denote the model of the alternative hypotheses H_2 . The likelihood ratio is thus represented

by $p(X/\gamma_{hyp})/p(X/\gamma_{\overline{hyp}})$. Taking the logarithm on both sides gives the log-likelihood ratio and thus represented as [24]:

$$\Lambda(X) = \log p(X/\gamma_{hyp}) - \log p(X/\gamma_{\overline{hyp}}) \tag{3.2}$$

The model γ_{hyp} represents the speaker claimed S and thus can be extracted from the statistical models of the speakers, whereas the alternative model $\gamma_{\overline{hyp}}$ is not defined and represents all other speakers. In general if N speaker models $\{\gamma_1, \gamma_2 \dots\dots, \gamma_N\}$ are given, then the alternative model is given by:

$$p(X/\gamma_{hyp}) = \mathcal{F}(p(X/\gamma_1), p(X/\gamma_2), \dots\dots\dots\dots p(X/\gamma_N) \tag{3.3}$$

Here, $\mathcal{F}(\ )$ is some maximum or average function. There are two major approaches for alternative hypotheses modelling: cohort model and universal background model. In cohort model the members are the speakers who have the similar sound as the target speaker. In this approach the input speech signal is compared to the target speaker (claimed identity) and the cohort, the main benefit is that it does not involve the large population size [1].

The second approach, universal background model is based on the database from a large sized population. Thus a single model is trained γ_{bkg} using the samples of speech from number of speakers. The main subject of concern in this approach is the selection of speakers, the size and the composition. Also it is possible to train multiple background models, using different background model for different set of speakers. But a single background model is more advantageous as it can be used for every hypothesized speaker in the system task.

3.3 Performance Measures

The measures of accuracy in a speaker verification system are False Acceptance Rate (FAR) and False Rejection Rate (FRR) [4]. Before defining the performance measure, we must know about the match scores obtained in the verification system. Genuine or authentic score which is the measure of similarity between two samples belonging to same speaker. Imposter score is the measure of similarity between two samples belonging to different speakers. The decision in verification system is given by comparing the scores with a threshold value η. In theory, we know that highly similar templates has higher score value therefore the genuine score value should always be higher than the imposter score value. But in practical, some of the imposter score values are higher than genuine scores and vice versa, which makes it difficult to choose an optimal threshold value.

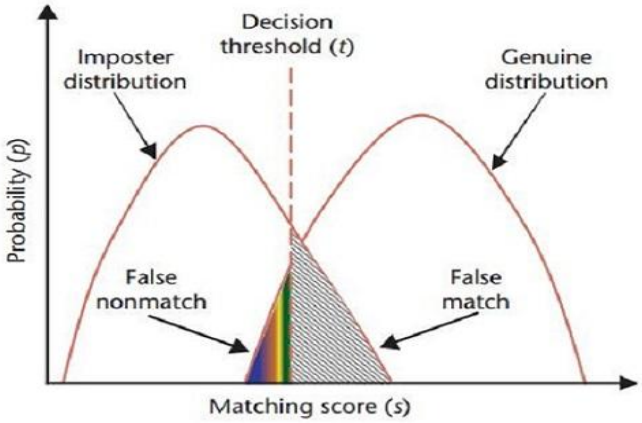


Fig 3.4 Genuine and Imposter Match score distribution

False Acceptance Rate (FAR): It is the probability of some users claiming a false identity is accepted or verified as true by the verification system. It is the proportion of imposter scores equal or greater than the threshold value μ. It is calculated using the formula:

$$FAR(\mu) = p(s \geq \mu/w_1) = \int_{\mu}^{\infty} p(s/w_1)\, ds \tag{3.4}$$

, w_1 here denotes the imposter class and $p(s/w_1)$ represents the probability density function of the imposter scores. FAR is also termed as False Match Rate (FMR).

False Rejection Rate (FRR): It is the probability of some users claiming his/ her true identity is rejected by the system. It can be defined as the fraction of genuine scores values lower than the threshold value μ. It is calculated using the formula:

$$FRR(\mu) = p(s < \mu/w_2) = \int_{-\infty}^{\mu} p(s/w_2)\, ds \tag{3.5}$$

, w_1 here denotes the genuine class and $p(s/w_2)$ represents the probability density function of the genuine scores. FRR is also termed as False Non Match Rate (FNMR).

The FAR and FRR are threshold dependent, and it can be seen from the Fig 3.5 that if we increase the threshold value then FRR increases where as FAR decreases and vice versa. When we plot FAR and FRR against threshold it intersects at some point because of the overlapping of scores, the value at that point of intersection is called EER. Also the values of the two rates FAR and FRR are equal at this point. The EER is shown in Fig [3.5].

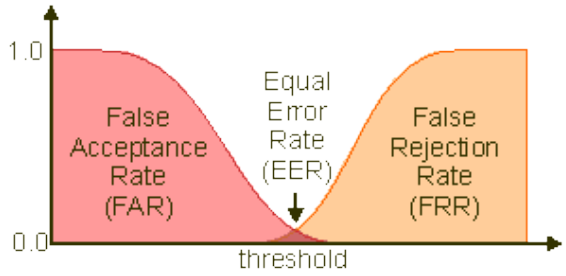


Fig 3.5 EER value

EER is also an important performance measure of the verification system which is independent of the threshold value. The lower EER value implies the system is more accurate.

ROC curves: These curves provide the complete specifications about the performance of the verification system. The value of Genuine Acceptance Rate (GAR) is plotted against the value of False Acceptance Rate (FAR) at different threshold values. Genuine Acceptance Rate is the value of FRR subtracted from one $[GAR(\eta) = 1 - FAR(\eta)]$.

CHAPTER 4

METHODOLOGY

4.1 Feature Extraction

The first step is to extract the features from the speech sample which contains specific characteristics of the speaker [35]. There are several speech features indicating the speaker identity. These include Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Predictive (PLP) coefficients, Linear Frequency Cepstral coefficients, Mel Frequency Cepstral coefficients (MFCC) and many others. Linear Predictive Coefficients (LPCs) received much popularity as these can be derived directly from the speech production model of the speakers; also the human auditory processing based features PLP are popular. In this work, MFCC and GFCC feature extraction techniques are used which outperforms other techniques. These are the spectral based features which are acquired using Fourier transformation (FT) or short term fourier transformation (STFT).

4.1.1 Mel Frequency Cepstral Coefficients (MFCC)

From the last many decades MFCC is the most popular technique of feature extraction. In most of the cases MFCC outperforms LPCC, under clean and matched environment. The blocks involved in this technique are shown in Fig [4.1].

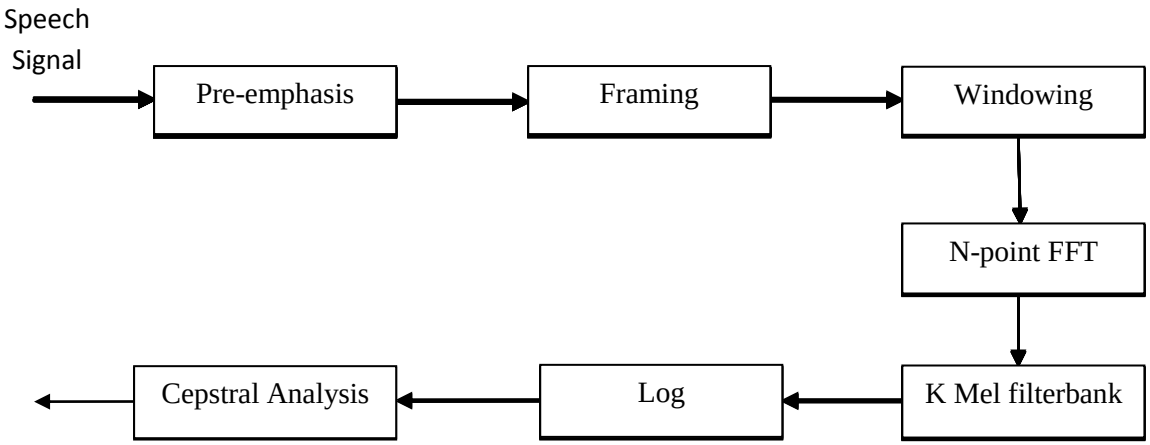


Fig4.1 Block diagram of MFCC feature extraction technique [35]

4.1.1.1 Pre-emphasis

The input speech signal is first pre-emphasized, which implies that the signal waveform is allowed to pass through a high pass filter. This amplifies the energy at high frequencies and even out the natural audio signal production process which has less energy at high frequencies [36]. Thus the variation in the power components of the speech signal reduces. The filter here can be used in any domain time or frequency. The filter equation in time domain is given as:

$$z(t) = x(t) - \beta x(t-1), \qquad 0.9 \leq \beta \leq 1 \tag{4.1}$$

, the value of the coefficient β used is 0.97 mostly. Here, x(t) is the input speech and z(t) is the output. Fig [4.2] shows the effect of pre-emphasis on the speech signal. The spectral density of the original signal has dropped at higher frequencies whereas in pre-emphasized signal the power distribution is better at higher frequencies.

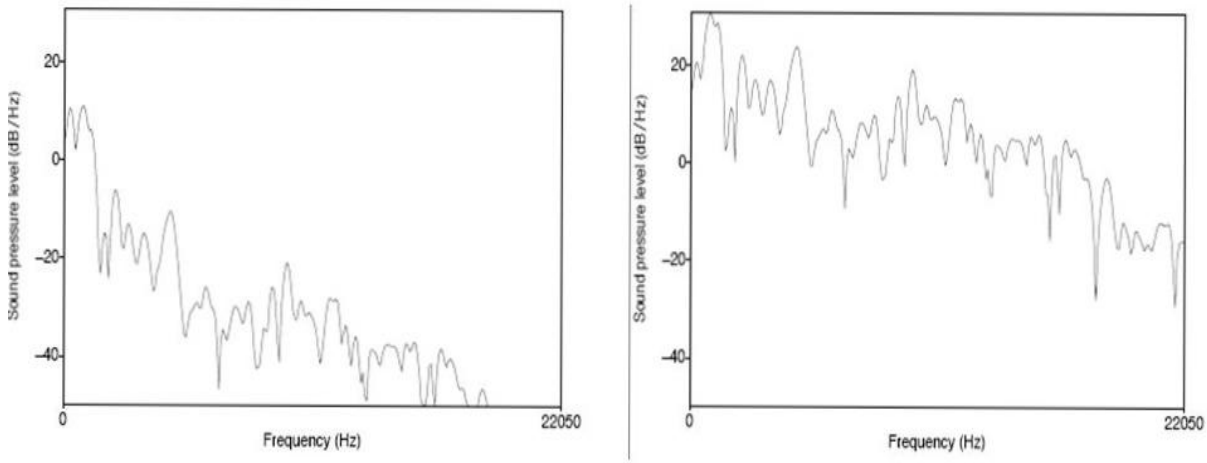


Fig 4.2 Pre-emphasized input speech signal

4.1.1.2 Framing

We all know that the speech signal varies continuously, so it is divided on short time scales as it becomes statistically stationary over these short frames. The frame size is chosen in a way that is neither much shorter nor much longer, if frame size is very small then it's difficult to get a reliable spectral estimation and if it is longer then there's a lot of variations in signal throughout the frame. The continuous speech signal is therefore segmented into N small frames, with frame step size of M samples where M is smaller than N. The first frame containing N samples start from sample zero, the second N sample frame starts with Mth sample and so on, thus the frames overlaps by N-M samples. In our experiments we took N= 256 and M= 128. Discussing in terms of time, the signal is framed over 20-30 ms, standard value is 25ms, with a step size of 10ms.

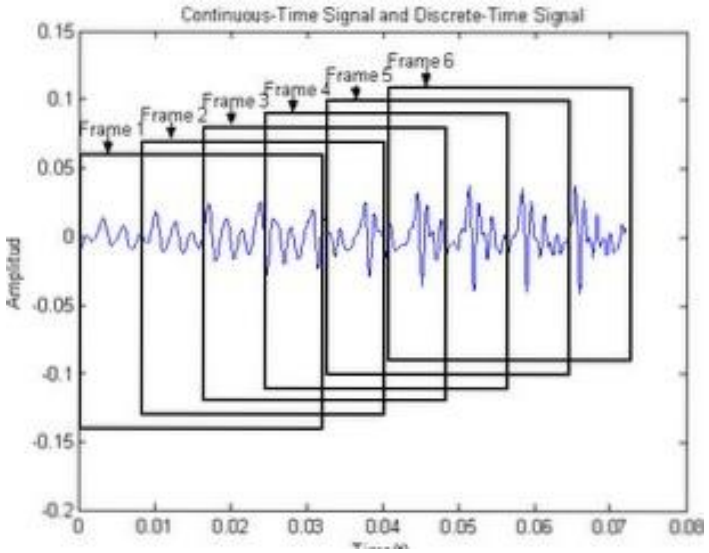


Fig 4.3 Frame blocking of an audio signal

4.1.1.3 Windowing

In this step, each frame of the signal is multiplied with the window function. Windowing is done to decrease the discontinuities of the signal at the beginning and end edges of each frame. If $w(n)$ is the window function, 0 ≤ n ≤ N-1 (N here is the frame length) then the general equation of windowing is given by:

$$z(n) = x(n) * w(n), \qquad 0 \leq n \leq N-1 \tag{4.2}$$

Many different types of window functions are available such as Blackman, Rectangular, Kaiser, Bartlett, Hamming, Turkey and Hanning window functions. In our study we use Hamming window function which is very much popular; it is continuous and smoothes out the sharp edges [37]. The equation describing the Hamming window is given by:

$$w(n) = \sigma - \rho \cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \leq n \leq N-1 \tag{4.3}$$

, here σ=0.54 and ρ = 1- α = 0.46.

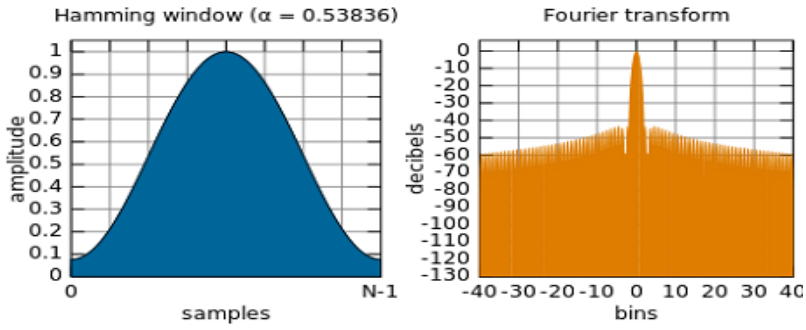


Fig 4.4 Hamming Window in time and frequency domain

The signal frame before and after applying the hamming window function is shown in Fig [4.5].

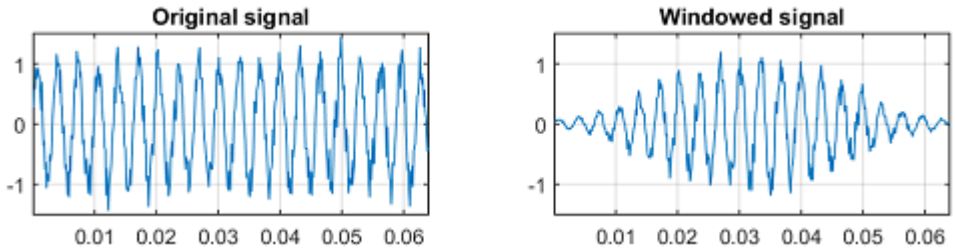


Fig 4.5 Windowed Signal

4.1.1.4 Fast Fourier Transform (FFT)

In this step, FFT of each frame is taken which converts the time domain signal into frequency domain signal. On choosing the algorithm the number of point for the FFT calculation has to be fixed. The number N is always a power of 2 and is usually taken as 512. On applying FFT we get 512 complex spectral values representing the magnitude and phase values of the signal. Since the power spectrum obtained is symmetric across 0, therefore only 256 spectral values uniformly spaced from 0 to $F_s/2$ (F_s here is the sampling frequency) are kept considering only the magnitude spectrum and ignoring the phase information. This step output is referred as power spectrum or periodogram estimate.

4.1.1.5 Mel Spaced Filter bank

The spectrum obtained from above has a lot of fluctuations and also contains a lot of details which is not required by the system. Therefore in this step we multiply the spectrum with the filter bank to reduce the size of the spectrum and to get the spectral envelop which provides the optimum information needed. The process is also called frequency warping as here the magnitude spectrum is converted into mel spectrum. A filter bank consists of a number of bandpass filters and they are characterized by the filter shape and frequency spacing. In MFCC we generally use triangular shaped filters which are localized using Mel scale [34].

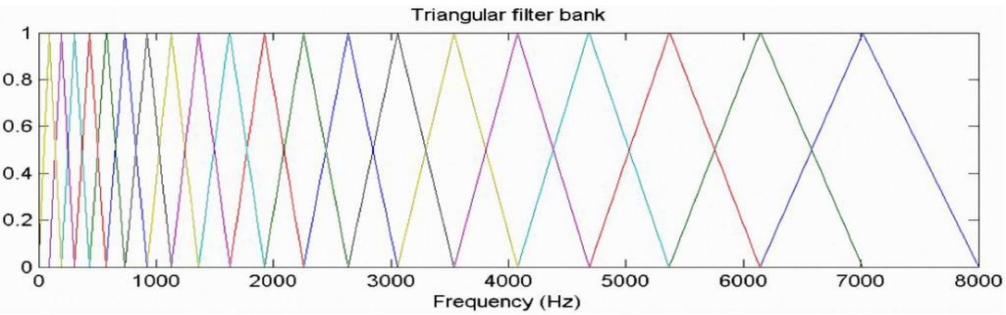


Fig 4.6 Mel scaled 26 triangular filters

Mel scale is based on human auditory system and localizes the filter in a way same as human ear frequency scale. Human ears interpret pitch in a non linear manner, discriminating even small pitch changes at lower frequencies than at higher frequencies. The equation for spacing the central frequencies of the triangular filters according to mel scale is given by:

$$f_m = 1000 \,.\frac{\log(1 + f/1000)}{\log 2} \tag{4.4}$$

Mel spaced filterbank is a series of 26 triangular filters which are distributed over the length of spectrum. Each filter vector is non zero for a particular frequency and is zero on other frequencies. Each filter in the filter bank is multiplied with the signal spectrum one by one and then the coefficients are added to get the energy value at each frequency band, denoting the kth filter bank output as Y_k. At the end of this process we have 26 values which indicate the filterbank energies. Fig [4.7] picturise this process with the help of an example. Further, the log of the filter bank output is taken $(\log Y_k)$ and 20 is multiplied to each coefficient to obtain the values of spectral envelope in dB. This step is also encouraged by human hearing process and it also makes the further calculations easy. The output obtained after this stage are the spectral vectors.

4.1.1.6 Cepstral Analysis

The log Mel spectrum obtained is transformed back to time domain in this step. This conversion is done using DCT. The spectral features are highly correlated because of the overlapping filter banks, thus DCT decorrelates these spectral features to give output in the

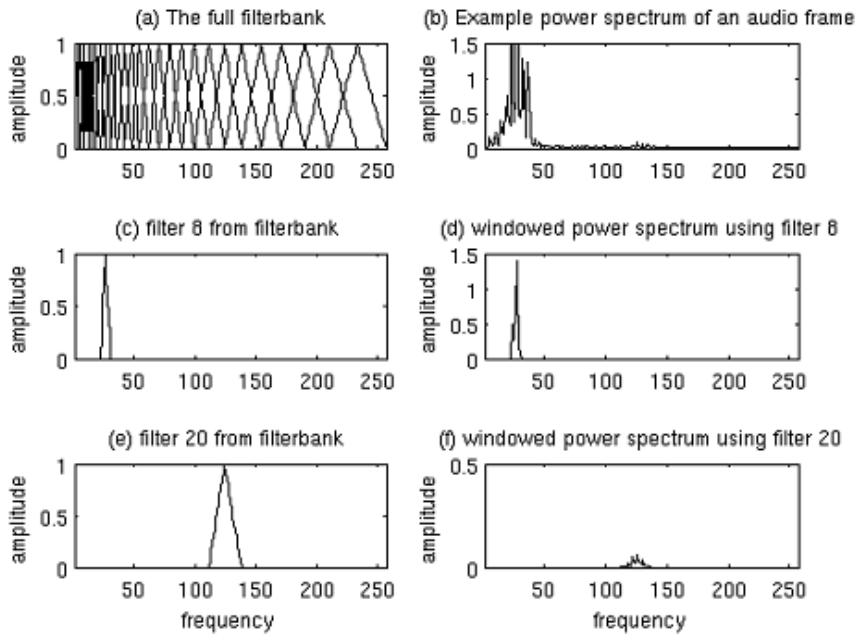


Fig 4.7 Frequency warping process example

form of MFCC. The cepstral coefficients provide a good estimate of the spectral attributes of the signal. The K spectral values are converted into cepstral coefficients using the equation given in [34] is:

$$C_j = \sum_{k=1}^{K} \log(Y_k) \cos\left[j\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \qquad j = 1,2,3 \ldots\ldots\ldots K \tag{4.5}$$

From the 26 cepstral coefficients obtained only the lower 12-13 coefficients are kept this is because the fast filter bank energy changes are represented in higher coefficients which actually degrades the system performance [1]. Also the first cepstral coefficient C_0, is excluded as it contains very less speaker specific information.

4.1.2 Gammatone Frequency Cepstral Coefficients (GFCC)

The major disadvantage of the MFCC features is that its performance decreases in noise and in mismatched condition between enrolling and testing environment. So, to improve the noise robustness of the system the new auditory features GFCC are used. GFCC features which are based on the human auditory system perform better than MFCC in noisy and mismatched conditions. The flow chart describing the process of GFCC feature extraction is shown in Fig [4.8]:

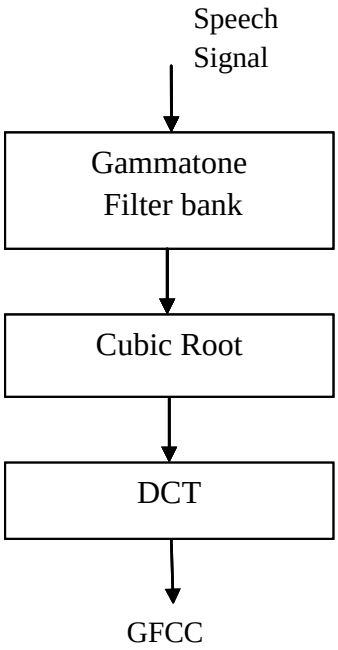


Fig 4.8 Block diagram of GFCC feature extraction technique

4.1.2.1 Gammatone filter bank

The input speech signal is decomposed into T-F domain by performing auditory filtering, which is done using gammatone filter bank [38]. Gammatone filter modelling is based on the structure of auditory periphery. It models the human cochlea, which extract the significant features from the speech signal by its frequency selectivity property. Gammatone filter bank is a set of overlapping bandpass filters. Each filter follows the shape of the gammatone function and its impulse response is represented as:

$$h(t) = bt^{n-1}e^{-2\pi bt}\cos(2\pi F_c t + \theta) \tag{4.6}$$

, here b=1, n is the order of the filter, F_c is the centre frequency, θ is the shift in phase and b is the bandwidth in Hz [39]. Each Gammatone filter's centre frequency and bandwidth is determined by ERB. The bandwidth of the actual auditory filters required to model the human cochlea is derived from ERB. ERB tries to model the auditory filter parameters in a way the signals are present in the human auditory nerve cells or channels. The ERB can be defined using a mathematical formula given by Galsberg et al. [40]:

$$ERB = \int_0^\infty |h(f)|^2 \, df \tag{4.7}$$

, here $|h(f)|$ is the filter transfer function. The ERB and centre frequency relation is given by:

$$ERB(F_c) = 24.7(4.37\frac{F_c}{1000} + 1) \tag{4.8}$$

The value of the bandwidth 'B'of the gammatone filter is derived from ERB, the equation for this was given by Patterson:

$$B = 1.019\, ERB = 1.019\,(24.7\left(4.37\frac{F_c}{1000} + 1\right)) \tag{4.9}$$

The spacing between the gammatone filters in the filterbank is determined by ERB scale. The equation to calculate the centre frequency of the Gammatone filter is given below.

$$F_c = \frac{-1000}{4.37} + \left(F_H + \frac{1000}{4.37}\right)\exp\left(\frac{m}{M}\left(-ln\left(F_H + \frac{1000}{4.37}\right) + \ln\left(F_L + \frac{1000}{4.37}\right)\right)\right) \tag{4.10}$$

, here F_L and F_H denotes the lower and upper frequencies of the gammatone filterbank in Hz, m denotes the filter number and M is the number of filters in the filterbank [39].

The input signal is passed through a gamatone filterbank consisting of 64 filters (channels). The centre frequency of these filters depends on the input data's sampling frequency and varies in the range of 50Hz to 4000Hz.

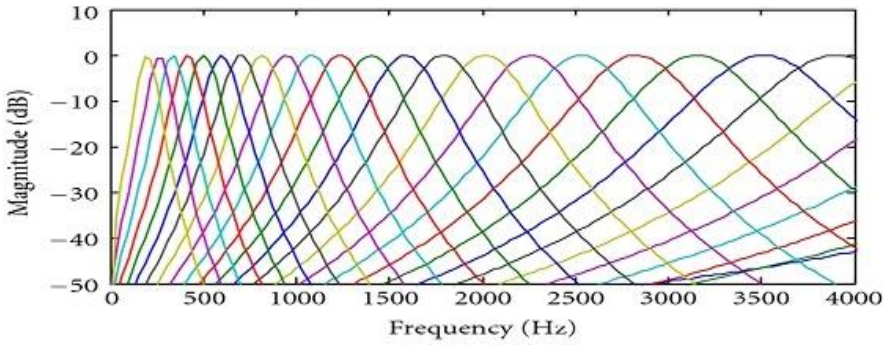


Fig 4.9: 64-channel Gammatone filter bank

The 64-channel filter response is then fully rectified by taking the absolute of each channel output. Since the original sampling frequency is retained in the filter output therefore each channel output is decimated to 100Hz along the dimension of time. The decimation process also cuts down the excess of the information in the signal [41] and produces the time-frequency (T-F) representation of the signal which is a form of cochleagram. Unlike, spectrogram which provides linear frequency resolution, cochleagram presents a much higher resolution at lower frequencies as compared to higher frequencies.

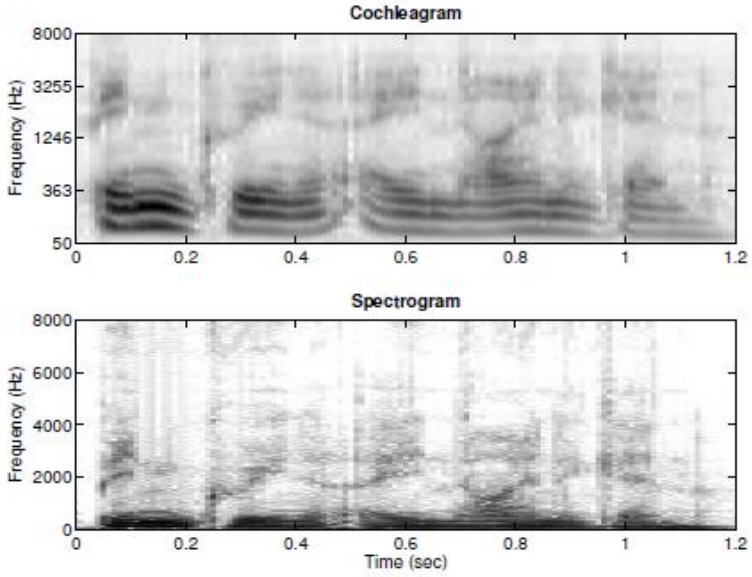


Fig 4.10 Cochleagram and Spectrogram of clean speech

4.1.2.2 Cube Root

In this step cube root operation is performed on the decimated output magnitudes. The equation of cube root operation is given as:

$$G_k[j] = ||g|_{decimate}[j,k]|^{\frac{1}{3}} \tag{4.11}$$

, here $j = 0,1 \dots \dots N-1$, $k = 0,1 \dots . . M-1$, N denotes the number of filter or frequency channels and M denotes the number of the decimated time frames obtained. The resulting output is called the Gammatone feature (GF) vector, consisting of 64 components of frequency and has dimension greater than MFCC vector.

4.1.2.3 Discrete Cosine Transform

GF feature components obtained from above are highly correlated because of the frequency overlapping of the filter channels. Therefore, in order to decrease the dimension and to decorrelate GF components DCT is applied. The output so obtained is called GFCC. The equation for deriving the cepstral coefficients from the Gammatone Features (GF) is given as:

$$C|n| = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} G[i] \cos\left(\frac{n\pi}{2N}(2i+1)\right), \quad n = 0,1, \dots \dots \dots N-1 \tag{4.12}$$

Note for n=0 the coefficient $C|0|$ is equal to the sum of all the components of GF and gives the total energy of GF vector [38]. The coefficients derived here are not the actual cepstral coefficients because the cepstral coefficients are derived by taking the log of the filterbank output. GFCC coefficients are called cepstral coefficients because its functionality is similar to the cepstral coefficients derived by typical cepstral analysis. The lower 23 coefficients are utilised as the GF feature frame's majority of the information is retained in them [26].

4.2 Gaussian Mixture Modelling (GMM)

The speaker verification system is based on likelihood ratio detector (discussed above), and for implementing this LLR detection we need to select the likelihood function $p(X/\gamma)$. Depending on the features being used the likelihood function is selected and generally for text-independent systems the most winning likelihood function is GMM [24]. It is one of the most generic and widely used modelling paradigms. A GMM describes each feature vector in the form of Gaussian distribution, having a characteristic mean and a deviation. The GMM of speaker 'i' is given as the weighted sum of N unimodal Gaussian densities.

$$p(x_k/\gamma_i) = \sum_{k=1}^{N} v_k \; M(x_k : \mu_k, \sum_k) \tag{4.13}$$

, here x_k is the R dimensional feature vector (of an utterance of speaker i), v_k are the mixture weights having value $\sum_{k=1}^{N} v_k$=1, μ_k is the mean vector having dimension Rx1, $\sum_k$ is the covariance matrix having dimension R x R. The $M(x_k : \mu_k, \sum_k)$ are the N unimodal Gaussian densities, which are calculated using equation 4.14 given in:

$$M(x_k{:}\,\mu_k, \sum_k) = \frac{1}{2\pi^{R/2}\sum_k{}^{1/2}} exp\left\{-\frac{1}{2}(x_k - \mu_k)(\sum_i)^{-1}(x_k - \mu_k)\right\} \quad (4.14)$$

The density model parameters are denoted as $\gamma_i = (v_k, \mu_k, \sum_k)$, where k=1,2,3……N and N is the number of the unimodal densities which is equal to number of samples of a speaker [24]. In GMM process as shown in Fig [4.11] all the speech data or utterances from a single speaker are pooled and hence all the possible sound variations from whatever is spoken by a single speaker are modelled into a single model for each speaker. The parameters of the maximum likelihood model $(v_k, \mu_k, \sum_k)$ are calculated using iterative parameter estimation paradigm called EM algorithm. The GMM parameter estimation is done by EM algorithm in a way that the likelihood of the model increases monotonically. Most often five iterations are adequate for the parametric convergence.

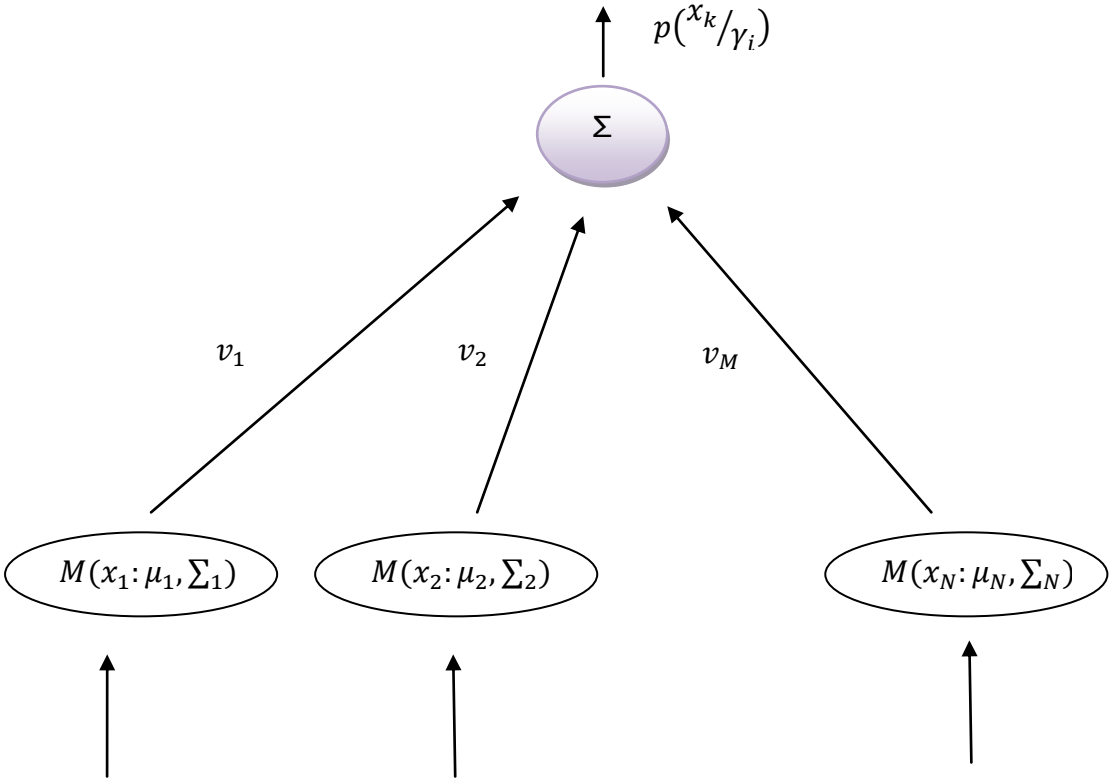


Fig 4.11 GMM mixture computation

We calculate the log likelihood of a speaker model by taking the log of the model density as given below:

$$\log p\left(\frac{X}{\gamma}\right) = \sum_{k=1}^{N} \log\, p(\frac{x_k}{\gamma}) \quad (4.15)$$

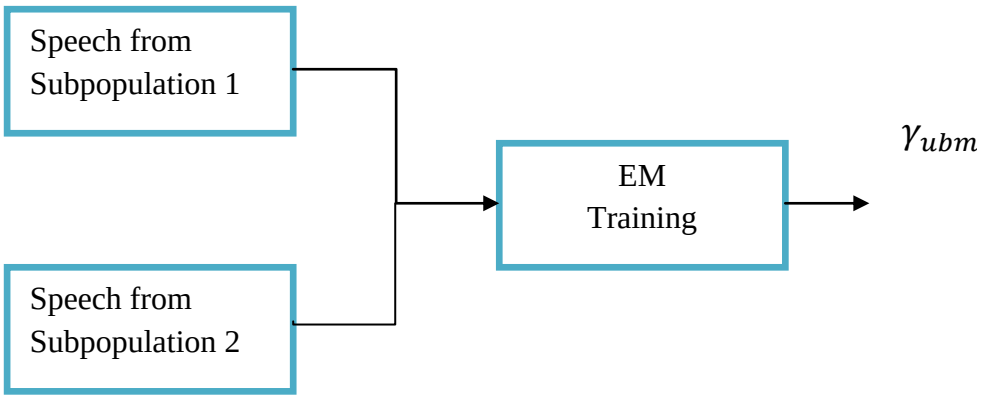
The GMM is a parametric probability density model but it also has some of the properties of non parametric density model. As a parametric model, it possess definite structure and defined

parameters which controls the density model behaviour in some known ways and does not have any constraints on the data that it should have specific distribution type. As a non parametric model it allows the arbitrary density modelling and has various degrees of freedom. There are many advantages to use GMM as a likelihood function such as it has no computation expenses, its model is easy to understand, and also it is unaffected by the temporal aspects of speech. Although GMM is a very successful modelling technique but still it suffers from drawbacks [4]. One is that it needs a large training data for the optimum parameter estimation of the models. The second problem which is very common in every modelling paradigm is that the data which appears in the test sample remains unseen in the training data generates low scores and thus lower down the overall performance of the system. The solution to it is the use of the varied training data.

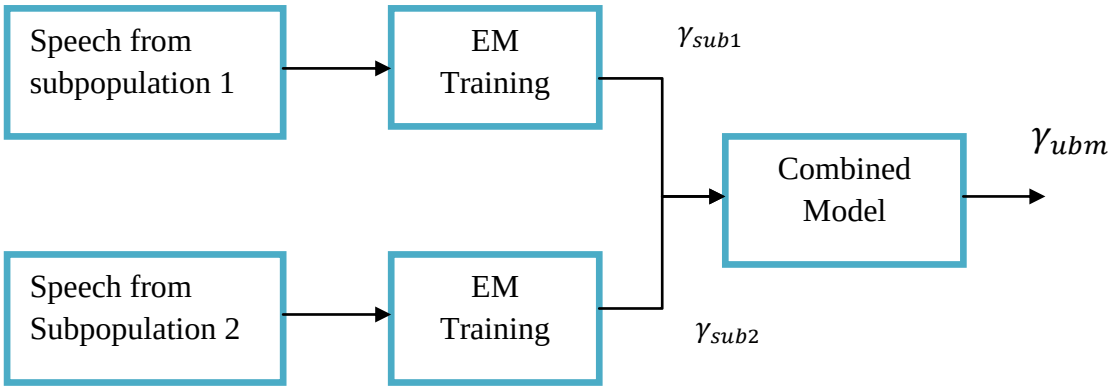
4.2.1 Universal Background Model (UBM)

A single background model is used to represent $p\left(X/\gamma_{\overline{hyp}}\right)$ or $p\left(\frac{X}{\gamma_{UBM}}\right)$ in GMM-UBM system. The UBM is nothing but a GMM trained using a very large data. The speech data is selected in a way that it represents the expected speech which can be encountered during the process of recognition. Also the speech data selection depends on the type of speech, its quality and the composition of speakers. Other than this there is no measure on the speaker number or on the amount of speech data used to train a UBM.

When given the training data, there are two approaches to obtain the final UBM model. First one is by simply pooling the data and then using EM algorithm to train UBM. While pooling the data one should be cautious that subpopulation within the data is balances. For Example when using the data which is gender independent there should be the balance between the female and male speech used. Otherwise there would be biasing toward the dominant speech (female or male) in the final model. The second approach is that for every subpopulation in the data an individual UBM is trained, such as one UBM is trained for female speech data and one is trained for male speech data and then combining the two to form a single UBM. In our study we used the first approach to train the UBM model.



(a)



(b)

Fig 4.12 Two approaches for UBM training

4.2.2 Speaker Model Adaptation

The GMM model is not only estimated using the powerful EM algorithm but also with fewer amounts of data the parameters is further adapted according to new data using MAP adaptation. The basic thought behind this adaptation approach is to update the well trained parameters in that single UBM model while deriving the hypothesized speaker model via adaptation. Thus the adaptation offers a tighter coupling among the UBM and the speaker's model which in turn increases the system performance. The process of adaptation is similar to EM algorithm only the combination of new statistic estimates with the old statistic of the UBM parameters is done [34]. The general MAP adaptation equation is given in [42]:

$$\hat{\mu}_{new} = \frac{L_s}{L_s + \tau}\mu_{old} + \frac{\tau}{L_s + \tau}\mu_{old} \tag{4.16}$$

, here L_s is the likelihood of the hypothesized speaker data, μ_{old} old UBM mean, τ is the weighted parameter and $\hat{\mu}_{new}$ is the new adapted mean of the UBM mixture.The equations to

update the old statistic estimates of UBM with the new statistic estimates of the training data are given below:

$$\hat{v}_k = [\frac{{\beta_k}^v m_k}{T} + (1 - {\beta_k}^v) v_i]\delta \tag{4.17}$$

$$\hat{\mu}_k = {\beta_k}^\mu E_k(x) + (1 - {\beta_k}^\mu)\mu_k \tag{4.18}$$

$${\hat{\rho}_k}^2 = {\beta_k}^\rho E_k(x^2) + (1 - {\beta_k}^\rho)({\rho_k}^2 + {\mu_k}^2) - {\hat{\mu}_k}^2 \tag{4.19}$$

, here $[{\beta_k}^v, {\beta_k}^\mu, {\beta_k}^\rho]$ are the adaptation coefficients representing weight, mean and variance respectively, which controls the balance among the new and old statistic estimates [24]. Also, $E_k(x)$ are the statistic estimates of the hypothesized speaker vector X, and δ is the scale factor. The new updated UBM parameter values of weights, means and variances are denoted by $\hat{v}_k, \hat{\mu}_k$ and ${\hat{\rho}_k}^2$ respectively. In GMM-UBM system, a single value for the adaptation coefficient which is efficient for all the parameters is used $({\beta_k}^v = {\beta_k}^\mu = {\beta_k}^\rho = {m_k}/{m_k + r})$, m_k is the probabilistic count of the speaker vector X and the value of the relevance factor is fixed as 16.

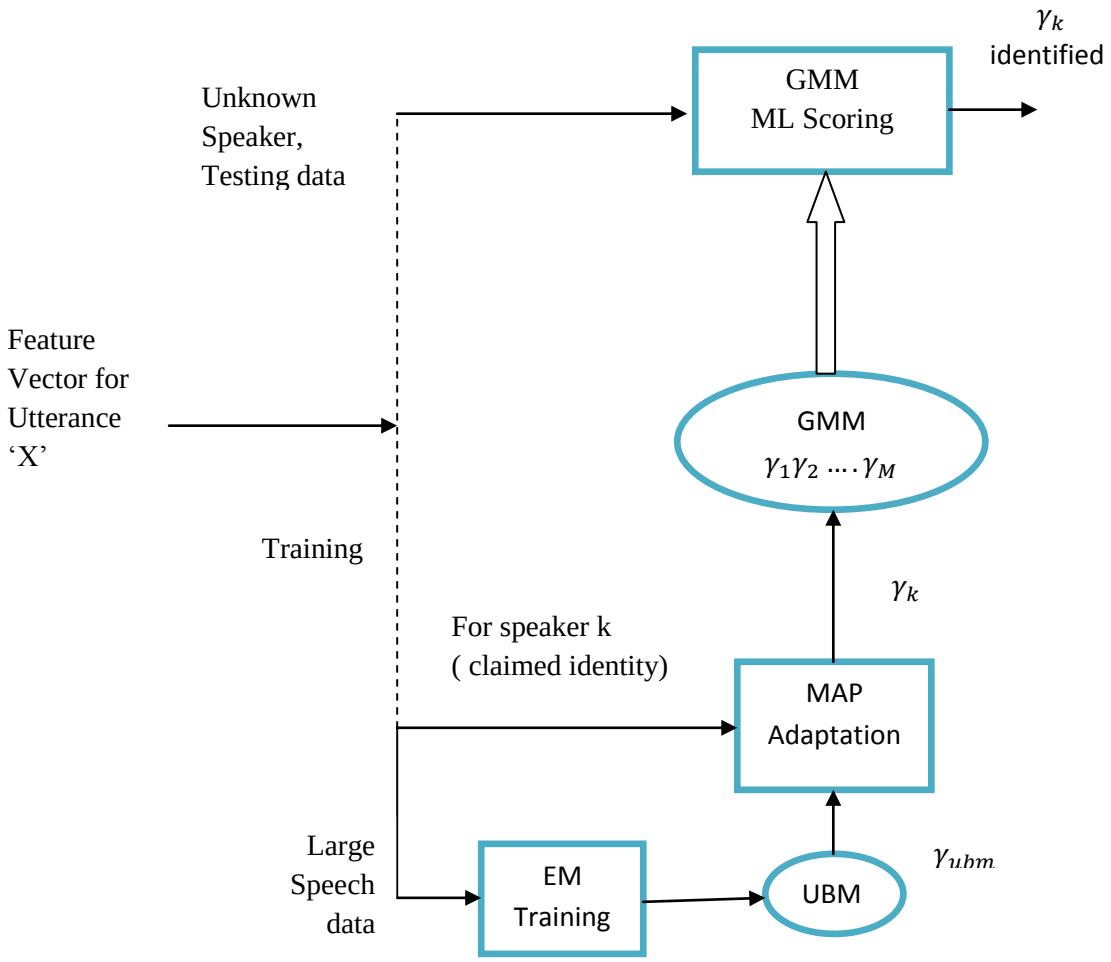


Fig 4.13 Model representing GMM-UBM process

It's proved in [24], that the performance of the GMM-UBM system is very much better than the performance of the GMM system. One of the reasons for this is that the unseen acoustic events that occur in the recognition speech do not affect the adapted models used in the log likelihood ratio. Also the number of the mixtures in a GMM-UBM is more due to the large training data than the GMM system. In an individual GMM system if the unseen data (data unseen while training the model) is encountered then the zero likelihood ratio is obtained, where as in GMM-UBM system the unseen data is covered by UBM and generate low likelihood ratio, thus giving better results.

4.2.3 GMM Classification

GMM classifier is one of the simplest classifier based on unsupervised learning process. In testing phase a set of feature vector X is extracted from the unknown speaker Z. From the M GMM speaker models the claimed speaker model is extracted and the log-likelihood ratio is computed as given in equation [3.2] above

$$\Lambda(X) = \log p(X/\gamma_{identity}) - \log p(X/\gamma_{UBM})$$

The log of the two GMMs is evaluated as shown in equation [4.15] one by using the hypothesized (claimed) speaker model and the other using the single speaker independent UBM model as $logp(X/\gamma_{identity})$ and $logp(X/\gamma_{UBM})$. The MAP adaptation of the hypothesized (claimed) speaker model and the UBM has lead to fast score calculations. The Λ(X) gives the log-likelihood scores which are then compared against the threshold value and the decision is made as true or false. The log-likelihood scores are computed using both MFCC and GFCC features separately.

4.3 K- Nearest Neighbour Classification

KNN classification algorithm is the most simple among all machine learning algorithm. KNN classification is a non parametric method as it does not make any assumptions of the parameters of input data. Also this classifier is based on instance based learning which is also named as memory base learning. In this type of learning the new observations are compared directly with the training instances stored, no generalization or model is trained by the learning concept. The computation cost of learning is zero all the computation cost is in the prediction process. Therefore it is also called lazy learning.

The test data is classified according to its distance from the training samples. The closer the test sample from the training sample the more is the probability that the two samples belong to the

same class. The errors caused due to the noise or due to the less training data can be suppressed and the robustness can be enhanced by finding more number of neighbours (i.e. increasing the value of k). Then the final decision is taken by applying the majority rule. It is a very simple method but performs very well even for complex approximation of the decision function.

The KNN classifier requires three things: the distance measure d, the parameter k, and the input reference templates. The value of the parameter k should be neither too low nor high. A smaller value of k suffers from noise effects and a larger value of k reduces noise effects but the class boundaries become less distinct. Therefore the value of k should be selected carefully depending on the input data size. The commonly used distance metric used is Euclidean distance measure. The input data for training should be dimensionally reduced before applying the KNN because calculating the Euclidean distance becomes very complex over a large data. Therefore the feature vectors obtained are dimensionally reduced before applying the KNN. The Gaussian Membership Function (GMF) is applied to both MFCC and GFCC features to reduce their dimension. The feature vectors are reduced to size 1x100 for each sample [43].

Training Phase: In training of KNN a reduced set of features are stored and then clustering is done. In clustering a random vector is chosen as a centre and concentric rings are constructed with radius *2kd* (k=0,1,2….), assuming that the features from same speaker being similar have the distance less than d. So every instance occurring in the same ring is said to be a set. A class label (w_i) is given to each set, i= number of set or concentric rings.

Testing Phase: A new input vector say y is to be classified, then the k – feature vectors which are closest to test vector on the basis of the distance d are selected. Assume, k_i are the number of feature vectors out of the k vectors that belongs to class w_i. Thus the vector y will belong to class w_i which has the highest class probability. The mathematical representation of the class probability is given below:

$$P\left({}^{w_i}/_{y}\right) = \frac{k_i}{k}\, p(w_i) \tag{4.20}$$

The output form required in verification process is in the form of scores. Genuine scores are calculated by measuring the distance between the test speech sample and other samples from same speaker, whereas the imposter scores are calculated by finding the Euclidean distance between the test sample and other speaker samples. Although KNN is a simple and benchmark method still it suffers from many drawbacks: as the input samples are simply stored therefore it requires a large memory space, requires a lot of computations and calculations during the test phase, which makes this classification process slow and it is assumed that the two samples that

are closest belong to the same class which is not always true [44]. Because of these drawbacks we used GMM as classifier which is faster than the KNN.

4.4 Problem statement

Firstly, in real world applications the speaker verification systems suffer from various challenges such as additive noise, mismatched conditions and channel/handset variability. Also the conventional method MFCC used for feature extraction do not perform well in noisy and mismatched conditions. Secondly, the unimodal systems have various problems such as spoof attacks, noisy sensor data and unacceptable error rates. So there was a need to develop a robust system perform well in a wide range of conditions.

4.5 Proposed Method

In this section we describe the method or process used in our study to develop a robust verification system. Proposed method used MFCC which extracts meaningful speaker characteristics and GFCC which is noise robust. The advantages of two individual systems (MFCC and GFCC) are combined by score level fusion to get multimodal system which is free from the drawbacks of unimodal systems. The method consisting of various stages is shown in the Fig [4.14].

4.6 Fusion of Two Systems

Fusion of two individual systems can be done at four distinct levels of information. Here in our study we used Score level fusion. It's an appropriate approach and most preferred because the best trade-off of the information content is offered in it and also it is easiest to apply [32].

4.6.1 Score Level Fusion

There are several merits of score level fusion which are stated as: The next level of rich information is contained in the match scores after features of the input data. The multimodal biometric system can easily utilize the matching scores obtained from the existing individual (unimodal) systems. Thus the information (in the form of scores) evaluated by the unimodal system are easily accessed and combined. This also eliminates the testing phase in multimodal system. All this motivates the use of score level fusion technique. Score level Fusion further has two categories: (a) Transformation-based and (b) Classifier based. And we have used both these approaches of score level fusion.

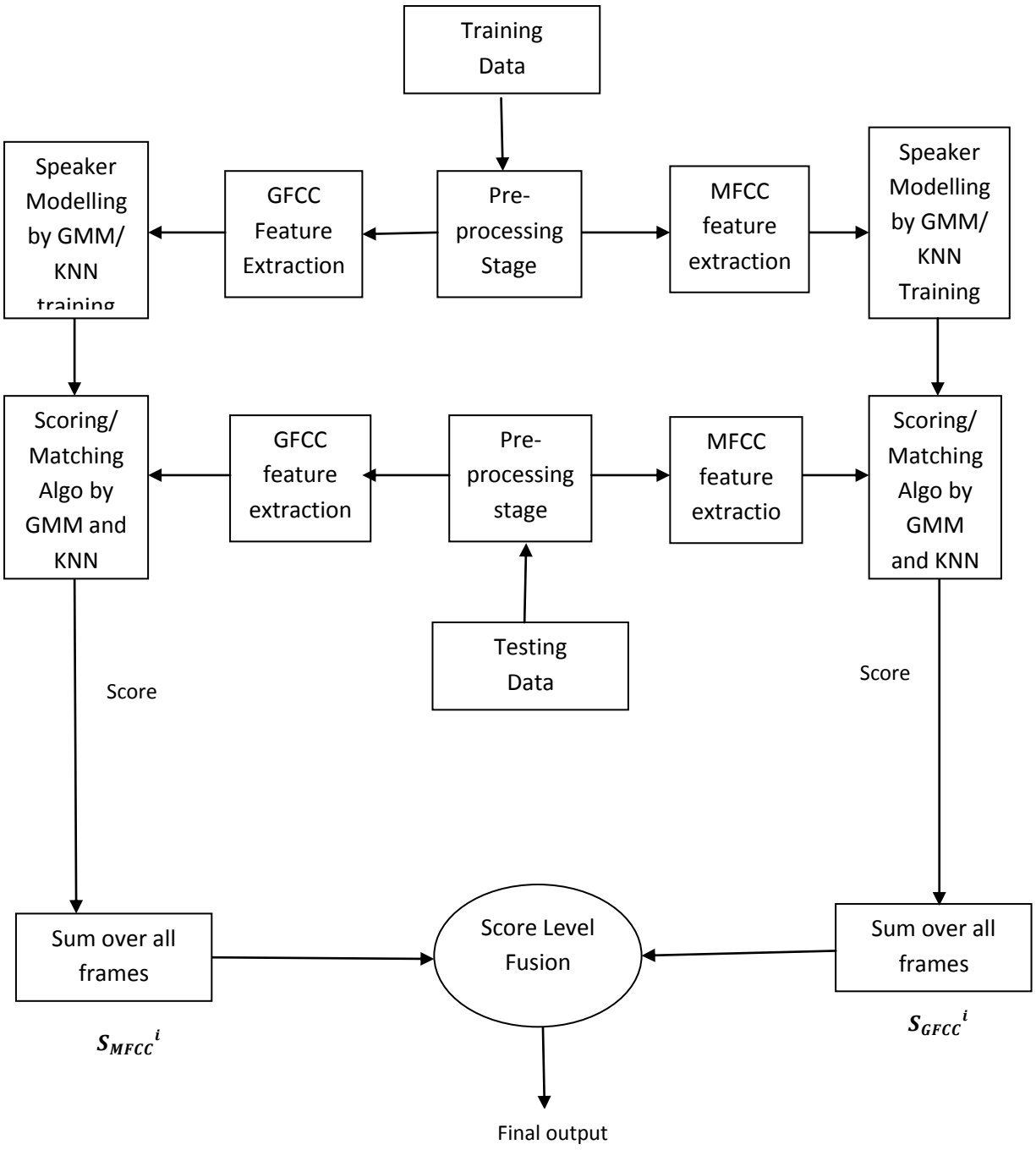


Fig 4.14 Block diagram of the proposed methodology

Transformation based score level fusion: In this approach the matching scores are obtained individually from MFCC and GFCC features using the KNN classifier (discussed above). These scores obtained are normalised by the Min-Max normalisation technique and then various score level fusion rules are applied over the normalised scores. Further the decision is made using threshold based method.

Normalization: Scores from different models are brought to a common scale range so that their combination proves to be meaningful; this process is called normalisation [45]. The normalisation

scheme used in our work is Min Max Normalisation, it is the simplest of all other maximum values) of the match scores are known. In this scheme the raw matching score's interval shifts to [0,1] (minimum value 0 and maximum value 1) and the rest of original distribution of the scores is retained apart from a scaling factor. Let s_k represents a set of raw scores, k=1,2…..n, and s_k' represents the normalise scores Also let the maximum and minimum value of the raw scores are represented as *max* and *min.*

$$s_k' = \frac{s_k - min}{max - min} \tag{4.21}$$

Score Level Fusion Rules: If ${s_k}^m$ and ${s_k}^g$ are the two scores for kth speaker generated from the two individual models (MFCC and GFCC respectively) and S_k is the fused score, k=1,2…..n. Then the various fusion rules applied are given by:

$$1.\, Sum\ Rule: S_k = {s_k}^m + {s_k}^g$$

$$2.\, Product\ Rule: S_k = {s_k}^m * {s_k}^g$$

$$3\ Hmacher\ t-norm: S_k = \frac{{s_k}^m\, {s_k}^g}{{s_k}^m + {s_k}^g - {s_k}^m\, {s_k}^g}$$

$$4\ Frank\ t-norm: S_k = log_p \frac{(1 + (p^{{s_k}^m} - 1)(p^{{s_k}^g} - 1)}{p - 1}$$

, here *p*= frank t-norm function parameter.

Classifier Based Score Level Fusion: In this approach no normalisation of scores is needed before combining the scores. The classifier used for this fusion approach is GMM. Thus two speaker sets were generated ${S_{MFCC}}^k$ and ${S_{GFCC}}^k$, k=1, 2…n (n are the number of speakers) each from MFCC and GFCC models. The scores from two different models are generated using same classifier, so no normalisation of scores is needed. The fusion of the scores is done using the weighted sum rule which is expressed as:

$${S_{com}}^k = w{S_{MFCC}}^k + (1 - w){S_{GFCC}}^k \tag{4.22}$$

, here ${S_{com}}^k$ represent the combined scores. The equation above is showing the general process. But in actual the scores evaluated from GMM of each model (MFCC and GFCC) is in the form of genuine scores and imposter scores. So the formula used is:

$${S_{com\_gen}}^i = w{S_{MFCC\_gen}}^i + (1 - w){S_{GFCC\_gen}}^i, \quad for\ genuine\ scores \tag{4.23}$$

$${S_{com\_imp}}^j = w{S_{MFCC\_imp}}^j + (1 - w){S_{GFCC_{imp}}}^j, \quad for\ imposter\ scores \tag{4.24}$$

, here i= number of genuine scores and j= number of imposter scores.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Database

The experiments are done using VoxForge database. These are text independent speech utterances in English language. It consists of 348 speakers with 10 samples of each speaker. Out of these 10 samples of each speaker 9 are used for training the classifier and 1 sample is used for testing. Also, the test data is mixed with the noise signals at various SNR levels (-5, 5, 10 and 20 dB).

5.2 Experimental Setup

The first stage of the experiment was to extract 13-dimensional MFCC feature vector from a pre-emphasized speech signal. The second stage was to extract 22-dimensional GFCC features using the same speech signal. The last stage is the evaluation phase which is performed using two different classifiers separately.

In one case K-nearest neighbour (KNN) classifier is used to evaluate the scores between training samples and testing samples. The KNN classifier is trained using both the features MFCC and GFCC individually, but before this the feature vector size is reduced to length 100 using GMF (Gaussian membership function). The scores obtained are in the form of Euclidean distance which are further of two type: genuine scores and imposter score (discussed above in KNN). These scores obtained are normalised and fused using various score level fusion rules and then the verification of the system performance is done using the receiver operating curve (ROC) between FAR (false acceptance rate) and GAR (genuine acceptance rate).

In another case the GMM-UBM model is used to evaluate the likelihood scores using both MFCC and GFCC features separately. No feature reduction method is used to train GMM-UBM model. The likelihood scores obtained are also of two type: genuine scores (348) and imposter scores (347*348=120756), which further are score level fused using weighted sum rule. Its performance is also verified using the ROC curve between FAR and GAR and by computing the error rates.

5.3 Results

1. Using KNN classifier

Table 5.1 describes the verification results of MFCC, GFCC and combined system in terms of error rates. Since the error rates FAR and FRR can't be minimised simultaneously. Therefore the results are specified by taking the minimum FAR values (as in column 2 and 3) and the corresponding GAR rate (verification rate) is shown to demonstrate the system performance. The third column depicts the value of EER, lesser is the value of EER better the system performs. Thus the performance of the system keeps on increasing and the fused system outperforms the other two.

Table 5.1: MFCC, GFCC and fused MFCC-GFCC (F-MFCC-GFCC) results in clean speech

Modality	False Acceptance Rate (FAR%)		EER%(Equal Error rate)
	0.1	1	
MFCC	85	91.85	3.02
GFCC	86.1	92	0.88
F-MFCC-GFCC	90.1	94.6	0.854

The figure [5.1] depicts the ROC curve between FAR and GAR values by varying the threshold with a step size of 0.01.

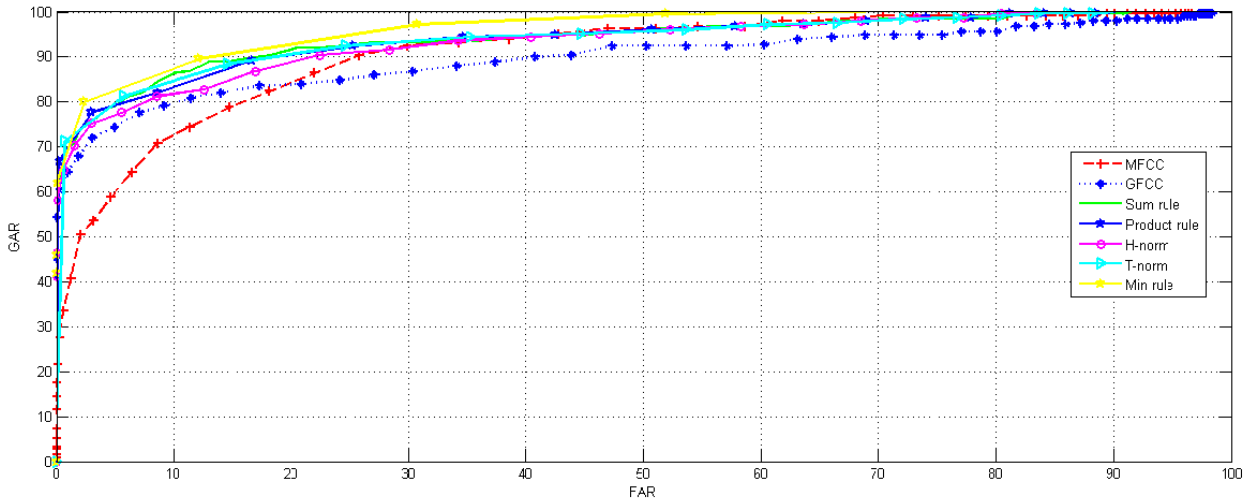


Fig 5.1: ROC of MFCC, GFCC and F-MFCC-GFCC

2. Using GMM-UBM classifier

Table 5.2 is same as table 5.1 the only difference is the change in classifier, in the third row it can be seen that for 1% FAR the GAR achieved 100% rate (i.e. for FAR 1% the false rejection is 0).

Here also in last column the EER is decreasing thus the system performance is increasing and the fused system outperforms the other two.

Further for evaluating the new features performance in noise, three types of noise: babble noise, destroyer engine noise and factory noise are added to the test speech samples in different SNRs from -5dB to 20 dB. Table 5.3, 5.4 and 5.5 shows the EER achieved for three different systems with different SNR levels. Figure 5.2 to 5.14 shows the DET curves with all the features at appropriate SNR levels. The baseline is to compare the unimodal system with the combined system.

Table 5.2: Results using GMM-UBM classifier under clean speech

Modality	**FAR %(False Acceptance Rate)**		**EER%(Equal Error Rate)**
	0.1	**1**	
MFCC	58.9	88.7	2.80
GFCC	85.06	99.71	0.61
Fused MFCC-GFCC	89.94	100	0.57

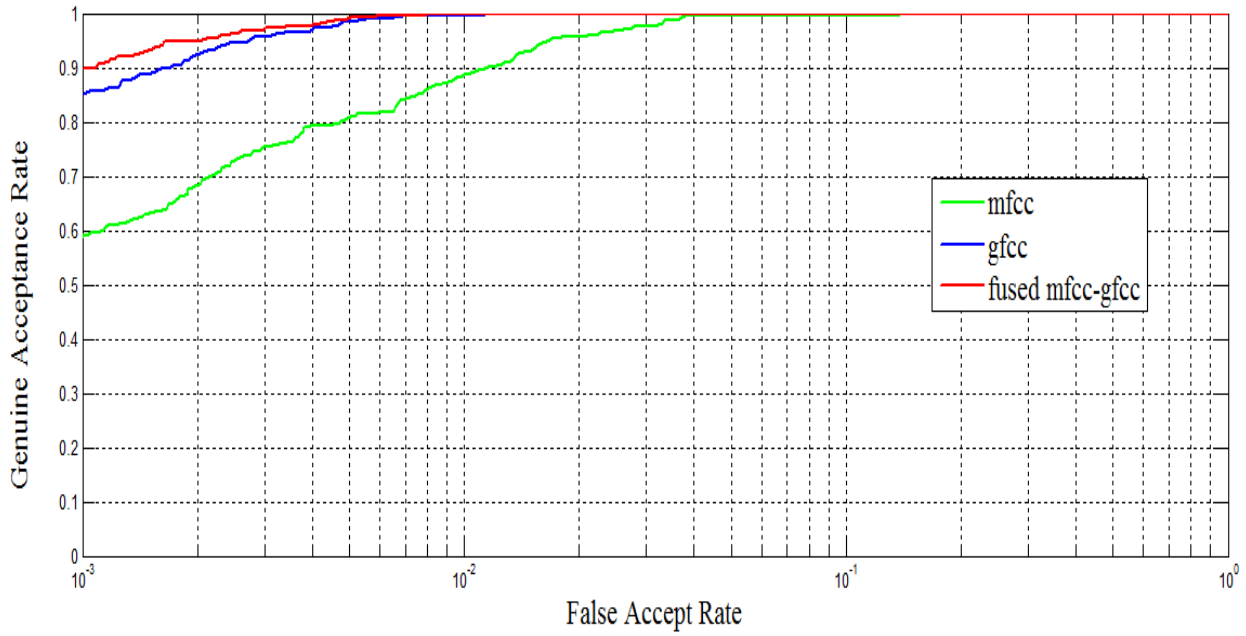


Fig 5.2 ROC Curve in clean speech (using GMM-UBM)

Table 5.3: Value of EERs of MFCC, GFCC and F-MFCC-GFCC at babble noise with different SNR levels

SNR(dB)	Equal Error Rate (EER %)		
	MFCC	GFCC	F-MFCC-GFCC
-5	36.01	32.83	30.53
5	18.17	13.72	9.93
10	10.48	9.88	8.67
20	4.34	3.43	2.81

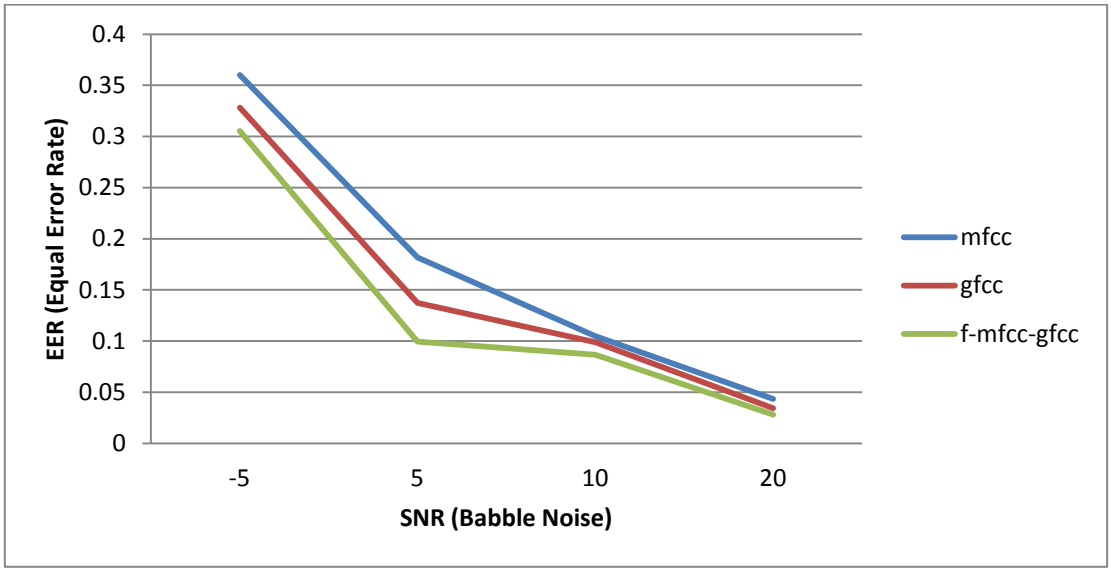


Fig 5.3 Plot of SNR vs EER values of three systems at Babble noise

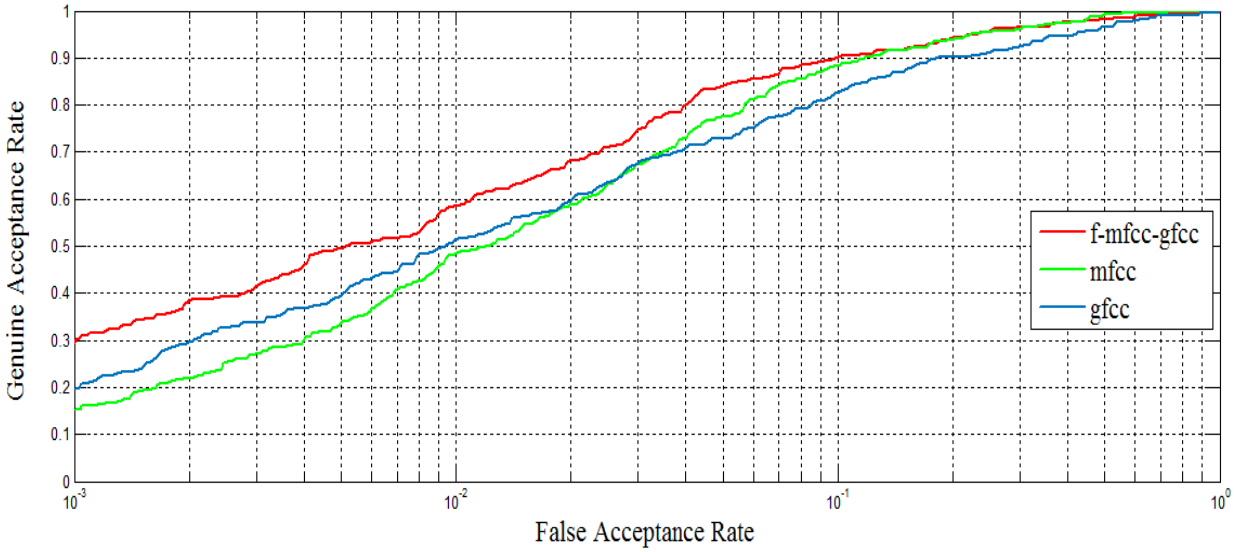


Fig 5.4 ROC Curve at 5dB SNR level (Babble Noise)

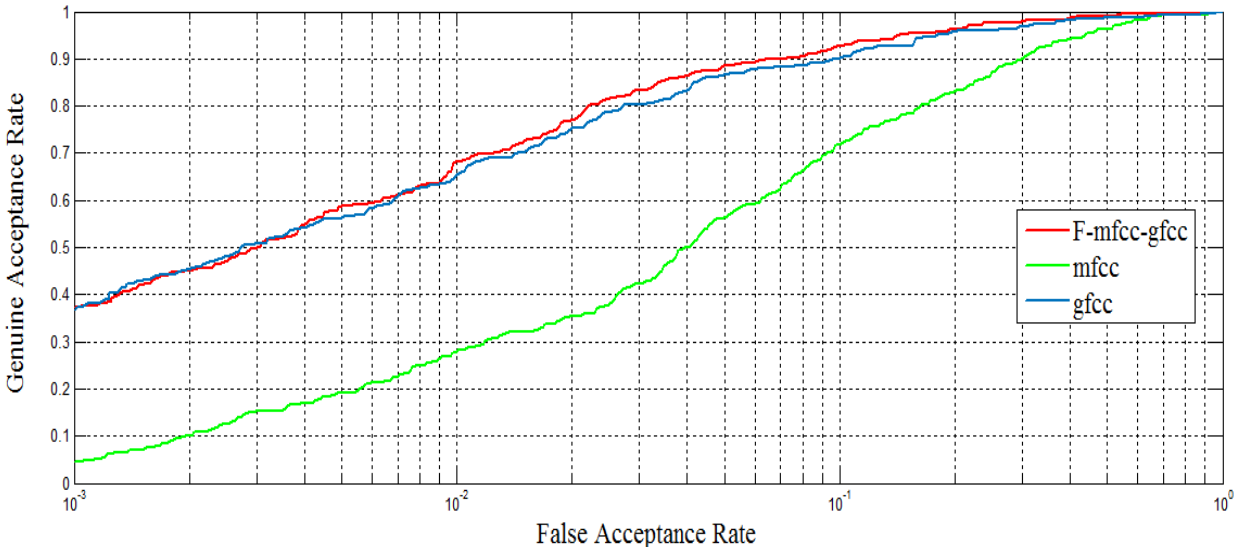


Fig 5.5 ROC Curve at 10dB SNR level (Babble Noise)

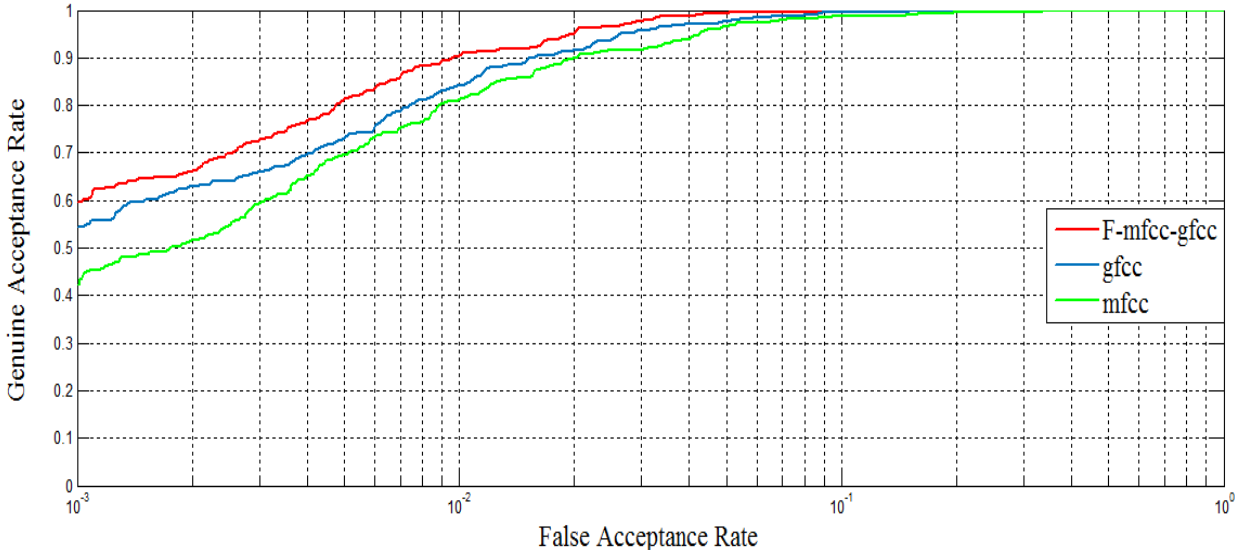


Fig 5.6 ROC Curve at 20dB SNR level (Babble Noise)

Table 5.4: Values of EER of MFCC, GFCC and fused MFCC-GFCC at destroyer engine noise with different SNR levels

SNR (dB)	Equal Error Rate (EER %)		
	MFCC	GFCC	F-MFCC-GFCC
-5	44.04	49.45	40.44
5	35.89	34.71	31.7
10	25.18	23.34	21.42
20	8.52	8.12	6.64

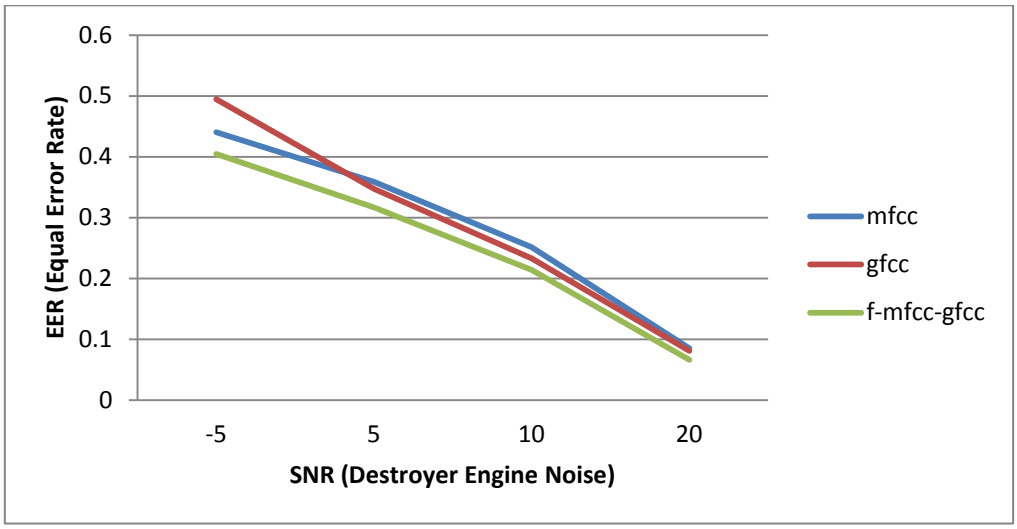


Fig 5.7 Plot of SNR Vs EER of three systems at Destroyer Engine Noise

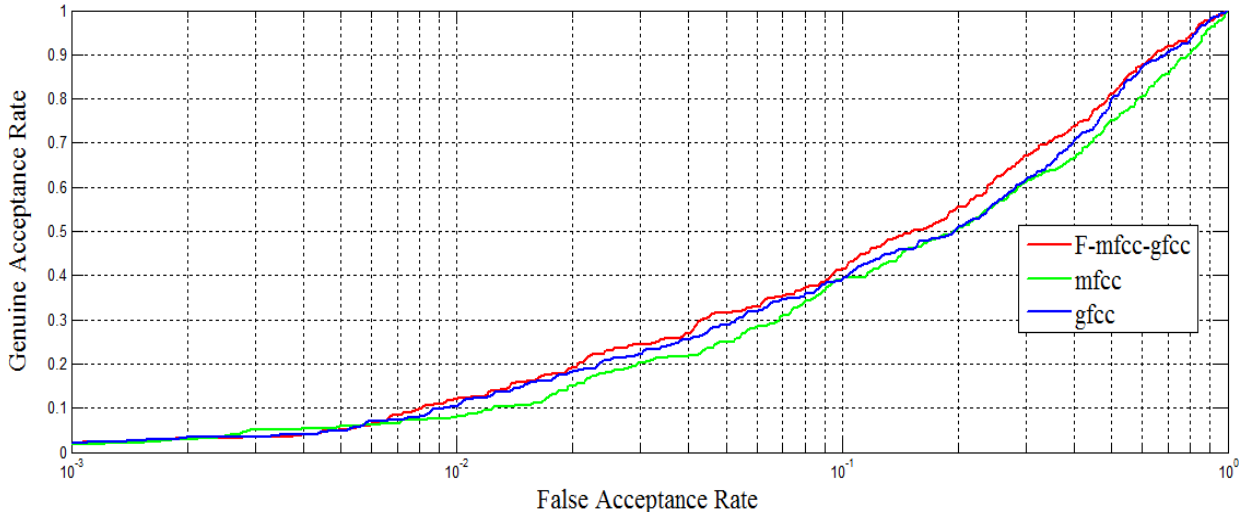


Fig 5.8 ROC Curve at 5dB SNR level (destroyer engine noise)

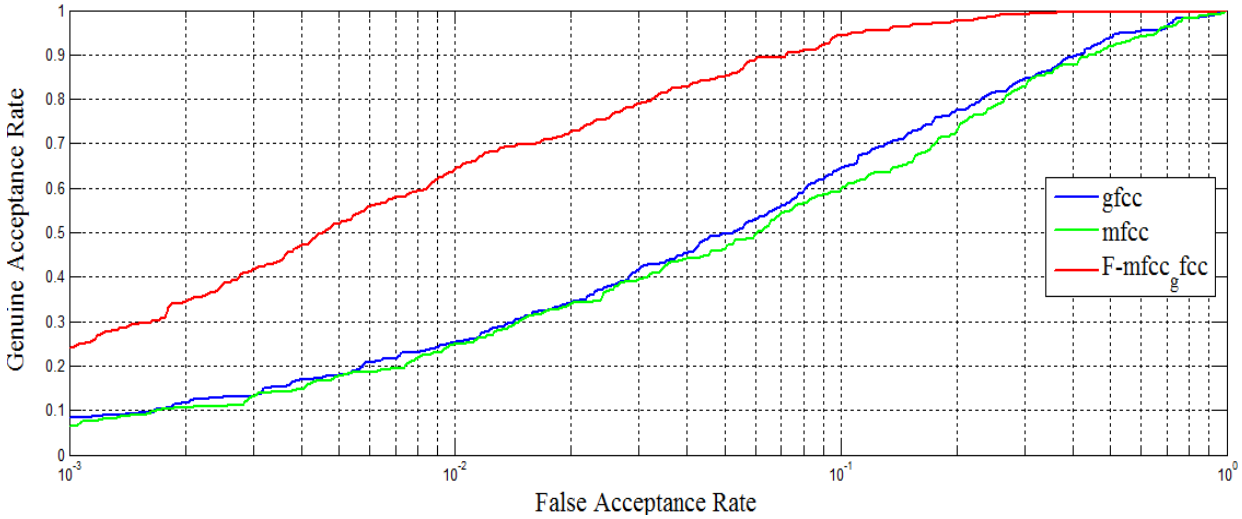


Fig 5.9 ROC at 10dB SNR level (destroyer engine noise)

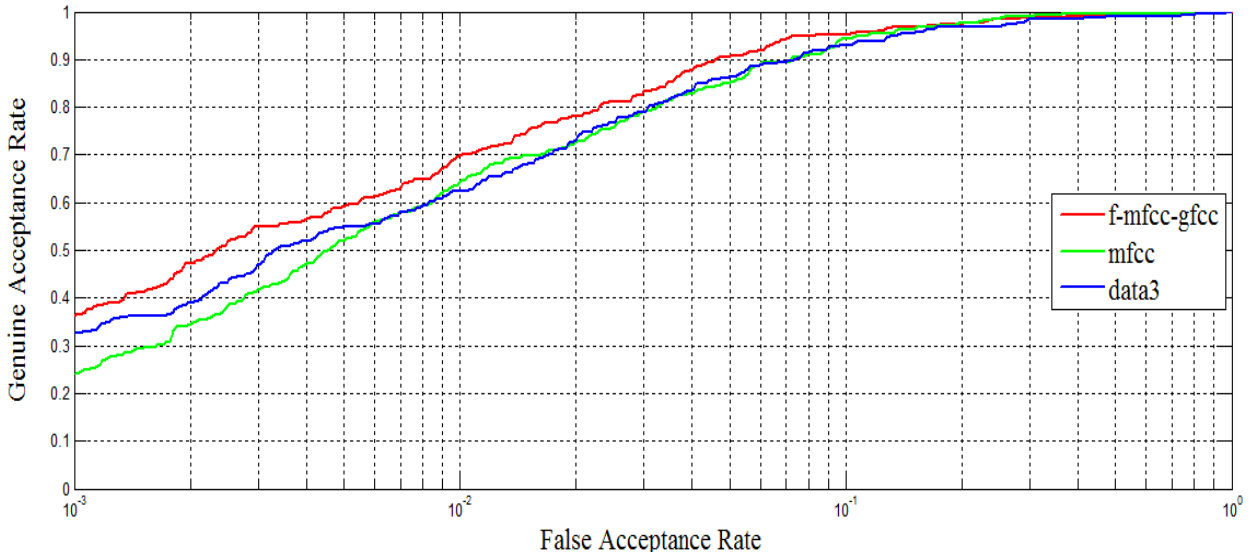


Fig5.10 ROC at 20dB SNR level (destroyer engine noise)

Table 5.5: Values of EER of MFCC, GFCC and fused MFCC-GFCC at factory noise with different SNR levels

SNR (dB)	Equal Error Rate (EER %)		
	MFCC	GFCC	F-MFCC-GFCC
-5	44.83	42.4	40.39
5	23.1	21.51	19.08
10	13.18	13.07	10.92
20	4.36	3.93	2.6

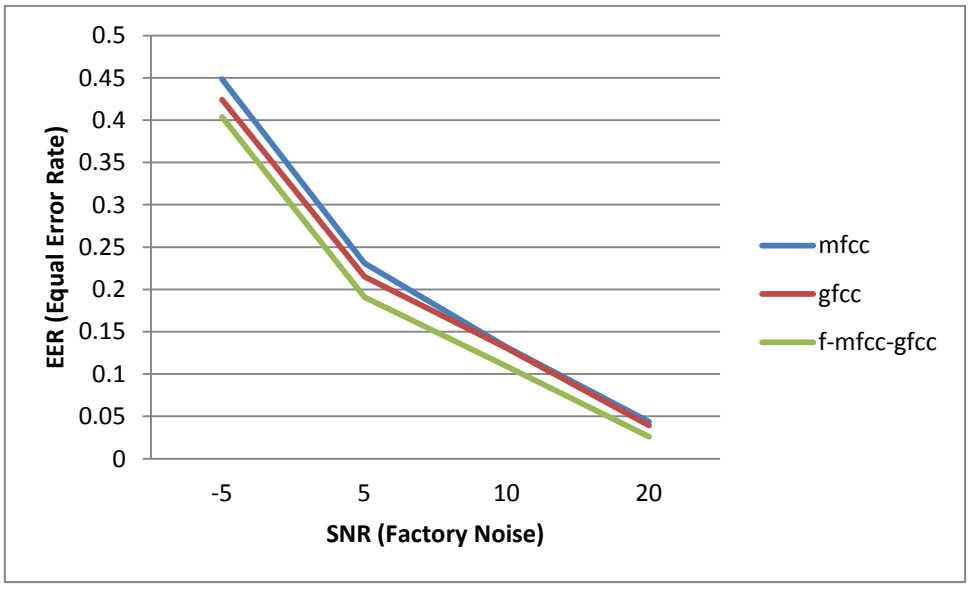


Fig 5.11 Plot of SNR Vs EER of three systems at factory noise

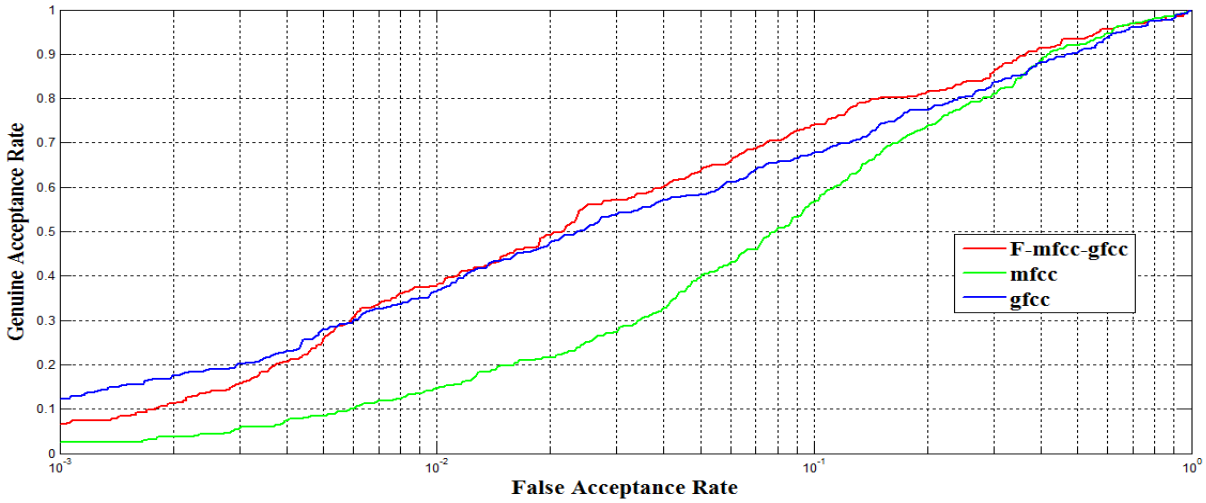


Fig 5.12 ROC Curve at 5dB SNR level (factory noise)

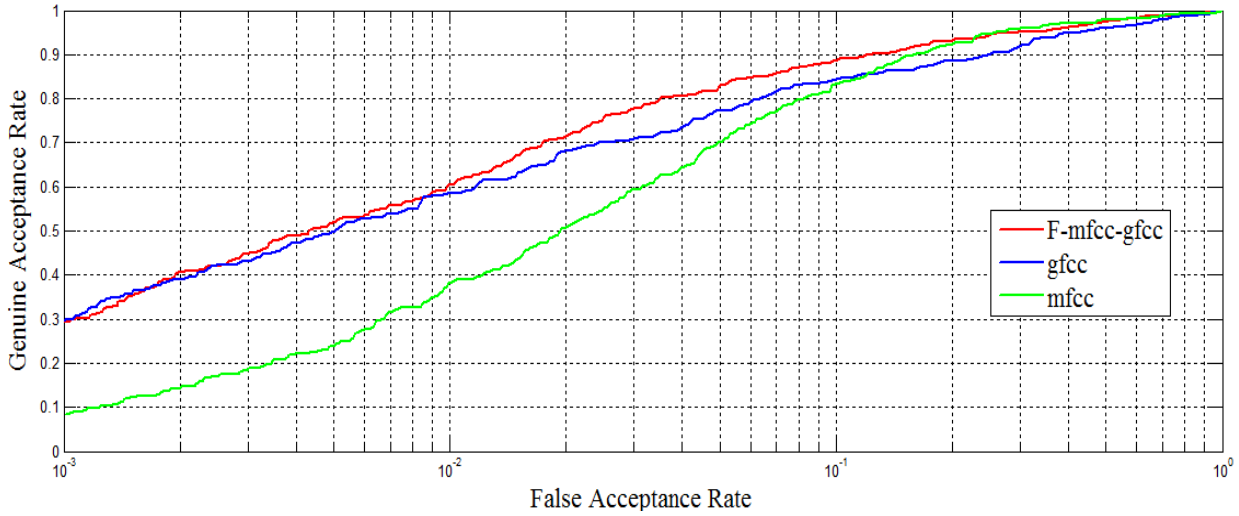


Fig5.13 ROC Curve at 10dB SNR level (factory noise)

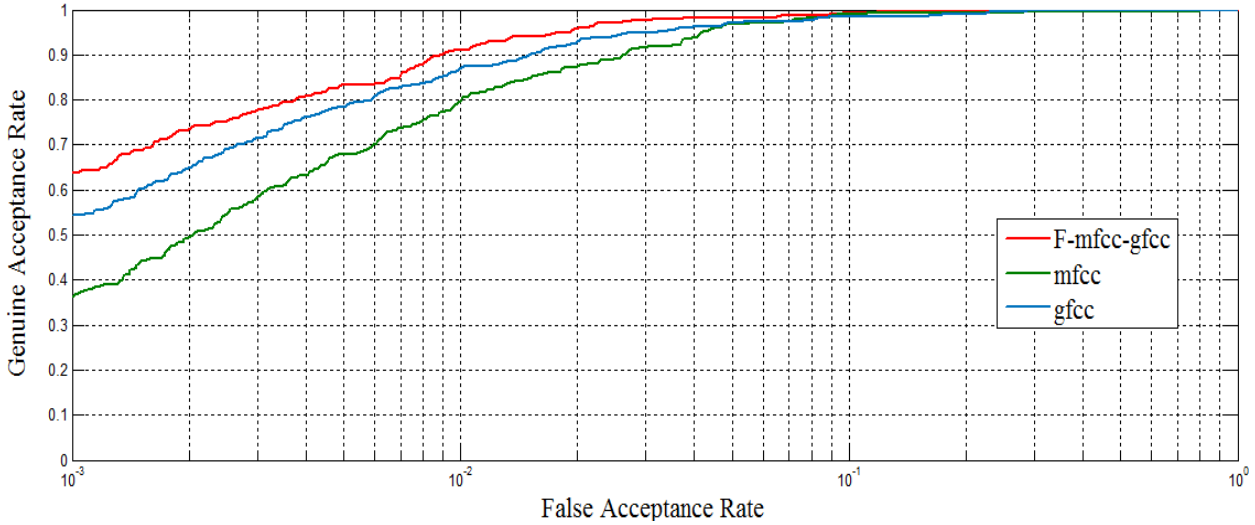


Fig 5.14 ROC Curve at 20dB SNR level (factory noise)

The results in Table 5.6 shows the percent decrease in error rate for the combined system when compared with individual systems. The highest improvement is found against MFCC at clean speech i.e. 78.21% decrease in error rate.

Table 5.6: The improvement in EER as compared to single features

SNR (dB)	% improvement against MFCC	% improvement against GFCC
-5	15.21	7.005
0	78.21	6.56
5	45.34	27.62
10	17.27	12.24
20	35.25	18.07

CHAPTER 6

CONCLUSION AND FUTURE WORK

The unimodal biometric systems containing the complimentary information are combined to get the multimodal biometric system which is free from the problems suffered by single model systems. In the present work, MFCC and GFCC features obtained are modelled and classified separately using GMM and KNN. The two individual systems MFCC and GFCC are thus fused and are compared against the single model systems. Also the comparison is done by adding three types of noise in the test samples. The results are shown for the individual feature sets and combined feature set over a single speech VoxForge database. The proposed combination has shown an improved performance and robustness as compared to individual MFCC and GFCC features under the mismatched and noisy conditions.

The future work may include the improvement of performance by the proper choice of the mixing portions of the two streams and also by the use of other feature combination methods. Features can be combined at feature level, sensor level and decision level other than score level. Also other features can be investigated for combining to get better results.

References

[1] H. Beigi ,“ Fundamentals of Speaker Recognition,” Springer Science & Business Media, December, 2011.

[2] Speaker Recognition, "Available at https://en.wikipedia.org/wiki/Speaker_recognition".

[3] P. Saikia, D. Bora, A. F. Syiemlieh and P. K. Dutta, "Real Time Speaker Recognition System using PCA and ICA," in *Michael Faraday IET India Summit*, Kolkata, 2012.

[4] Roberto Togneri , D. Pullella, "An Overview of SpeakerIdentification:Accuracy and Robustness Issues," *IEEE Circuits and System Magazine*, vol. 11, no. 2, pp. 23-61, 2011.

[5] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," *Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 354-358, 1963.

[6] S. Pruzansky and M. V. Mathews, "Talker-Recognition Procedure Based on Analysis of Variance," *Acoustical Society of America*, vol. 36, no. 11, pp. 2041-2047, 1964.

[7] K.P. Li, J.E. Dammann and W. D. Chapman, "Experimental Studies in Speaker Verification, Using an Adaptive System," *Acoustical Society of America*, vol. 50, no. 5, pp.966-978, 1966.

[8] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *Acoustic Society of America*, vol. 50, no. 2, pp. 637-655, 1971.

[9] G. R. Doddington, "A Method of Speaker Verification," *Acoustical Society of America*, vol. 49, no. 1A, pp. 139-139, 1971.

[10] W. Haberman and A. Fejfar, ""Automatic ID of Personnel through Speaker and Signature Verification - System Description and Testing," *Carnahan Conference on Crime Countermeasur*, 1976.

[11] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 29, no. 2, pp. 254-272, April 1981.

[12] J. M. Naik, L. Netsch and G. Doddington, "Speaker verification over long distance telephone lines," *International Conference on Acoustics, Speech and Signal Processing,glasgow*, vol.

1, pp. 524-527, May 1989.

[13] A. Rosenberg and F. Soong, "Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes," *IEEE International Conference on ICASSP*, vol. 11, pp. 873-876, 1986.

[14] S. Furui, "40 Years of Progress in Automatic Speaker Recognition," *Third International Conference on Biometrics*, Alghero, 2009.

[15] T. Matsui and S. Furui, "Text-independent Speaker Recognition using Vocal Tract and Pitch Information," in *The First International Conference on Spoken Language Processing, (ICSLP)*, Kobe, 1990.

[16] D. A. Reynolds, "A Gaussian mixture modeling approach to text independent speaker identification," Phd Thesis, Georgia Institute of technology, September 1992.

[17] R.C.Rose, J. Fitzmaurice, E. M. Hofstetter, D. A. Reynolds, "Robust Speaker Identification in noisy environments using noise adaptive speaker models," *Proc. Int. Conf Acoust., Speech, and Signal Processing*, vol. 5,no. 3, pp. 137-140,1990.

[18] D. A. Reynolds and R. C. Rose, "An integrated speech-background model for robust speaker identification," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 185-188, March 1992.

[19] E. M. Hofstetter, D. A. Reynolds, R. C. Rose, "Integrated Models of Signal and Background with Application to Speaker Identification in Noise," *IEEE Transactions on Speech And Audio Processing*, vol. 2, no. 2, pp. 245-257, April 1994.

[20] D. A. Reynolds, "Experimental Evaluation of Features for robust speaker identification.," *IEEE Transactions on speech and audio processing*, vol. 2, no. 4, pp. 639-643,1994.

[21] R. C. Rose, Douglas A. Reynolds, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and audio processing*, vol. 3, no. 1, pp. 72-83, January 1995.

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

[23] L.F. Lamel and J. L. Gauvain, "Speaker verification over the telephone," *Speech Communication*, vol. 31, no. 2, pp. 141-154, 2000.

[24] Douglas A. Reynolds, Thomas F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, no. 1, pp. 19-41, 2000.

[25] C. Barras and. J. L. Gauvain, "Feature And Score Normalisation For Speaker Verification of Cellular Data," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* vol. 2, pp. 11-49, April 2003.

[26] Yang Shao, S. Srinivasan and DeLiang Wang, "Incorporating auditory feature uncertainties in robust speaker identification," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* vol. 4, pp. 277-280, 2007.

[27] Roy Patterson, I. N. Smith, J. Holdsworth and P. Rice" An Efficient Auditory Filterbank Based on the Gammatone Function," In *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, December 1987.

[28] Yang Shao and DeLiang Wang, "Robust Speaker Identification Using Auditory Features and Computational Auditory Scene Analysis," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 31 March 2008.

[29] L. Hong, A. Jain and S. Pankanti,"Can Multibiometrics Improve Performance" in *Proceedings of IEEE Workshop on Automatic Identification Advanced Technologies*, New Jersey,USA, 28 October 1999.

[30] A. K. Jain, S. C. Dass, K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Proceedings of International Conference on Biometric Authentication,Springer*, Hong Kong, July 2004.

[31] M. C. Cheung, M. W. Mak and S. Y. Kung, "A two-level fusion approach to multimodal biometric verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. v-485, March 2005.

[32] Karthik Nandakumar, Sarat C. Dass and A. K. Jain, "Likelihood Ratio Based Biometric Score Fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 342-347, 2007.

[33] S. Chakroborty and G. Saha, "Improved Text-Independent Speaker Identification using

Fused MFCC & IMFCC Feature Sets based on Gaussian Filter," *International Journal of Information and Communication Engineering,* vol. 5, no. 1, pp. 11-19, 2009.

[34] Frederic Bimbot, and D. A. Reynolds, "A Tutorial on Text Independent Speaker Verification," *EURASIP Journal on Applied Signal Processing,* vol. 4, pp. 432-451, 2004.

[35] Saurabh Bhardwaj and Smriti Srivastava, "Gfm-based methods for speaker identification," *IEEE Transactions on Cybernetics,* vol. 43, no. 3, pp. 1047-1058, June 2013.

[36] H. P. Combrinck and E. C. Botha, "On the Mel Scaled Cepstrum," Department of Electrical and Electronic Engineering, University of Pretoria, 1996.

[37] Wikipedia, "https://en.wikipedia.org/wiki/Window_function#Hamming_window".

[38] Xiaojia Zhao, Y. Shao and D. Wang, "CASA-Based Robust Speaker Identification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 5, pp. 1608-1616, July 2012.

[39] El Bachir Tazi, A. Benabbou and M. Harti, "Efficient Text Independent Speaker Identification Based on GFCC and CMN Methods," *IEEE International Conference on Multimedia Computing and Systems (ICMCS),* pp. 90-95, May 2012.

[40] Brian R Galsberg,Brian C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research,* vol. 47, no. 1, pp. 103-108, August 1990.

[41] X. Zhao and D. Wang, "Analyzing Noise Robustness of MFCC and GFCC Features in Speaker Identification," *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 7204-7208, 26 May 2013.

[42] J.L. Guavain, C. H Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process,* vol. 2, no. 2, pp. 291-298, 1994.

[43] G. Chaudharya, S.Srivastava, S. Bhardwaj and PreetKiran, "Gaussian membership function based speaker identification using score level fusion of MFCC and GFCC," *Proceedings of the International Congress on Information and Communication Technology*, pp. 283-291, Springer Singapore, 2016.

[44] J. Kacur, R. Vargic and P. Mulinka, "Speaker identification by K-nearest neighbors," Systems, Signals and Image Processing (IWSSIP), Sarajevo, June 2011.

[45] A. Jain, K. Nandakumar, A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, pp. 2270-2285, 2005.

Publications

1) G. Chaudharya, S.Srivastava, S. Bhardwaj and PreetKiran, "Gaussian membership function based speaker identification using score level fusion of MFCC and GFCC," *Proceedings of the International Congress on Information and Communication Technology*, pp. 283-291, Springer Singapore, 2016.

Preet_Thesis

ORIGINALITY REPORT

6%	1%	5%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Advances in Intelligent Systems and Computing, 2016. Publication	1%
2	Frédéric Bimbot. "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal Processing, 2004 Publication	1%
3	Encyclopedia of Biometrics, 2015. Publication	1%
4	Lecture Notes in Computer Science, 2006. Publication	<1%
5	www.lx.it.pt Internet Source	<1%
6	www.sersc.org Internet Source	<1%
7	Reynolds, D.A.. "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, 200001 Publication	<1%
8	ijceronline.com Internet Source	<1%