TEXT-INDEPENDENT ROBUST SPEAKER IDENTIFICATION

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

MASTER OF ENGINEERING

in

Electronic Instrumentation & Control Engineering

Submitted by

Vishu Sharma
Roll No.- 801451028

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)

CERTIFICATE

I hereby certify that the work which is presented in dissertation entitled, "**Text-Independent Robust Speaker Identification**", in fulfilment of the requirements for the award of the degree of **Master of Engineering** in **Electronic Instrumentation and Control**, submitted to Electrical & Instrumentation Engineering Department of Thapar University, Patiala is an authentic record of my own work carried under the supervision of **Dr. Saurabh Bhardwaj**. 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: Patiala

Date: 14 July 2016

(Vishi Sharma)

Roll No: 801451028

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge and belief.

(Dr. Saurabh Bhardwaj)

Assistant Professor, EIED

Countersigned By:

(Dr. Ravinder Aggarwal)

Professor & Head

Electrical & Instrumentation Engg. Deptt.

Thapar University, Patiala

(Dr. S.S. Bhatia)

Professor & Dean

Academic Affairs

Thapar University, Patiala

ACKNOWLEDGEMENT

First and foremost I take the privilege to offer my deepest sense of gratitude to **Dr. Saurabh Bhardwaj**, Assistant Professor, EIED, Thapar University, Patiala for his commendable support and constant motivation throughout this work. With deep humility, I thank him for all the insightful conversations and his valuable time. His guidance has helped me improve my knowledge and perspective towards the work. I will always be indebted.

I am thankful to **Dr. Ravinder Aggarwal**, Professor & Head, EIED for constantly encouraging each student to put their best foot forward in whatever field of work they take up. I express my gratitude to **Mr. Nirbhowjap Singh**, Assistant Professor & PG coordinator for his constant motivation and support.

I would like to extend my sincerest thanks to all the faculty members and staff of Electrical and Instrumentation Department, Thapar University, Patiala, who have bestowed their guidance at appropriate times without which it would have been difficult to proceed my work. I express heartfelt gratitude to my parents, family and friends who have constantly helped me keep my morale high through the work.

Vishu Sharma

(801451028)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NIST - National Institute of Standards and Technology

MFCC - Mel Frequency Cepstral Coefficients

FFT - Fast Fourier Transform

STFT - Short Time Fourier Transform

DCT - Discrete Cosine Transformation

GMM - Gaussian Mixture model

VQ - Vector Quantization

EM - Expectation Maximization

ML - Maximum Likelihood

SVM - Support Vector Machine

SNR - Signal to Noise Ratio

LIBSVM - Library for Support Vector Machine

DB - Decibel

ABSTRACT

As an underlying topic in speaker recognition, speaker identification aims to identify the speaker in a speech sample. Although identification of speaker is a complex problem, with the development of signal processing algorithms the problem has been simplified and the system performs better if the training and testing conditions are identical. However, the real world environment like background noise, room reverberation, crosstalk, etc., degrades the system performance. Achieving robustness in speaker identification has become a major concern ubiquitously. There are number of existing approaches to this problem such as proposing a robust feature set, introducing noise to models created by clean speech and using methods for speech enhancement to restore the clean speech characteristics. This dissertation aims to address the robustness problem in identification of speakers by proposing a new feature set and then introducing noise in the clean speech models. A new methodology is proposed in which root compression based features are extracted and the order of root is increased for analysing the apt value of root that gives the best identification results.

Another methodology extracts the useful information from these feature set, making another set of features which performs substantially better than the conventional speaker features under the influence of noise. The system has been tested under a wide range of signal-to-noise ratio (SNR). The two methods of classification has been used namely, gaussian mixture models and support vector machine.

To justify the results of identification, the proposed methodology have been verified on NIST-2003 and Vox-forge 2015 database in presence of noise. The analysis has been done with improving the order of root in the feature compression stage and the evaluation results have shown that features extracted by taking the square root have outperformed others in a noise dominant system while the cube root compression based features have shown best identification accuracy in case of signal dominant system.

Chapter -1

INTRODUCTION

1.1 Overview

There is thousand years of convention to use body characteristic such as voice, face, gait to recognise one- another. Alphonse Bertillon developed the idea of using body measurements to recognise criminals in 19^{th} century. Based on his idea and practical discovery of the uniqueness of human body based on fingerprints, biometric system was developed. Biometric system is a branch of pattern recognition that uses specific physiological or behavioural characteristic of human to accomplish the personal recognition by machine. The examples of the biometric system includes voice recognition, finger-print recognition, face recognition, palm print and iris recognition. Voice recognition based biometric system has emerged as a viable technology in the recent decade spurred on by advancements in signal processing, algorithms, architectures, and hardware. Also voice can be easily recorded using microphones and hardware of low costs that makes the biometric system simple and economical. Depending on the application, there are two subtypes of voice recognition system, first one is Speech Recognition, that is used to identify the contents of speech signal and other one is Speaker Recognition that establishes the identity of speaker. Speaker Recognition has been used ubiquitously in many applications like security system, login, and forensic science. The speaker recognition is categorised into text-independent and text-dependent speaker recognition on the basis of their dependence on the text utterances. Text independent recognition system means that system should work for any type of text in training and testing phase whereas in other one the text has to be identical in both training and testing phase. This work is focused on text-independent speaker recognition where recognition doesn't depend on text.

The recognition task can be further dichotomized into identification and verification. In identification, one-to-many comparison is performed to establish the identity of person[1]. Relating to speaker identification, the aim is to determine which speech in an unknown group of speech sample matches best with the speaker[2]. While, in speaker verification only one comparison is made to give consent to the claim of identity by user. The verification performs the binary decision task that whether the individual speaking is same he claims to be. In recognition, it is usually done to verify the identity after identification of person is done.

The speaker identification can also be done in two modes of operation based on number of speakers. One is open set and other is closed set operation for identification. In open set mode there can be any number of speakers in system which is limited in case of closed set operation. This work concentrates upon development and analysis of text independent close set operation for identifying speaker.

The canonical speaker identification system is two-step methodology that involves feature extracting followed by signal modeling and feature matching. In feature extraction useful information is extracted from the speech signal by analyzing it. There are many types of features that can be extracted from a speech sample, namely high -level features, prosodic & spectro-temporal features and short-term spectral & voice source features. High level features include phones, semantics, idiolect and accent, are difficult to extract but robust against the effects of noise. Pitch, energy, rhythm, duration, temporal features comes under the Prosodic and spectro-temporal features. Typically extracted features are short term features and vocal tract information is captured by these features.

In humans, vibration of vocal folds generates speech signal, and every individual has its own identical vocal property because of their different voice production organs. This signal is passed further to vocal tract which consist of tongue, palate, pharynx, lips, jaws and nasal tract as its basic components as shown in Figure.1.1.The components of vocal tract acts as filter and its frequency depend upon the resonance of vocal tract, which carries the identity of sound. Thus the frequency components of vocal tract can be used as a feature to ascertain the identity of speaker as well. The normal range of frequency humans can hear is 20 Hz – 20 kHz. These frequency components present in speech signal can be estimated by the frequency-time representation also known as spectrogram of speech signal, thus categorised as Short-term spectral features. There are various methods of feature extraction that will be discussed in the ensuing chapters.

Feature modelling and matching is the next step in identification system, where the features of testing sample are matched to the model database. Gaussian Mixture Model is most commonly used for modelling the feature set. Another method used popularly is Support Vector Machine to classify the data. These techniques for signal modelling and matching will be discussed in this work. There are two major problem of the speaker identification system namely, noise and channel variability.

Noise in a system can be due to internal or external sources, and it distorts the signal. It becomes difficult to identify the person from a noisy speech signal. Another problem is channel variability that is caused due to the variations in the communication channel for the speech signal.

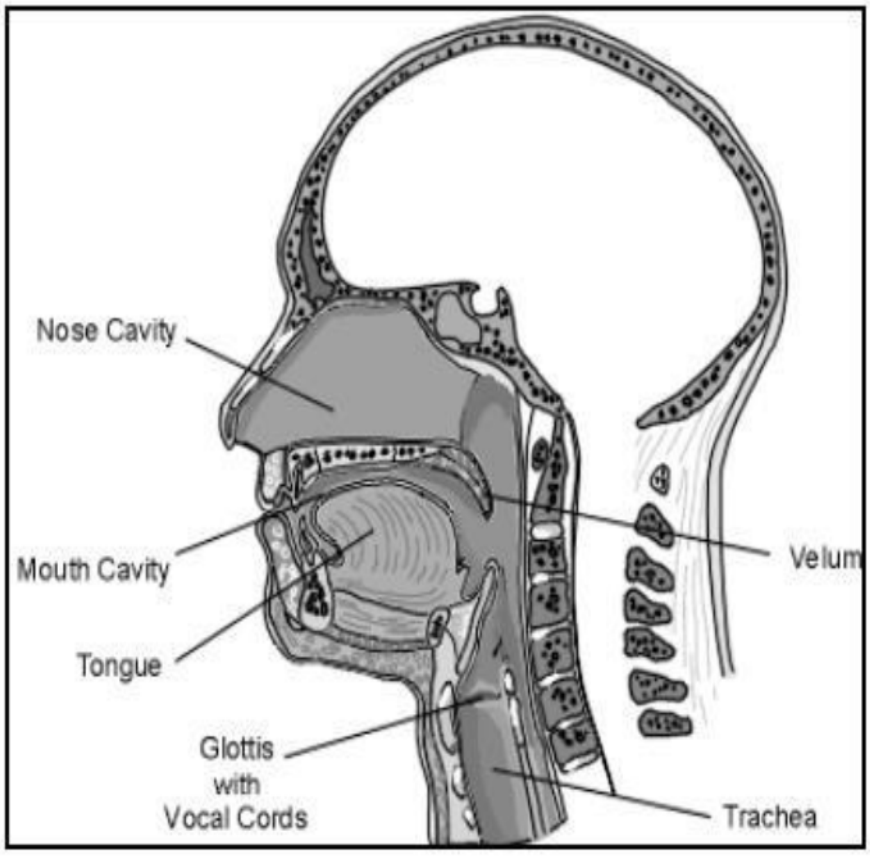

Figure1.1. Human speech production

The main purpose is development of a text independent speaker identification system by extracting a feature set from the available speech sample and then modelling it with different methodologies such that is robust to noisy environments.

1.2 Objective and Motivation

The main motivation throughout the work is to increase the robustness of the system. It has been accomplished by reducing the effects of noise in the process of identification. The entire work is carried out keeping the following objectives into consideration:

- To perform a survey of the work done in the field of speaker identification.

- To understand the basic methodology of speaker identification.
- To develop a new methodology that extracts the robust feature set that aids in enhancement of robustness of identification process.
- To test these feature set with different methodology of classification
- To scrutinize the performance of various methodologies used in this work.

1.3 Organization of the dissertation

The work carried out in this dissertation has been organised in five chapters.

- **Chapter 1** deals with introduction and overview of this work.
- **Chapter 2** briefs about the literature of speaker identification, it also discusses the various existing methodologies for noise robustness in identification process.
- The canonical methodology for speaker identification which includes feature extraction and matching is elucidated in **chapter 3**.
- **Chapter 4** discusses the experiment carried out, results are also discussed in the same.
- Conclusion is drawn in **chapter 5**, it also discusses the future scope of the work.

Chapter -2

LITERATURE REVIEW

Speaker recognition has been a topic of research and development since 1950's, where first known paper in this field was published in 1954 [3] . With the passage of time, the need of speaker recognition system was realised by Pollack, Shearme, et al.[4] and they started comparing speech formants to recognising the speakers. Although at that time the importance was given to the duration of speech signal for recognising speaker.

In 1960, the physiological components of producing speech were modelled by a Swedish professor, Gunnar Fant. This model was based on analysis of X- rays for recognition. During 70s, Fant model was extended by Dr. J. Perkell, tongue and jaw was included in this model using motion x-rays. The speaker recognition system used the analog filters with aid of human expert to perform the matching. Based on this, Texas instruments developed a model prototype that was used by U.S. air force for testing.

With the advancement in biometrics, pattern recognition was applied in the context of speaker recognition in 1963 by Prunzansky, et al.[5], the proposed system had achieved 93% accuracy for ten speakers. The speaker recognition was developed for the legal experts to perform forensic analyses; however, the need for automatic speaker recognition was soon developed. In mid-1980's, NIST Speech Group was developed by NIST to promote the speech processing.

There are many feature extraction techniques available for speaker identification, namely, Linear predictive coding (LPC)[6], Perceptual Linear Predictive Coefficients (PLPC)[7], Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients(MFCC). LPCC and MFCC are the short term features but MFCC is considered as a standard among the various techniques available for speaker identification. In LPCC, the cepstrum is obtained by auto-correlation of speech frame but LPCC coefficients are less effective in our case. MFCC was introduced in 1980 by Davis and Mermelstein [8], the cepstrum coefficients in MFCC are obtained by cosine transformation of the logarithm of the short- time energy spectrum on mel scale.

A practical biometric system should possess certain characteristics like high recognition accuracy, high computational speed, and system should be sufficiently robust to duplicity.

The major technical challenge in the speaker recognition system is effects of variation in communication channel for receiving speech. These variations may be in the form of artifacts like cross-talk, noise and distorts the pattern of feature space and thus leading to increased vagueness in identification. Till date, many techniques are available to deal with variability issue in speaker identification system.

Variants of MFCC such as delta-MFCC and Delta-delta MFCC [9] have been proposed to deal with the variation caused by the noisy environment. Many feature normalisation techniques like Cepstral Mean and Variance Normalisation (CMVN), and Relative Spectra filtering of log domain coefficients (RASTA filtering) [10] have been mused to encounter the effect of noise.

Alternatively, noise was modelled and combined with the clean speech models [11], feature warping [12] was introduced in 2001. Further, score- normalisation techniques like T-norm[13] have been used further on these normalised feature space to compensate noise effects.

The state of art in speaker verification employ the joint-factor analysis[14] or supervector [15] or simply i-vector for modelling speaker and channel characteristics.

Another method was proposed to deal with the stationary noise in signal using fourier transform with Epharim-Malah estimation[16]. This method reduced the noise by compacting the spectral coefficients.

Adaptive wiener filter[17] has also been used to accommodate noise with varying speech signal. In addition, to aforementioned methods, wavelets[18] was introduced in signal processing to diminish effects of noise and applied to speaker recognition application[19]. GFCC[20] was introduced to increase the robustness of system in presence of noise that employed the cubic root based rectification . [21] clearly shows that cubic root used in the derivation of scale invariant coefficients has provided robust features to encounter noise.

Another challenge in speaker identification is competing voice in real environment. In such case, the interfering speech signal usually exhibit similar traits as the target speech, which makes the classification difficult. When a single communication channel is used to record two voices simultaneously, the speech is called co-channel speech and the identifying speaker in these conditions is called co-channel speaker identification. Starting with [22] , many

methodologies were proposed, that extracted the usable speech from the target speech. Multi-pitch tracking algorithm was employed for usable speech identification[23].

GMM[24] is dominant technique for modelling in text independent recognition system. In this type of model, first a model is created for each subject by clustering the multi-dimensional data using the probabilistic based approach and then a test sample with multi-dimensional features is tried to adapt into the respective subject's model with the highest likelihood. Although, neural network has been used for classification, but not significantly.

In machine learning, many algorithms have been used for classification, before 1960's all these algorithms were based on the linear decision surface. The implementation of piece-wise linear separating surface became feasible with introduction of perceptron in 1962 by Rosenblatt[26]. For adapting the weights of neural networks to minimise the error, back propagation algorithm was introduced [27].

A new algorithm was developed by Vapnik and his group [28], that resolved the problem of choosing the decision boundary for classifying binary data. The introduction of Support Vector Machine (SVM) [25] has eased the classification of multi-dimensional data. Also improved SVM can be used for multi-class classification and regression analysis of data.

Many approaches for speaker identification have been proposed that uses SVM. It has been used in Hidden Markov Models for probability modelling[29].

The major problem with this approach is long training times and SVM did not gave the probability estimates[30]. So, in another approach SVM was combined with GMM in speaker recognition application. Although this work restricts to usage of GMM and SVM separately on features extracted from speech sample. The primary objective is to make the system robust in noisy environment.

Chapter -3

METHODOLOGY FOR SPEAKER IDENTIFICATION

The basic identification system is depicted in Figure.3.1. The canonical speaker identification is executed in two phases, training and testing phase. Features are extracted from the speech samples and modelled using some modelling techniques during training phase. These models are stored as speaker database or speaker models. Whenever there is an unknown speech sample for testing, features are extracted for this unknown speech sample using the same methodology for feature extraction as in training phase. These testing features are now matched to the speaker model created earlier in the training phase.

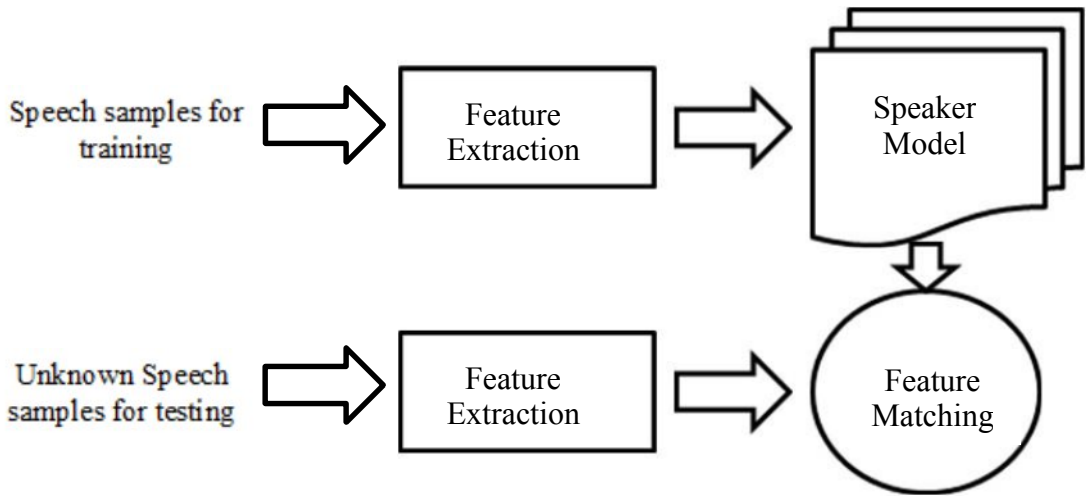

Figure 3.1. Methodology for Speaker Identification

The feature extraction and feature matching techniques are elucidated in the subsequent sections of this chapter.

3.1. Feature Extraction

Feature Extraction is first step of speaker identification system in which useful information from speech is extracted, discarding the redundant and unwanted information from the signal.

Feature extraction methodology involves transforming signal into appropriate form, for models used for classification. However, in this process some useful information may also be lost, so feature set should be chosen carefully.

A few properties of features is listed that would be desirable for speaker identification:

1) Variability among speakers should be low.
2) High discrimination among the classes
3) The feature set should be invariant to degradations in speech due to variability in channel and noise distortion.

The objective is to identify the set of properties from the utterance that correlates to the speech signal, that is, properties can be somehow estimated through signal processing. Such parameters are called features. Extracting features typically involve signal conditioning that is followed by measurement of certain parameter and then augmenting the measurements with derived measurements. Statistically conditioning these parameters is next step in feature extraction.

Many feature extraction techniques available are discussed in the preceding chapter. MFCC have been used as the standard method for extracting features from the speech signal. In this work, another methodology is projected that modifies the MFCC features by using the root based feature compression instead of log based feature compression. Also, another feature set where, information is extracted from standard MFCC. All these feature extraction methodologies are discussed next.

3.1.1. MFCC

MFCC is the standard methodology for extracting the features from speech sample. These coefficients are based on the cepstral analysis, the signal information from glottal excitation and vocal tract are separated from each other. The cepstrum, is given by inverse Fourier transformation of log magnitude of Fourier transformed signal. The cepstrum coefficients in MFCC are obtained by cosine transformation of the logarithm of the short- time energy spectrum on Mel scale. The basic MFCC procedure is explained in Figure.3.2.

The short utterances of speech signal, as shown in Figure.3.3, results in a large database, therefore, a large number of frequency components in the signal. It is practically challenging to directly identify the speaker on the basis of its frequency components. To avoid this

problem, speech signal is sampled at frequency of 8 kHz to reduce the no of samples, permitting a significant data reduction.

The signal is further passed through a high pass filter to make the signal comparable to noise. The short time Fourier transform of signal is then examined on this signal, which is second block of the MFCC feature extraction mechanism.

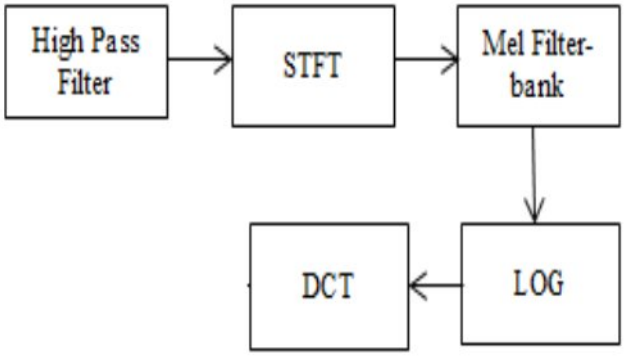

Figure3.2.MFCC feature extraction

The filter-bank energy of this signal is then obtained, which is compressed by the logarithm function and further cosine transformed to get the MFCC. Each step in methodology is elucidated in the following section.

Filtering

The first stage in MFCC feature extraction is filtering, where speech signal is passed through a high pass filter after being sampled at a frequency of 8 kHz. It is also called the pre-emphasis stage. Mathematically, filter is represented by (1), where x_n is samples and α is the pre-emphasis coefficient.

$$y_n = x_n - \alpha * x_{n-1} \quad (1)$$

where $0 < \alpha < 1$

The motive of this stage is to focus on the high frequency components to avoid the problem of spectral tilt. A spectrum of speech signal has more energy at low frequency as they have small amplitude at higher frequency in comparison to low frequency, which is because of the natural property of glottal pulse; thus the spectrum becomes tilted. Boosting high frequency

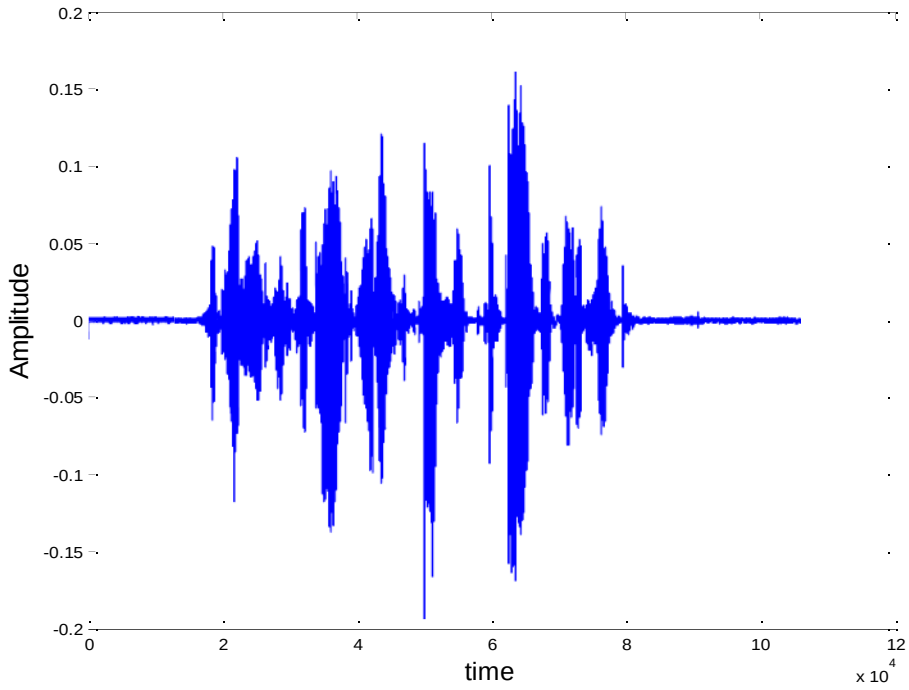

Figure.3.3. Original Speech Signal

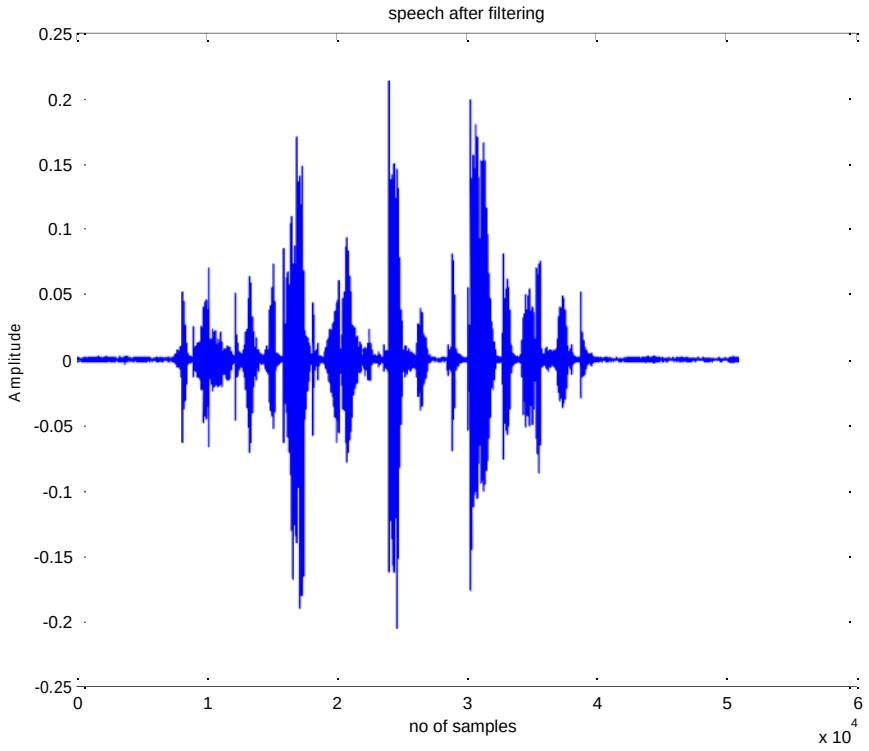

Figure.3.4. Filtered signal

components provides more information to the acoustic model [27], thus improving identification performance in terms of accuracy. Also, the signal after being filtered by a high pass filter becomes comparable to noise.

STFT

The emphasised signal thus obtained is framed into short time intervals of 20ms with a frame shift of 10ms. The frames considered are overlapping to avoid any information loss and speech signal is presumed to be stationary in this frame. The size of this frame should be chosen wisely as a shorter frame will not have enough samples to give a reliable estimate of the spectrum and in case of wider frame the signal will be no longer stationary. The frames so obtained are weighted by a windowing function given by (2), and hamming window is chosen as windowing function as it provides less stop band ripple.

$$\text{cos}(\quad = \{0.54 - 0.46 \quad \frac{2?(\blacklozenge 1)}{(\blacklozenge 1)}) \} \blacklozenge \quad (2)$$

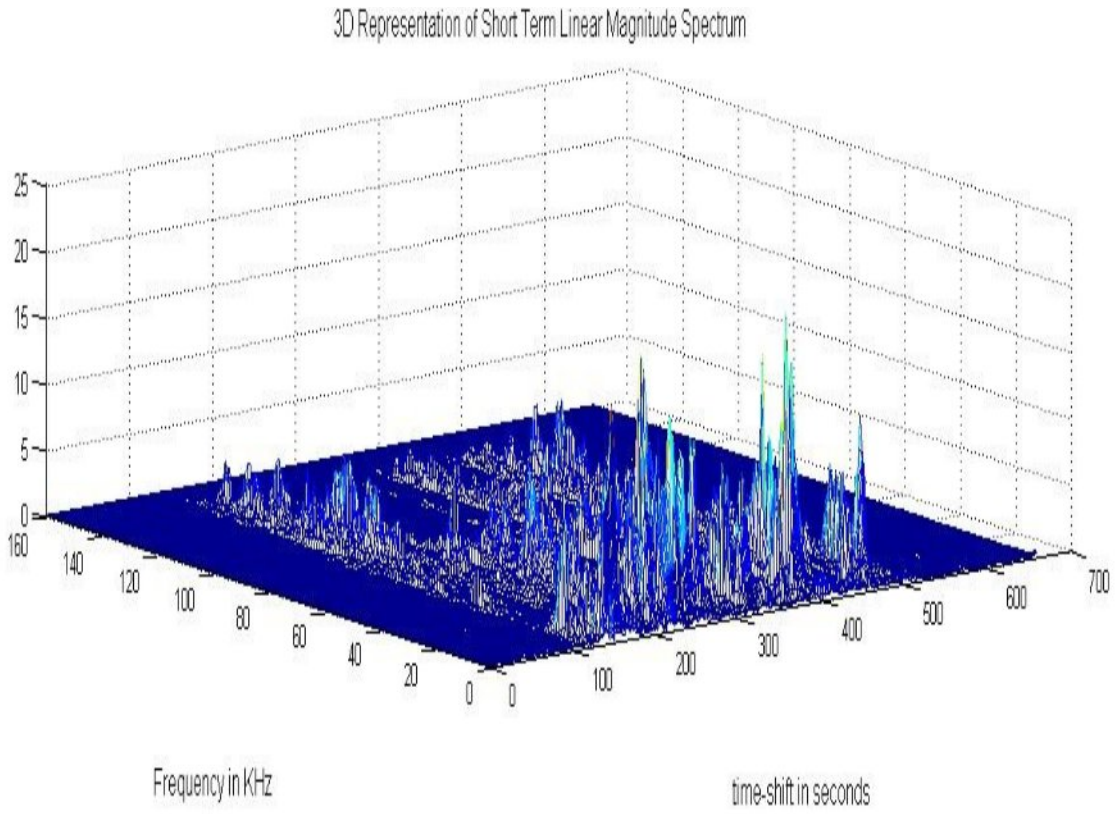

Figure3.5. 3D Representation of Short Term Linear Magnitude Spectrum

The magnitude of FFT of windowed signal is taken to get the spectral estimate of the speech signal in frequency domain. FFT algorithm is used to fastly compute discrete fourier transform (DFT) of signal $x(n)$ is given by (3),

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn} \quad (3)$$

Figure.3.5, shows 3-D representation of STFT of the speech signal that shows Frequency-time representation of the signal.

Mel filter bank

Subsequent step in the method is passing the signal through filter-bank. The design of filter-banks is on the basis of human perception of frequency, as human hearing is more sensitive to lower frequency bands. We can't discern the high frequency bands above 1 kHz thus triangular filters are placed on mel scale, given by (4), which is designed in such a way that it varies linearly below 1 kHz and approximately logarithmically above 1 kHz.

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \quad (4)$$

Motivated by standard behaviour of hair cells in cochlea of human ear, triangular filters are placed such that it has more no of filters in low frequency region in comparison to high frequency region [30], as shown in Figure.3.6.

The filter- bank energies of this signal is calculated and is sent to the next block of feature extraction as discussed next.

Logarithm

The logarithm is used for dynamic range compression of the filter-bank energies obtained in previous step; this makes the frequency estimates to be less sensitive to little variations in input [27]. Also the phase information is of no use in speaker identification task and therefore not considered in our case.

A big shortcoming of the logarithmic compression is that if the value of feature goes 0, the value of logarithm actually becomes negative infinity. But as per the standard process of speaker identification system using MFCC, the features are compressed with log function.

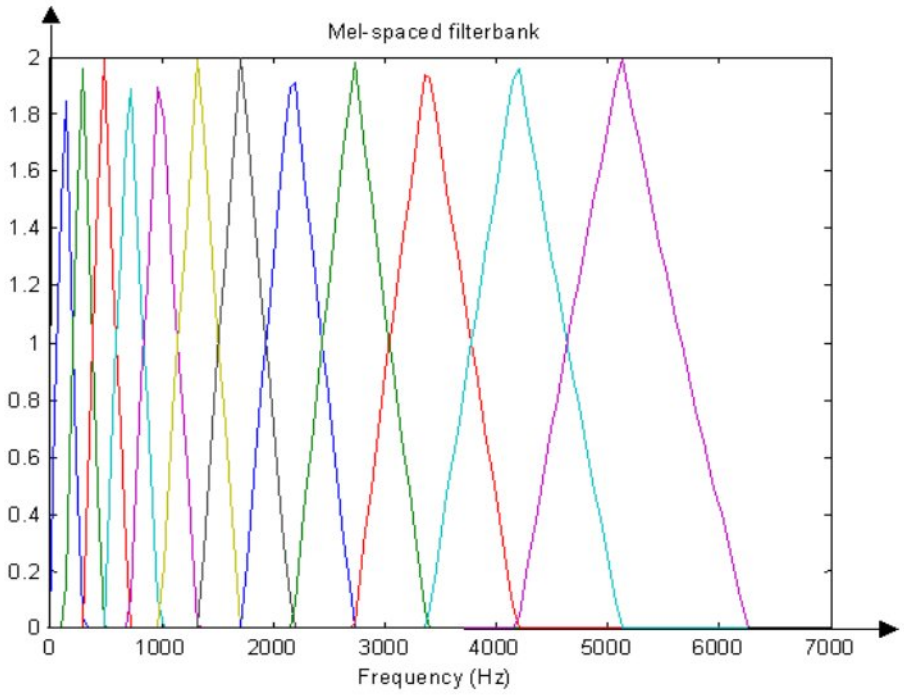

Figure3.6. Mel filter-bank

Discrete Cosine Transform

The compressed features obtained are overlapping, so to remove the information redundancy DCT is performed. Log-compressed filter-bank energies are decorrelated using DCT to produce the cepstrum coefficients given by (5)

$$?_{\square} = \sqrt{\frac{2}{\square}} \sum_{\square=1}^{\square} \square_{\square} \cos\left(\frac{?\pi}{\square}(\square - 0.5)\right) \quad (5)$$

The result of this conversion is Mel frequency Cepstrum coefficients.

3.1.2. Root Compressed Features

Any transformation on data is usually done to make the analysis of data easier. There are few desirable properties of any transformation, as given by [32], that makes the analysis of data easier.

Some of them are:

1) The effects on data should be additive,
2) The error variability should be almost constant and
3) The error distribution should be nearly normal.

In this methodology, the n^{th} root of the filter- bank energies is taken instead of logarithmic compression. A big shortcoming of the logarithmic compression is that if the value of x goes 0, the value of logarithm actually becomes negative infinity, while the values are mapped onto much smaller values in case of the roots. The compression based on cubic root is found to give more robust results as discussed in [21]. So in this work, another analysis is done by increasing the order of root and comparing the results for various levels.

Figure.3.7, compares the logarithm compression with root compression by increasing the number of root. Clearly, as no of roots increases beyond seven, the change of scale is negligible, so we have limited the no of roots to eight in analysis.

3.1.3. Information Extracted Features

The objective of this technique is to extract meaningful information from the feature vector obtained, as the size of feature vector is very large. In order to compact information as well as to reduce the computational complexity information is extracted from these feature vectors.

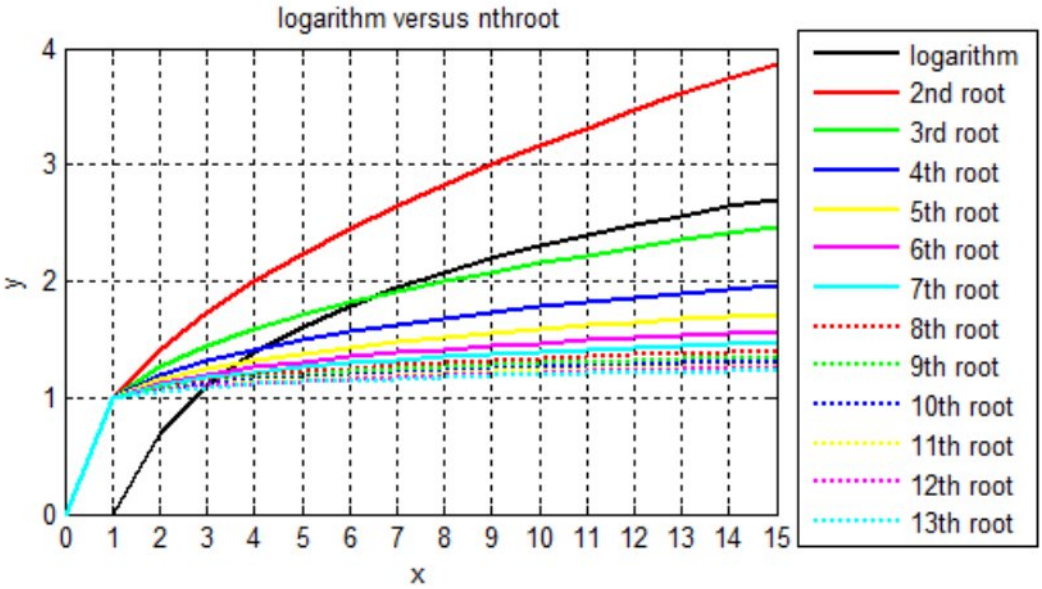

Figure3.7. Graph of nth order root with log

The basic idea is to fuzzify the information which is available in form of mean and variance obtained in row and column wise manner from feature set.

Each row of MFCC matrix represents the pseudo frequency bin (Cepstral coefficient) and each column represents a set of Cepstral coefficients (temporal frame). We consider each element of MFCC matrix ? as information source value $?_{ij}$, the columns of this matrix provide the temporal information rows provide the spatial variation of the signal. The MFCC feature vector obtained is given by X (�× ?) where a is dimensionality of feature set and b is no of frames.

$$? = \begin{bmatrix} ?_{11} & \cdots & ?_{1j} & \cdots & ?_{1b} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ ?_{i1} & \cdots & ?_{ij} & \cdots & ?_{ib} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ ?_{a1} & \cdots & ?_{aj} & \cdots & ?_{ab} \end{bmatrix}_{?\times?}$$

Algorithm:

Inputs: X: Feature vector set

Output: $?_?$: New Feature Vector Set

1) Compute mean ($?_?$) given by (2) and variance ($?_?$) given by (3) of frame $?_?$

$$?_? = \frac{1}{?}\sum_{?=1}^{?} ?_{??} \qquad ? = 1, \ldots, ? \tag{6}$$

$$?_? = \sum_{?=1}^{?} (?_{??} - ?_?)^2, \qquad ? = 1, \ldots, ? \tag{7}$$

Where $?_? = \begin{bmatrix} ?_{1?} \\ \vdots \\ ?_{??} \\ \vdots \\ [?_{??}] \end{bmatrix}, \qquad ? = 1, \ldots, ?$

2) The next step is to fuzzify the mean and variance on the data. For this Gaussian membership value for each value in j frame is computed and given by (4)

$$\frac{1}{16}\left(\frac{?_{??} - ?_?}{}\right)^2$$

(8)

The membership value obtained is then multiplied with the input feature value at that point as given by (5).

$$?(??) = ?? \times ? ?(??), \quad ? = 1, \dots, a \;, j = 1, \dots, b \qquad (9)$$

3) Then, compute mean ($??$) given by (6) and variance ($?_?$) given by (7) of $?_?$ which is obtained by taking $?_? = [?? \quad \cdots \quad ?? \quad \cdots \quad ??], \quad ? = 1, \dots, ?$

$$?? = \frac{1}{?} \sum_{?=1}^{?} ?? \qquad ? = 1, \dots, ? \qquad (10)$$

$$?_? = \sum_{?=1}^{?} (?? - ??)^2, \qquad ? = 1, \dots, ? \qquad (11)$$

4) The mean and variance so obtained are fuzzified on a Gaussian membership function given by (8) for each value in i^{th} dimension,

$$?? ?(??) = ? \; ^{-\frac{1}{2}\left(\frac{?? - ?}{?}\right)^2}, \qquad ? = 1, \dots, a \qquad (12)$$

The membership value obtained is then multiplied with the input feature value at that point as given by (9)

$$?(??) = ?? \times ? ?(??), \qquad ? = 1, \dots, a$$
$$j = 1, \dots, b \qquad (13)$$

5) The information obtained row-wise and column-wise is fused to give the effective information given by (10)

$$? = \frac{1}{2} \sum_{?=1}^{?} \{?(??) + ?(??)\}, \; ? = 1, \dots, ?$$
$$j = 1, \dots, b \qquad (14)$$

These features contain the cepstro-temporal information useful for discriminating the subjects. It is not possible to reconstruct the signal from this effective information but this holds the discriminating information of the subjects.

3.2. Feature Matching

Feature matching is next stage of identification process, where models created at training phase are compared to the testing speech sample.

There are two types of models that are widely used in speaker identification system:

1) Stochastic Models
2) Template Models

The stochastic models are based on the probability theory that presumes speech production as a random process. It also assumes that the parameters of this type of model can be estimated recisely. Therefore, it is also called parametric model. Whereas, the template models generate the model for speech production in a non- parametric manner and is free from any assumption for data generation. In the early work, the template models used to dominate but in recent work stochastic models have been used extensively due to its flexibility and generation of better models for identification.

Many feature matching algorithm have been used for feature matching namely, Dynamic Time Wrapping (DTW), Vector Quantisation (VQ) and GMM. VQ uses a unique style of representing each speaker by mapping vectors from a large space to finite no of regions in the space. Each region is represented by its centre called code word and is also termed as cluster. When VQ is completed, only a few vectors are left and collectively termed as speaker's codebook. This codebook is used for testing speaker in system.

However, GMM have been used more popularly than any other stochastic process of modelling due to certain advantages as discussed later. Also support vector machine has proved to provide good results in identifying person. The MFCC or root compressed feature set can't be used directly for classification using SVM, therefore, information is extracted from these feature set as discussed in the preceding section. SVM is applied on these new feature set. These two feature matching techniques are discussed in next section.

3.2.1. Gaussian Mixture Model

The Gaussian mixture speaker model was developed by D. Reynolds in late 90's [23]. It has manifested high identification accuracy for classifying short utterances of test signal from unobstructed speech.

Gaussian is a characteristic curve that is bell shaped and dies out quickly to zero practically. A gaussian curve is given by (15) and shown in Figure.3.8.

$$?(\blacklozenge) = \frac{1}{2??}\, \blacklozenge^{\frac{1}{2}\frac{?-?^2}{(\ ?\ )}} \qquad (15)$$

Mixture model is a probabilistic model that considers that the data underlying belong to the distribution of mixture. It can be given by the weighted sum of probability density function

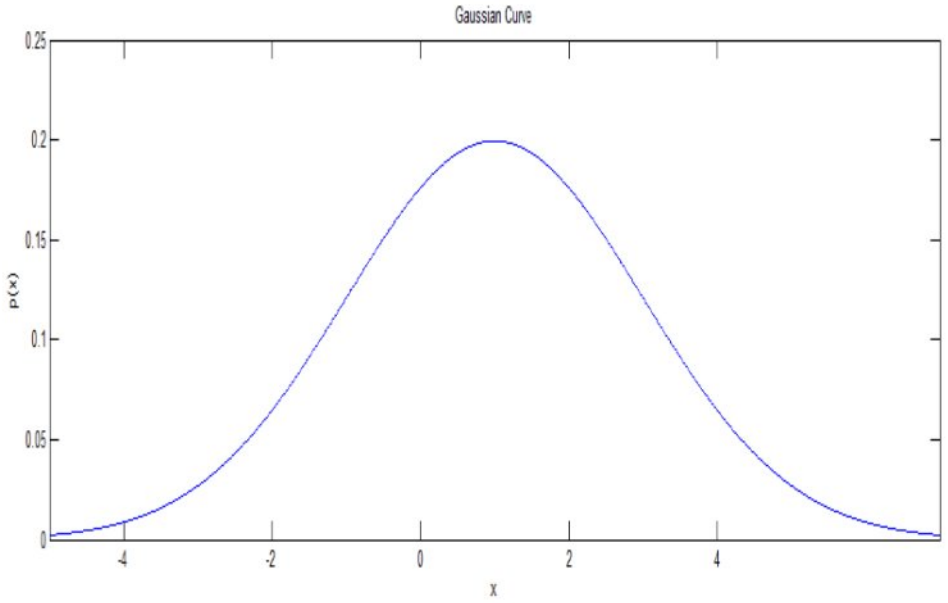

Figure 3.8.Gaussian Curve

$$p(x) = \pi_0 p_0(x) + \pi_1 p_1(x) + \pi_2 p_2(x) \ldots\ldots\ldots\ldots + \pi_k p_k(x) \quad (16)$$

or $$p(x) = \sum_{i=0}^{k} \pi_i p_i(x)$$

where, $$\sum_{i=0}^{k} \pi_i = 1$$

GMM is a commonly used technique for clustering of data. The assignment of components in clusters is done on the basis of posterior probability. The gaussian component that gives maximum value of posterior probability is selected as part of cluster. This type of clustering is also called soft clustering method. The posterior probability at each point show that every point in data has probability that it belongs to cluster.

GMM is basically a parametric probability density function represented as a weighted sum of densities of its multivariate Gaussian components [33], as shown in Figure.3.9., where each component has its own mean (μ), covariance (Σ) and weight (w).

Generalized GMM model is as shown in Figure.3.10, where density model for a speaker is represented by λ (p_i, μ_i, Σi) and mixture density used for likelihood is defined as (17)

$$p(x) = \sum_{i=0}^{M} w_i p_i \quad (17)$$

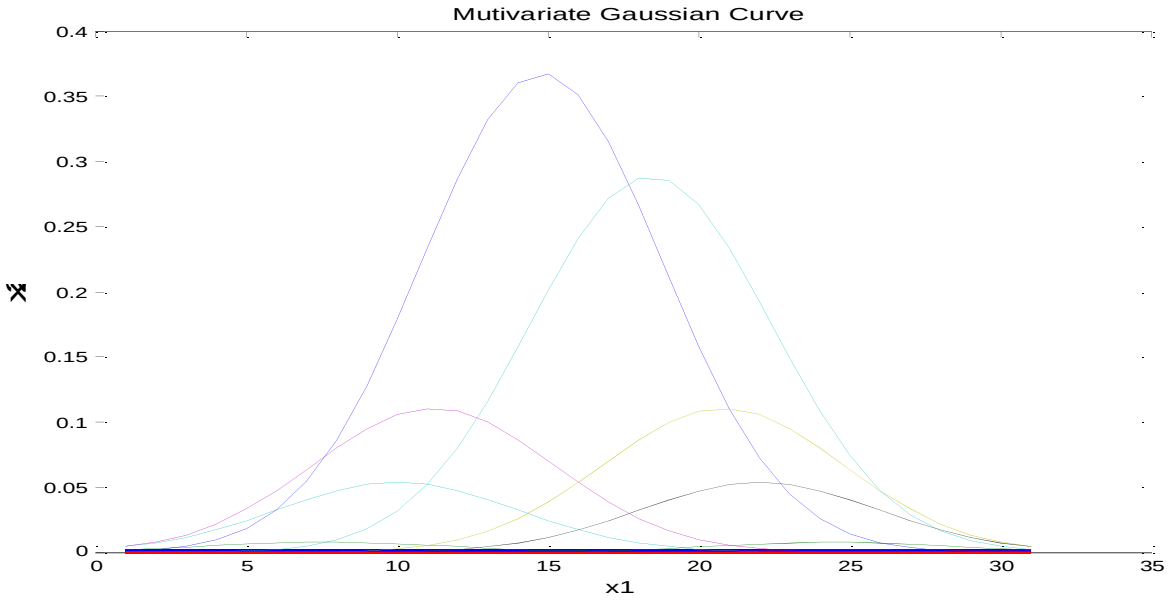

Figure 3.9. Multivariate Gaussian Curve

where w_i are mixture weights and M are the no of gaussian components used to represent the data. The density function is a weighted linear combination of M unimodal Gaussian densities, $p_i(x)$, each parameterized by a mean vector μ_i, and covariance matrix Σ_i given by (18) [33]

$$p_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_i)'(\Sigma_i)^{-1}(x - \mu_i)\right\} \quad (18)$$

While full covariance matrices are supported by the general model, only diagonal values of covariance is considered for analysis, as discussed in [33]. There are three major reasons for selection of diagonal covariance matrix. The primary reason is that the density modelling can be achieved equally using diagonal and full covariance matrix. Also the system with diagonal covariance are computationally effective than the one with full covariance matrix. Third, it is empirically observed that the system with diagonal matrix has performed better than the one with full matrix GMM.

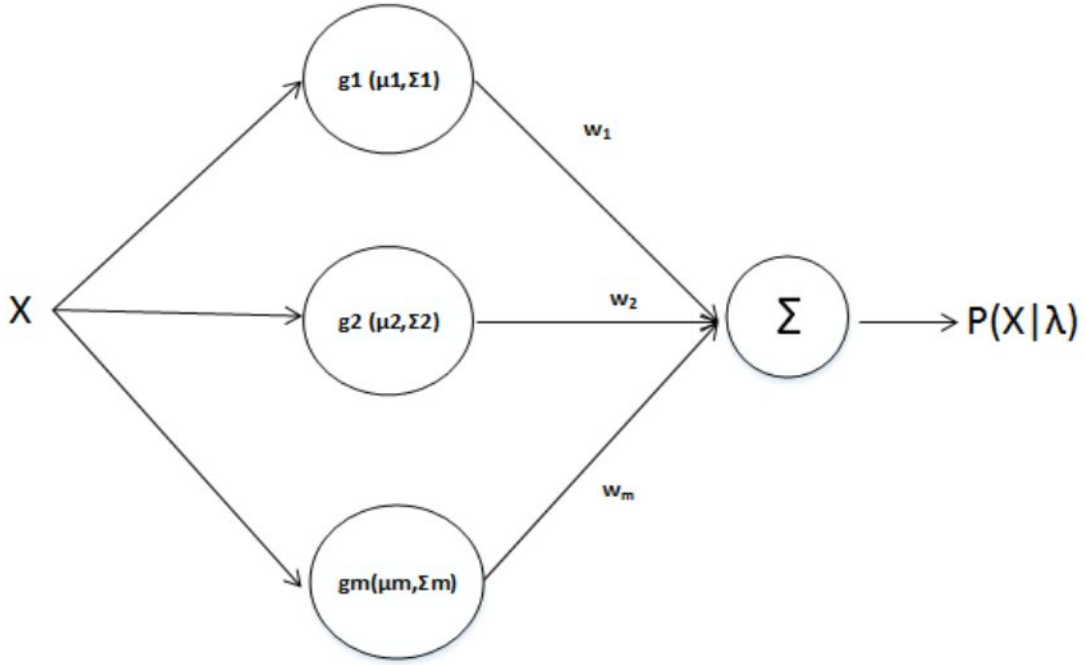

Figure3.10. Generalized GMM model

There are many techniques available for parameter estimation in GMM [35]. Maximum likelihood (ML) algorithm is popular and well established method to find the parameter that matches best with the distribution of feature vector under training in GMM. Given training vector set X={x1, x2….., x_T}, the GMM likelihood can also be defined as (19)

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \quad (19)$$

As the expression is non-linear, it is not possible to directly maximise the parameters. However parameter estimates can be found by using ML as a special case of EM algorithm (Dempster et al., 1977)[34]. The parameters of GMM are refined using iterative EM algorithm such that the maximum value of likelihood is achieved, as given by (20)

$$p_m(X \mid \lambda(k+1)) > p_m(X \mid \lambda(k)) \quad (20)$$

where the likelihood of $(k+1)^{th}$ iteration is greater than k^{th} iteration. The re-estimation formulas for the values of mixture weights, mean and covariance, that assure an increase in likelihood value of model are given by (21),(22), (23) respectively .

$$\bar{w}_i = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|x_t, \lambda) \tag{21}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} \Pr(i|x_t, \lambda)}{\sum_{t=1}^{T} \Pr(i|x_t, \lambda)} \tag{22}$$

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} \Pr(i|x_t, \lambda) x_t^2}{\sum_{t=1}^{T} \Pr(i|x_t, \lambda)} - \bar{\mu}_i^2 \tag{23}$$

where, posteriori probability for i^{th} component is given by (24) ,

$$p_i(i|x_t, \lambda) = \frac{w_i p_i(x_t|\mu_i, \Sigma i)}{\sum_{k=1}^{M} w_k p_k(x_t|\mu_k, \Sigma k)} \tag{24}$$

GMM is commonly used in biometric system due to its capability to represent a large no of classes in distribution. GMM is based on the statistical model that makes the system computation inexpensive and also oblivious to the temporal aspects of speech signal [2].

After modelling the various gaussian densities, the testing of speech signal is carried out on this model to get the identification accuracy.

3.1.2. Support Vector Machine

Support Vector Machine (SVM) was developed by Vapnik and his group at A T &T Bell laboratory[25], that resolved the problem of choosing the decision boundary for classification of binary data. A no of decision boundary are feasible in a linear perceptron problem, the major problem in these algorithms is to decide the apt decision boundary, as shown in Fig.3.11.

Here, the basic idea of classification is mapping input vector into high dimensional feature

space[23]. It is done through some non-linear mapping functions.

Consider training patterns with set of labels given by

$$(\vec{x}_1, y_1),\ldots..,(\vec{x}_n, y_n), \quad y_i \in \{-1,1\} \tag{25}$$

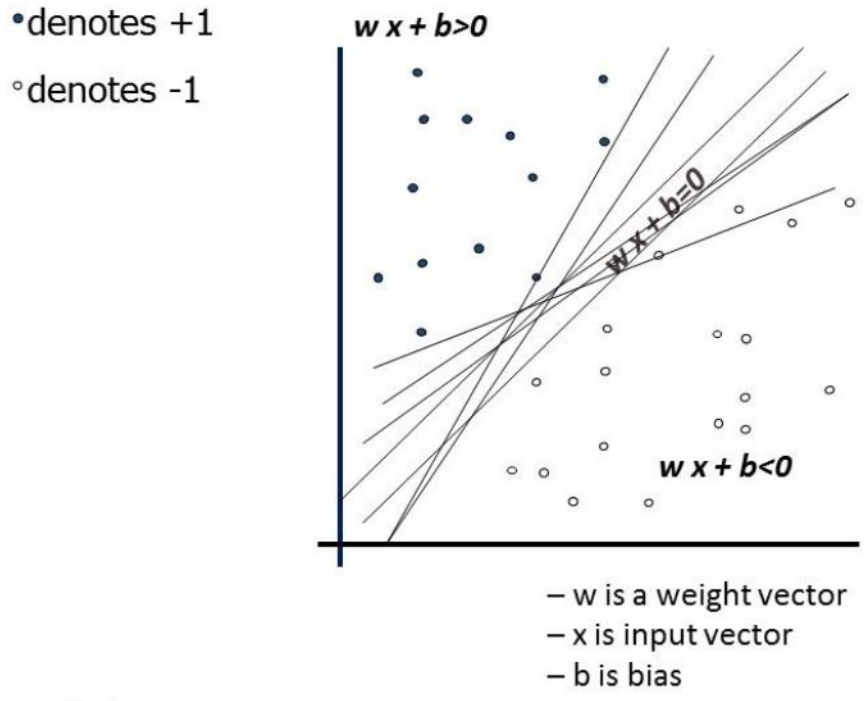

Figure 3.11. Possible Decision boundary with perceptron model

The training patterns are known to be linearly separable for vectors $\vec{w}$ and b such that the following inequalities (26) hold for all training set elements.

$$\begin{aligned} \vec{x}_i \cdot \vec{w} + b &\geq 1 \qquad y_i = 1 \\ \vec{x}_i \cdot \vec{w} + b &\leq 1 \qquad y_i = -1 \end{aligned} \tag{26}$$

Re-writing (26) ,

$$y_i(\vec{x}_i \cdot \vec{w} + b) \geq 1, \qquad i = 1, \ldots\ldots, n \tag{27}$$

This algorithm builds optimal hyperplane (28) in high dimensionality space which is selected such that it ensures maximum margin to adjoining data points. Linear SVM is as shown in Figure.3.12. This large margin ensures the lower generalisation error of classifier. Vectors for which $(\vec{x}_i \cdot \vec{w} - b) = 1$, are termed as support vectors.

$$w_0 x + b_0 = 0 \tag{28}$$

The margin, ? is given by
(29)

$$\rho = \frac{2}{|w|} \tag{29}$$

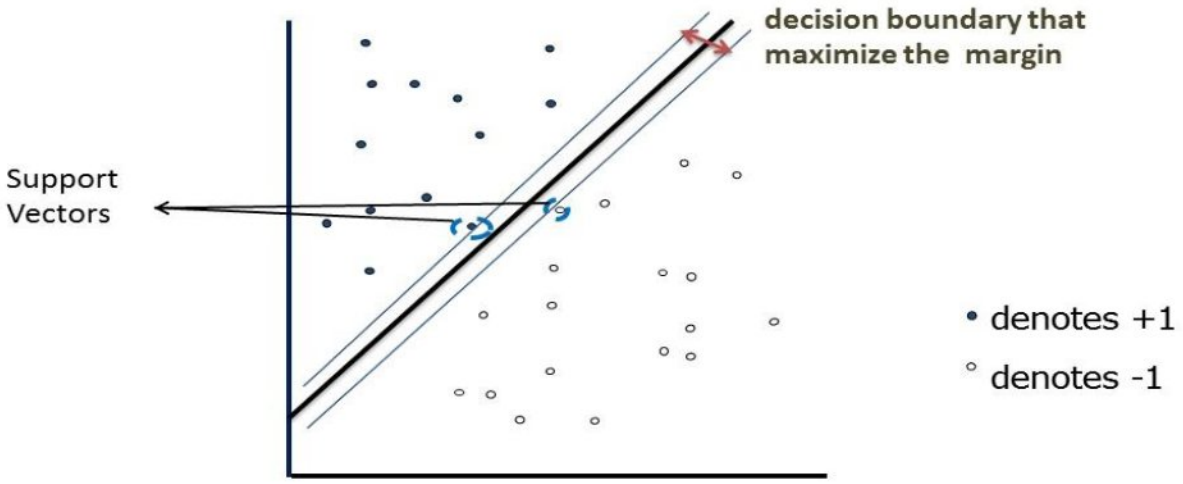

Figure 3.12. Linear SVM

he optimal hyperplane minimises w under conditions given by (27) . Quadratic optimisation technique is employed to identify the solution of optimisation problem, however, in quadratic programming it is easier to unfold dual problem and solution of which involves constructing a dual problem (30), where positive Lagrange multiplier ($?_i$) is associated with each constraint given by (31)

$$L_D = \frac{1}{2}\|w\|^2 - \sum_i ?_i(y_i(w \cdot x_i + b) - 1); \tag{30}$$

where , $w = \sum_i ?_i y_i x_i$ (31)

$$\sum_i ?_i y_i = 0$$

Substituting into (30) the values of constraints

$$L_D = \sum_i ?_i - \frac{1}{2}\sum_i \sum_j ?_i ?_j y_i y_j x_i \cdot x_j \tag{32}$$

which is maximised over $?_{\blacksquare} \geq 0$. The data points corresponding to non-zero values of $?_{\blacksquare}$ are called support vectors.

Classifying function is now given by (33)

$$\blacksquare\blacksquare = \sum_{\blacksquare}^{?} \blacksquare\blacksquare\blacksquare\blacksquare^{?}\blacksquare + \blacksquare \tag{33}$$

However, the data is not always linearly separable and in that case SVM uses soft margin technique that develops a hyperplane that separates many data points but not all. The standard formulation of soft margin involves adding variables *C* and $\in_{\blacksquare}$ given by (34) subjected to condition (35)

$$\min_{\in} \frac{1}{2}\blacksquare^{2}\blacksquare + ?\sum_{\blacksquare=1}^{\blacksquare} \blacksquare \tag{34}$$

subject to

$$\blacksquare(\blacksquare^{2}\blacksquare_{\blacksquare} + \blacksquare) \geq 1 - \in_{\blacksquare}$$

$$\in_{\blacksquare} \geq 0 \tag{35}$$

The constraint given by (35) can be more concisely written as

$$\blacksquare\blacksquare(\blacksquare\blacksquare) \geq 1 - \in_{\blacksquare} \tag{36}$$

Combining (34) with $\in_{\blacksquare} \geq 0$

$$\in_{\blacksquare} = \max(0, 1 - \blacksquare\blacksquare(\blacksquare\blacksquare)) \tag{37}$$

Hence, the problem now becomes equivalent to unconstrained optimisation problem over w, given by (38)

$$\min_{\blacksquare ?\,?} \|\blacksquare\|^{2} + ?\sum_{\blacksquare}^{\blacksquare} \max(0, 1 - \blacksquare\blacksquare\blacksquare\blacksquare) \tag{38}$$

where, $\|\blacksquare\|^{2}$ is regularisation parameter and $\max(0, 1 - \blacksquare\blacksquare\blacksquare)$ is loss function. The parameter C can be used to minimise the error because of Overfitting.

For easier calculations, dual problem is considered to this soft- margin formulation. Using the Lagrange multipliers $?_{\blacksquare}$ the final dual formation is given by (39)

$$\max_{?} \sum_{\blacksquare}^{?} \blacksquare - \frac{1}{2}\sum_{\blacksquare}\sum_{\blacksquare}^{?}\blacksquare\,?\,\blacksquare\,\blacksquare\blacksquare\blacksquare\blacksquare^{?}\blacksquare \tag{39}$$

subject to

$$\sum_{i} ? \, y_i = 0$$

$$0 \leq ? \, \alpha_i \leq ? \tag{40}$$

For non- linear transformations, there is a simplified approach to separate the hyperplanes Furthermore, $K(x_i, x_j)$ is kernel function which maps the training feature set x_i into higher dimensional space by function Θ and is defined by (41)

$$K(x_i, x_j) = \theta(x_i)^{?} \theta(x_j) \tag{41}$$

Four basic kernel functions $K(x_i, x_j)$ that can be used are as shown in Table 3.1. [35]. After deciding the kernel function, the kernel parameters ?, d, r are chosen, grid method to search the parameters is propitious to improve the accuracy of identification as selection can be parallelized, parameters being independent. Library for Support vector machine (LIBSVM)[36], is used for the implementation of the classifier.

Table 3.1. Kernel functions

Linear	$K(x_i, x_j) = x_i^{?} x_j$
Polynomial	$K(x_i, x_j) = (? x_i^{?} x_j + r)^{d}$, $? > 0$
Radial Basis Function (RBF)	$K(x_i, x_j) = \exp(-? \lVert x_i - x_j \rVert^{2})$, $? > 0$
Sigmoid	$K(x_i, x_j) = \tanh(? x_i^{?} x_j + r)$

Chapter -4

Results and Discussion

4.1. Database

The experiment was performed on standard Database namely NIST-2003 and Vox-forge 2015 with 100 speakers each. Switchboard NIST 2003[37] database actually consist of speech samples from 356 speakers recorded over a telephonic conversation for a duration of 2 minutes per speaker sampled at 8kHz, but we considered only 100 speaker by dividing each sample into 5 samples of almost 50 sec duration for analysis. Vox-forge 2015 database [38] is a collection of speech recorded from different regions of world at a rate of 8 kHz and only 100 speakers have been considered for analysis.

In this work, noisy speech signal was added to the speech samples of these two databases. The noise used here is of four types, white noise, babble noise, volvo noise, factory noise. Noise is added to two clean speech database at different signal-to-noise ratio with the help of MATLAB.

4.2 Experimental Analysis

The speech signal was sampled at a frequency of 8 kHz and standard MFCC features were calculated with a frame size of 20ms, a frame shift of 10ms considering standard 13 mel filter-banks. The identification accuracy was calculated by likelihood estimation model given by GMM. Now, new cepstral features as discussed in proposed methodology, were calculated by taking the n^{th} root and tested using GMM. Another feature set has been extracted by using the information content in terms of mean and variance. This feature set has been tested using SVM.

These methodologies were tested under the presence of noise signal of different types, namely babble, white, factory, Volvo noise. The noise was taken in the format of audio signal and was added to clean speech signal at different SNR and the order of root kept increasing each time, analysis was done till eighth root as mentioned earlier.

The performance of the system is tabulated in Table.4.1.A. for NIST 2003 Database and

Table.4.2.A. for Vox-forge 2015 database. It is evident from results of the identification as shown in Figure 4.1(a)-(d) (NIST 2003 database) and Figure 4.3.(a)-(d)(Vox-forge 2015 database), that the performance in terms of accuracy percentage of system is different for different values of noise in system.

For the system in which the signal is dominant to the value of noise i.e. for a high value of SNR, the highest accuracy is found to be on cubic root or 3^{rd} order of root. Also, 2^{nd} root gave best performance if the value of SNR is low i.e. noise is dominant in system. Precisely, for SNR of range 5-10 DB, square root based compressed features gave best accuracy results and for the system with SNR within 15-20 DB, cubic root compression based features gave best identification accuracy. However, the results were calculated on other values of root for analysis purpose only.

Also these results have been verified using Support vector machine and the results for NIST 2003 database is tabulated in Table4.1.B. and for Vox-forge 2015 database is as given by Table.4.2.B. Accuracy plot for NIST 2003 database with different methodology at different SNR tested on SVM is shown in Figure. 4.2.(a)-(d) with various noise types. Figure. 4.4(a)-(d) shows accuracy plot for Vox-forge 2015 database tested with SVM . In the figures, LOG represent the results calculated with standard logarithmic compression of feature set and 2, 3,… represent the results on feature set compressed by 2^{nd} root, 3^{rd} root and so on.

Table .4.1. A comparison of Accuracy (%) with different methodology on NIST- 2003 database

A. Identification Accuracy(%) tested with GMM

NOISE	SNR (dB)	METHODOLOGY							
		log	2nd root	3rd root	4th root	5th root	6th root	7th root	8th root
Clean		98	17	51	82	95	96	97	97
White	5	4	33	24	17	11	11	9	9
	10	19	64	60	50	43	39	39	36
	15	45	81	82	78	78	76	74	72
	20	76	83	90	88	89	86	86	85
Babble	5	9	9	7	6	7	9	9	7
	10	12	24	18	15	16	15	15	13
	15	23	50	52	42	39	36	39	40
	20	55	49	79	65	65	64	67	64
Volvo	5	8	38	7	13	12	13	11	7
	10	15	68	21	17	17	20	20	19
	15	29	46	79	36	36	37	40	39
	20	49	37	58	59	61	61	60	59
Factory	5	8	9	11	11	12	19	18	18
	10	15	19	23	26	27	30	23	19
	15	32	64	67	58	53	51	49	47
	20	63	65	88	79	77	76	78	76

B. Identification Accuracy(%) tested on SVM

NOISE	SNR (dB)	METHODOLOGY							
		Log	2nd root	3rd root	4th root	5th root	6th root	7th root	8th root
Clean		82	73	85	91	88	91	89	89
White	5	11	67	63	45	32	28	24	18
	10	19	74	72	69	59	50	46	42
	15	31	73	79	77	68	65	56	52
	20	39	74	82	82	74	73	69	66
Volvo	5	5	15	10	5	5	8	6	6
	10	6	33	16	14	10	11	9	8
	15	8	59	37	28	23	19	18	17
	20	15	72	66	56	46	38	33	30
Babble	5	6	25	18	5	5	5	5	3
	10	13	52	34	25	17	19	10	10
	15	18	66	66	53	44	37	33	32
	20	28	72	76	74	62	59	56	53
Factory	5	7	34	22	18	18	16	13	11
	10	8	50	32	34	26	23	21	18
	15	15	66	62	51	40	37	33	31
	20	30	72	75	75	66	63	56	56

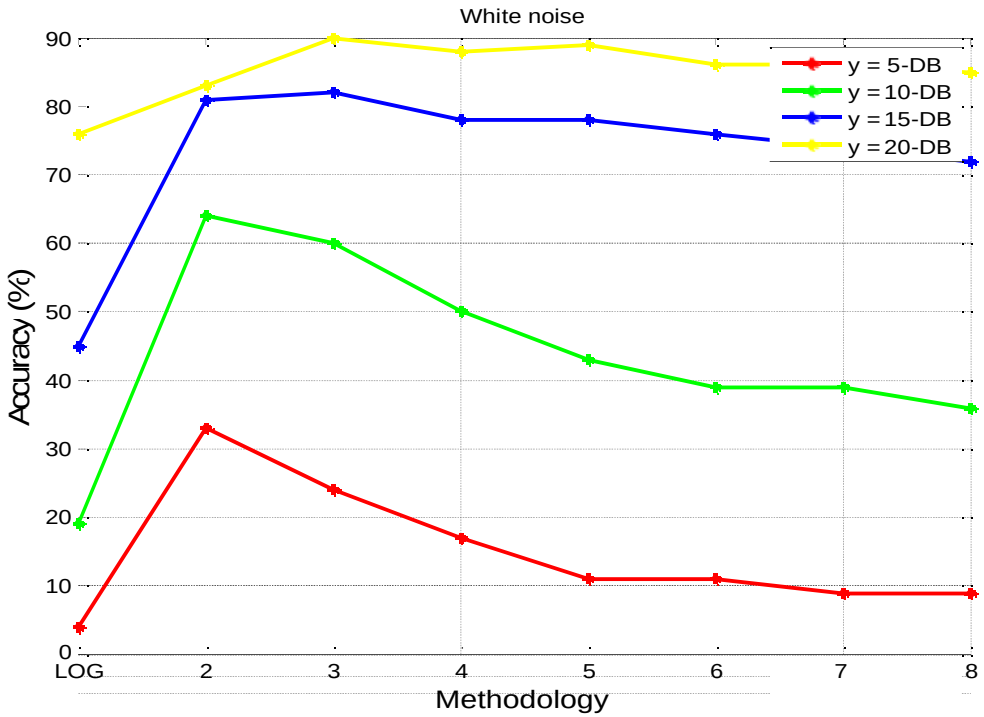

Figure. 4.1(a)

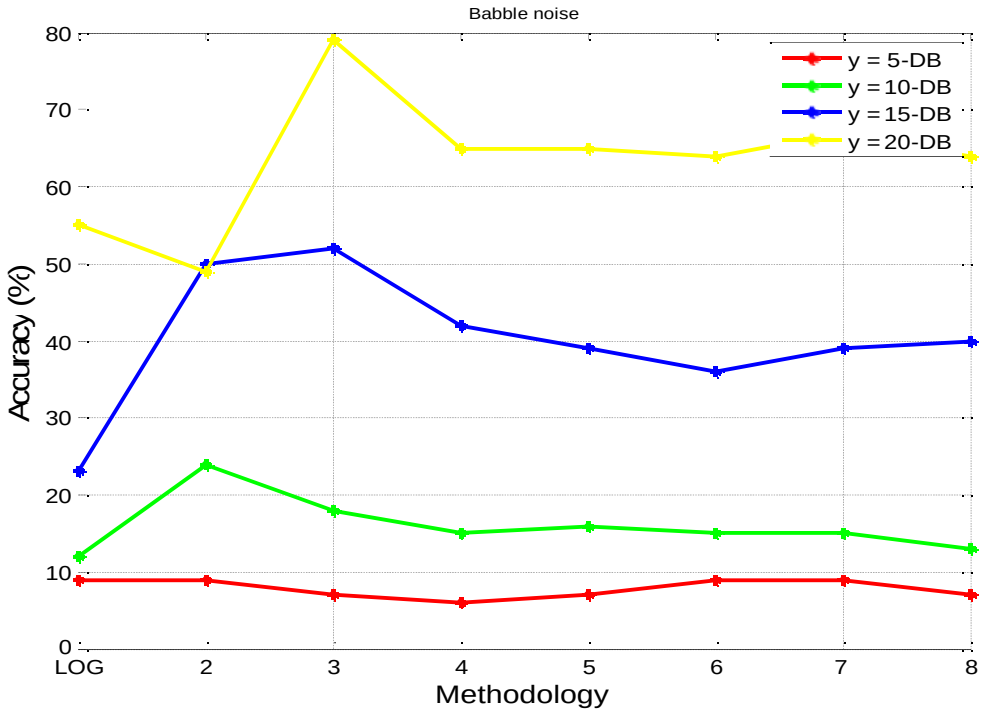

Figure.4.1(b)

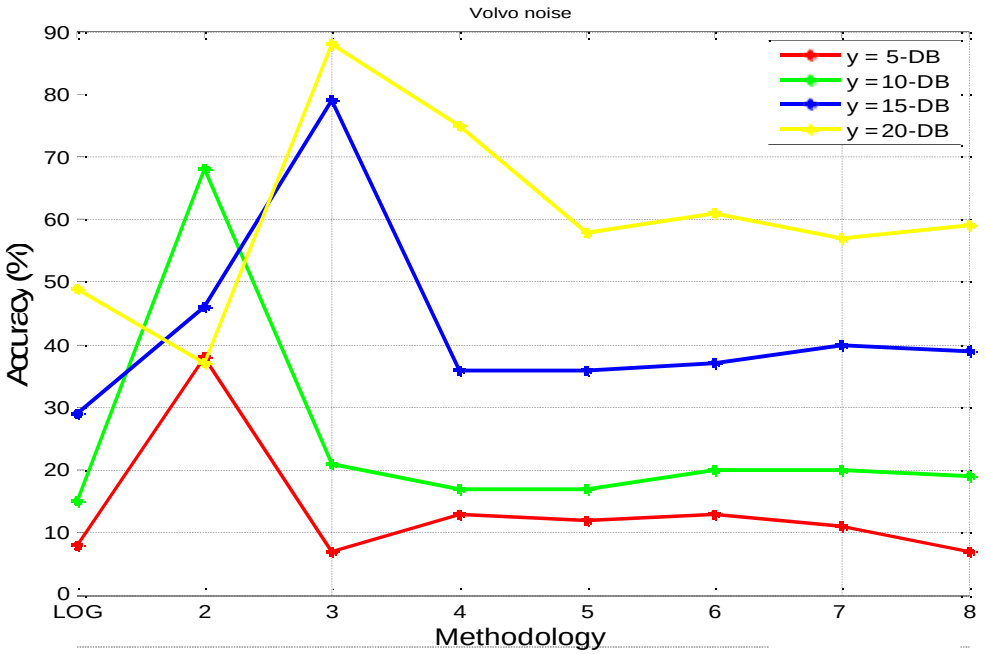

Figure.4.1(c)

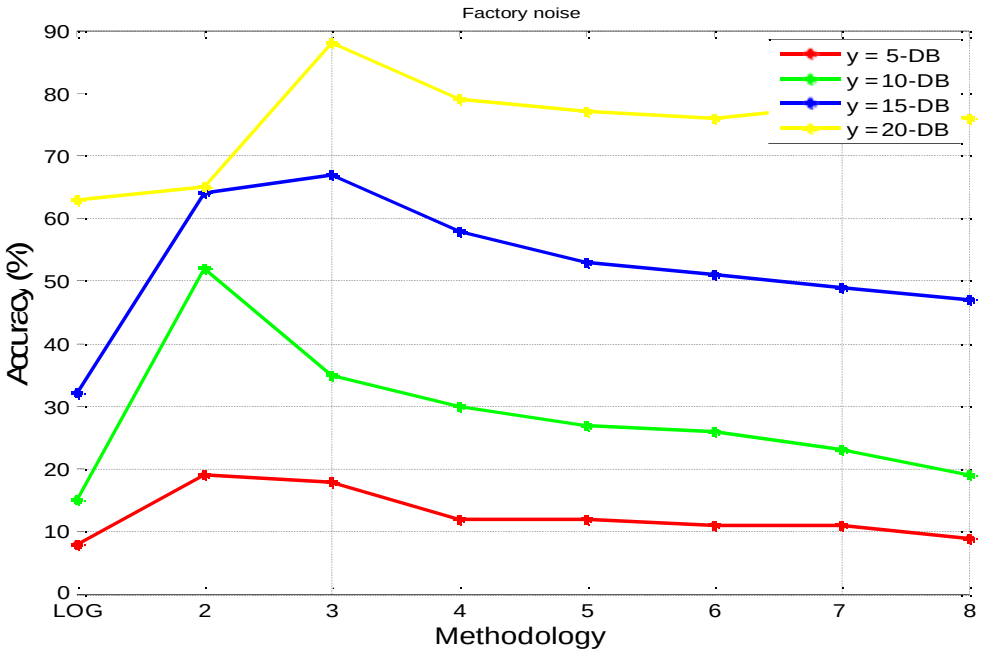

Figure.4.1(d)

Figure 4.1(a)-(d)Accuracy plot on NIST 2003 Database tested on GMM at different SNR

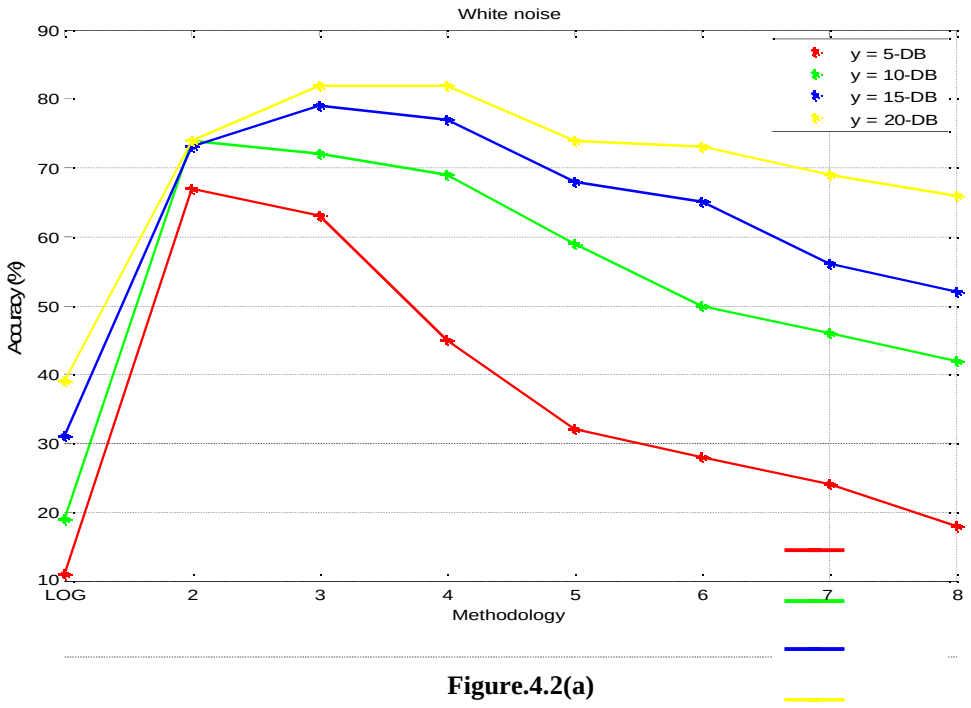

Figure.4.2(a)

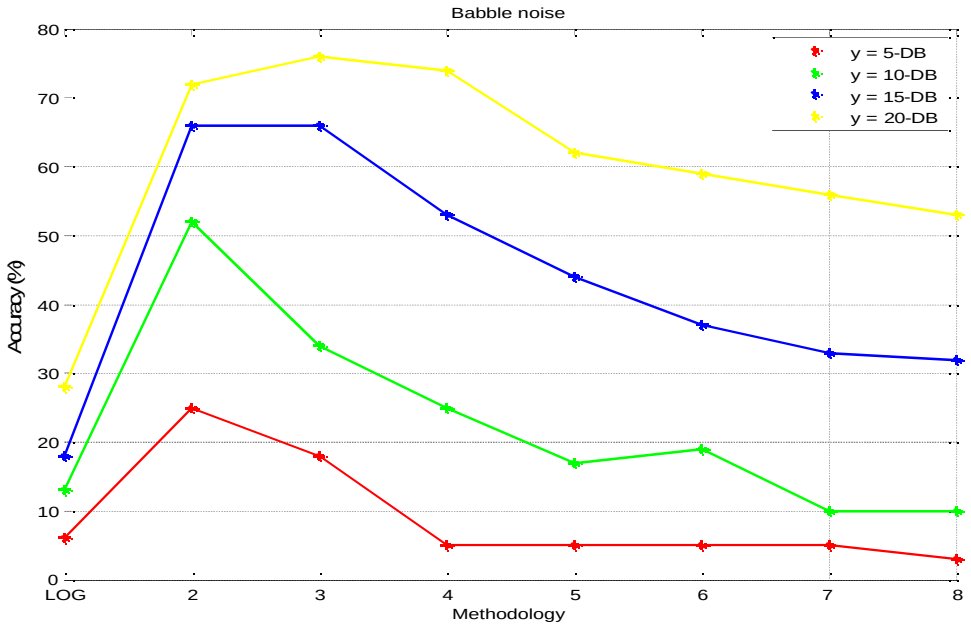

Figure.4.2(b)

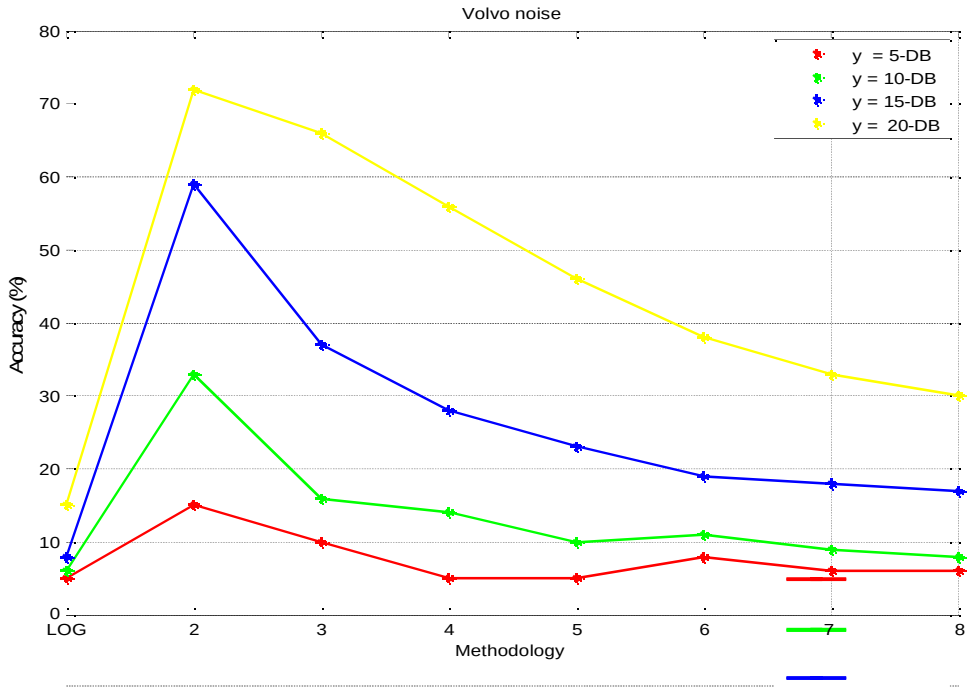

Figure.4.2(c)

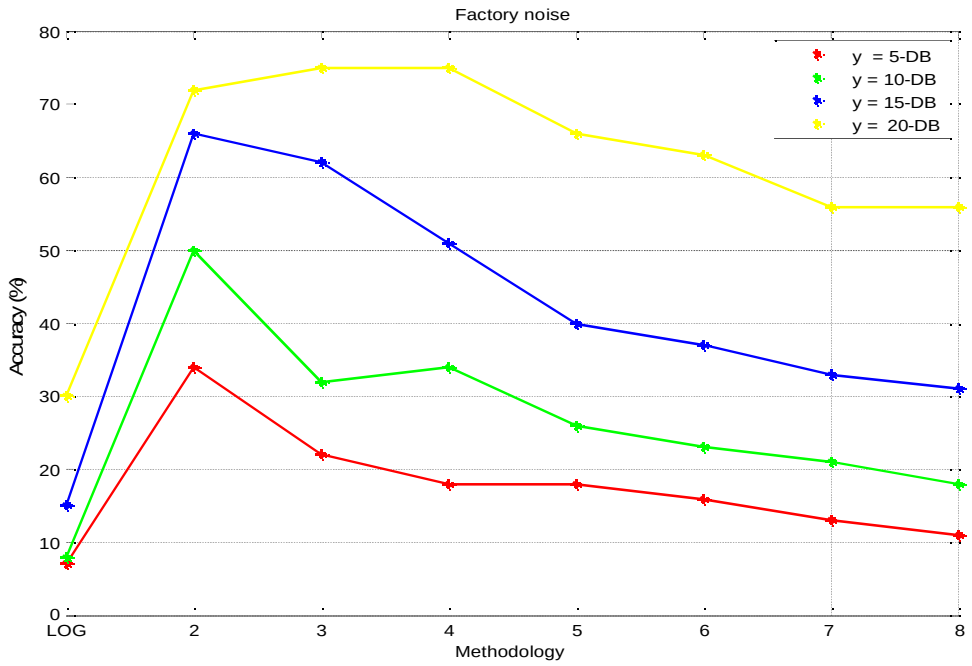

Figure.4.2(d)

Figure 4.2.(a)-(d)Accuracy plot on NIST 2003 Database tested on SVM at different SNR

Table .4.2. A comparison of Accuracy (%) with different methodology on Vox-forge 2015 database

A. Identification Accuracy(%) tested with GMM

NOISE	SNR (dB)	METHODOLOGY							
		log	2nd root	3rd root	4th root	5th root	6th root	7th root	8th root
Clean		100	83	96	99	99	100	100	100
Babble	5	35	45	43	30	41	42	38	32
	10	58	77	69	56	67	69	68	65
	15	77	83	89	70	85	86	84	85
	20	91	87	97	81	96	97	96	95
White	5	20	75	65	60	51	46	41	42
	10	53	91	92	91	88	85	83	79
	15	82	96	97	96	95	95	94	94
	20	93	94	96	96	96	96	96	97
Factory	5	11	38	31	31	23	22	23	23
	10	36	76	72	72	65	63	62	57
	15	69	87	91	89	88	85	81	80
	20	85	90	96	96	95	94	93	92
Volvo	5	48	22	29	30	37	39	37	37
	10	67	48	50	56	53	54	52	55
	15	78	73	70	70	72	69	68	70
	20	90	77	87	81	82	83	84	85

B. Identification Accuracy(%) tested with SVM

NOISE	SNR (dB)	METHODOLOGY							
		log	2nd root	3rd root	4th root	5th root	6th root	7th root	8th root
Clean		87	74	80	81	80	83	85	83
Babble	5	12	23	16	15	14	13	14	13
	10	24	50	39	36	30	30	29	27
	15	44	66	69	67	58	55	54	54
	20	66	72	75	79	77	70	69	68
White	5	26	68	58	48	44	34	31	29
	10	41	74	74	68	63	63	58	54
	15	56	76	77	74	73	73	70	69
	20	62	76	77	77	79	78	77	75
Factory	5	22	18	17	15	14	15	15	15
	10	32	35	31	31	30	36	35	35
	15	47	61	53	49	50	52	53	53
	20	62	76	66	68	69	66	69	66
Volvo	5	19	34	27	20	20	19	16	16
	10	35	57	53	47	46	45	38	38
	15	48	70	71	65	65	64	64	63
	20	66	72	77	74	78	73	71	70

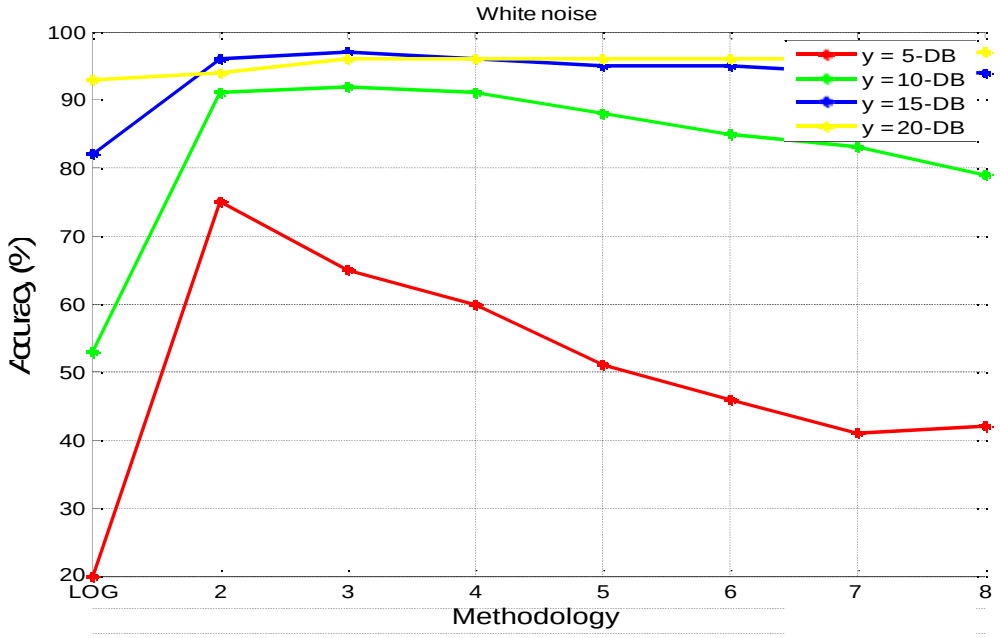

Figure.4.3(a)

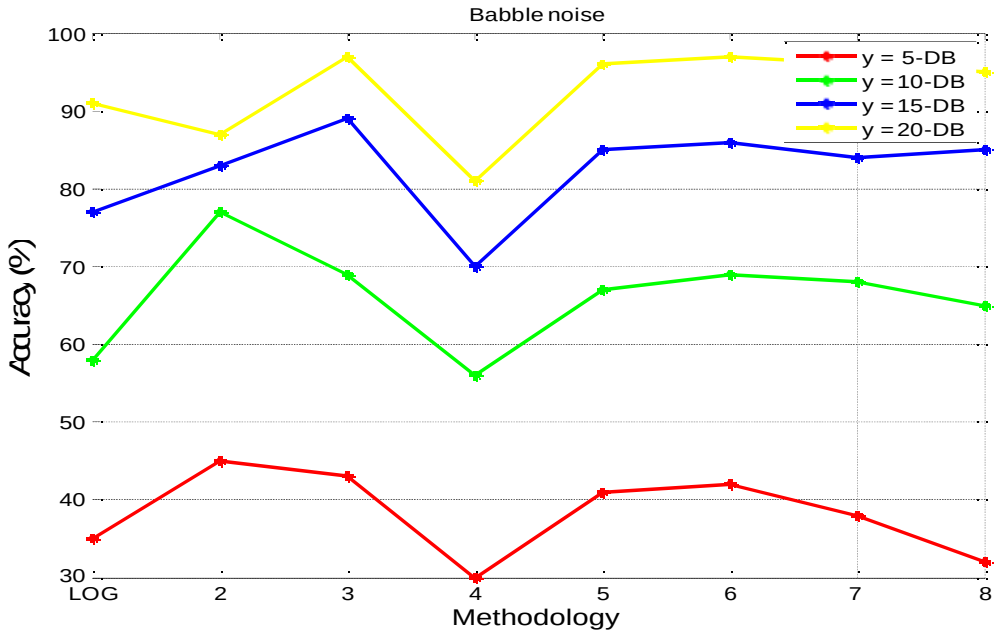

Figure.4.3(b)

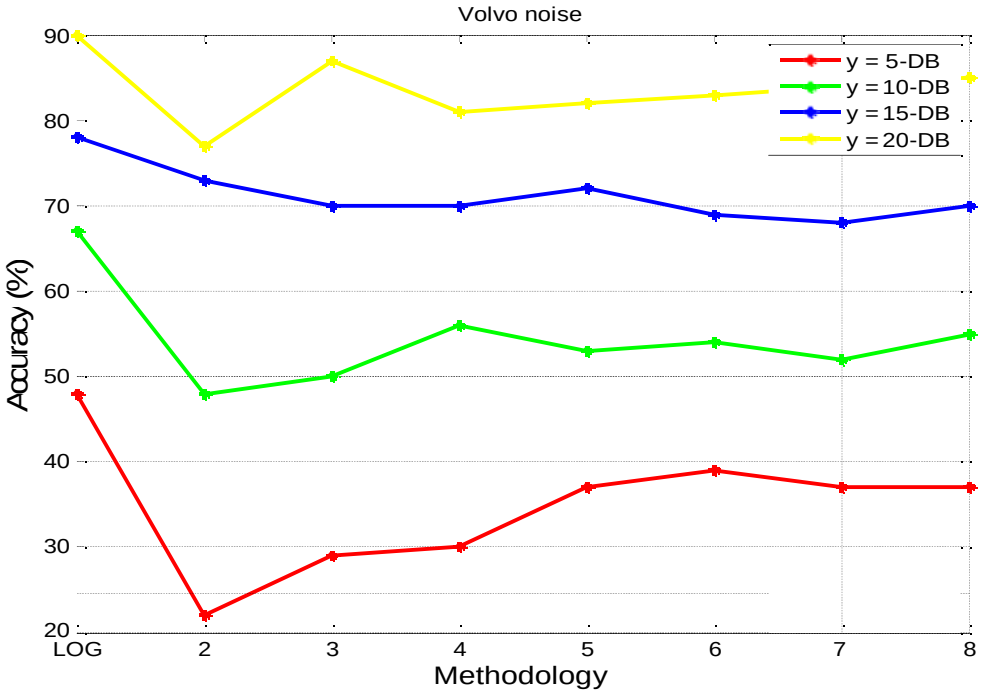

Figure.4.3(c)

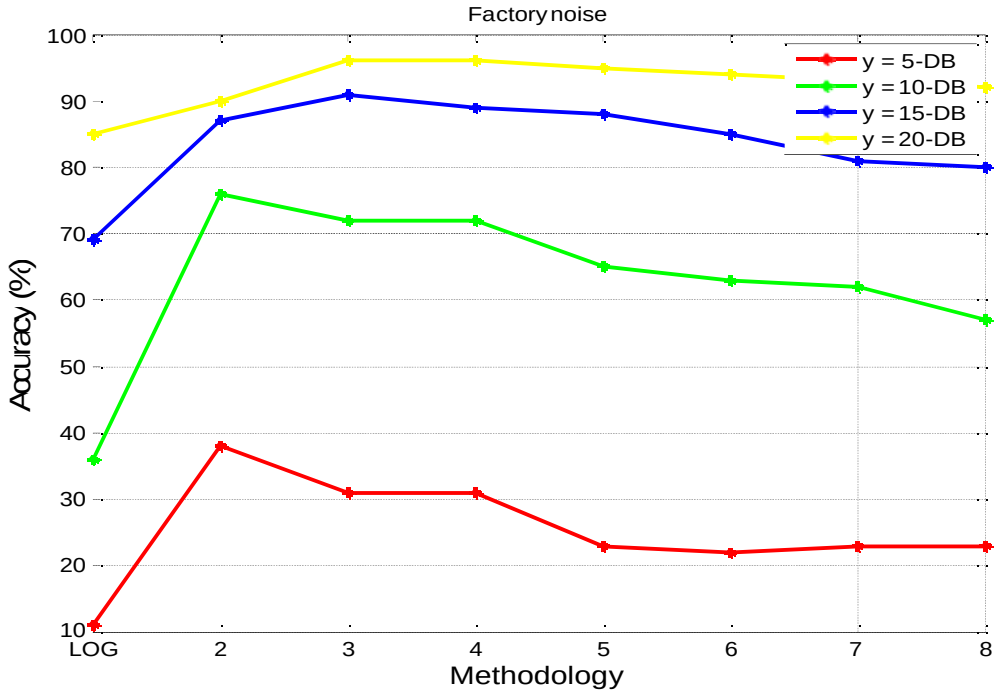

Figure 4.3(a)-(d). Accuracy plot on Vox-forge 2015 Database tested on GMM at different SNR

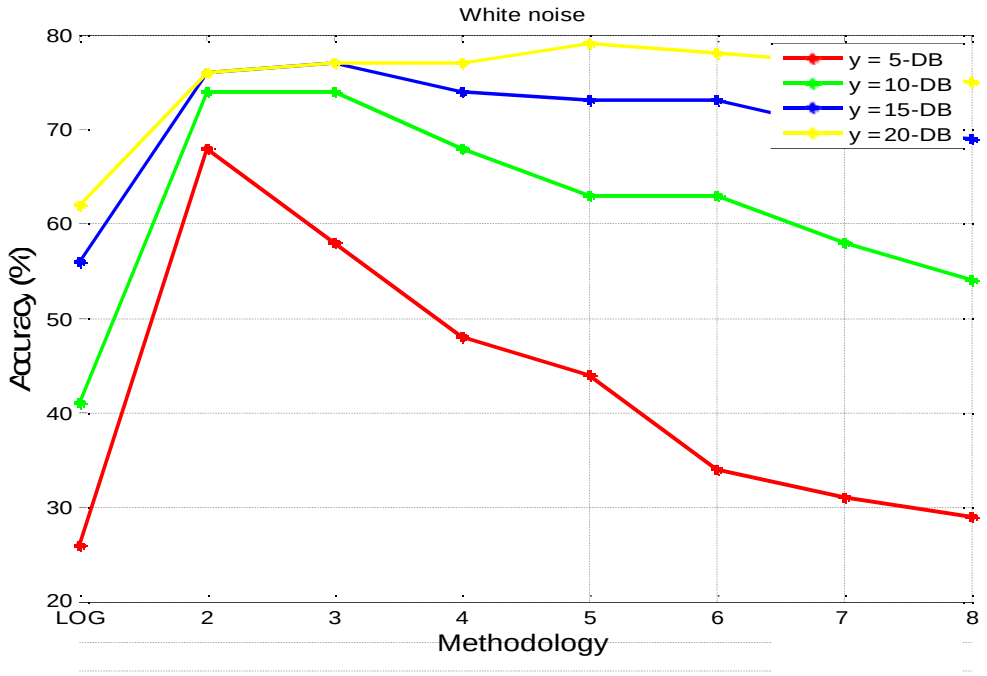

Figure.4.4(a)

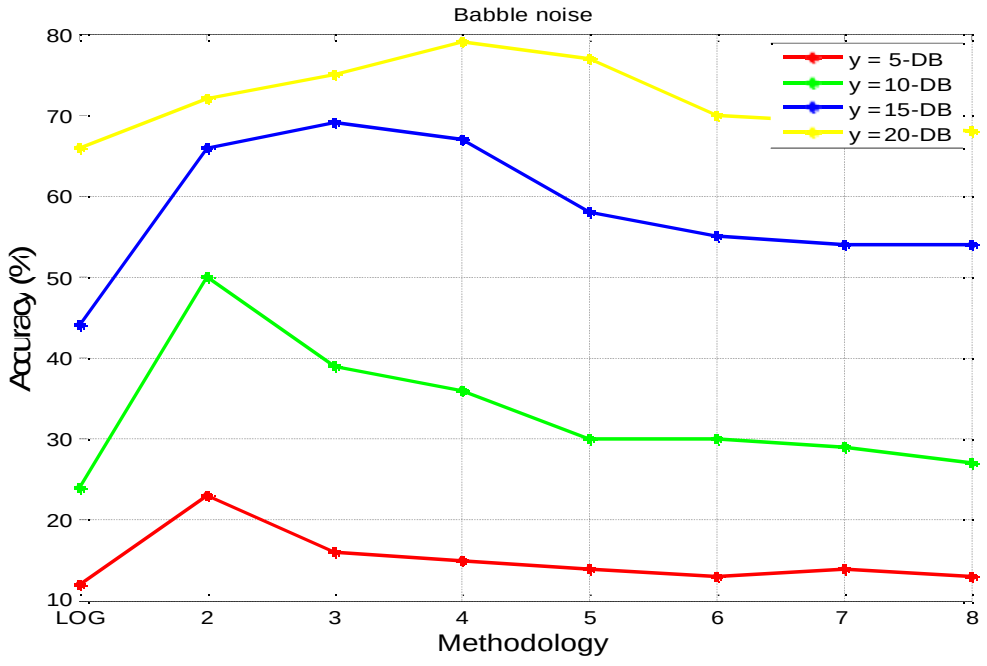

Figure.4.4(b)

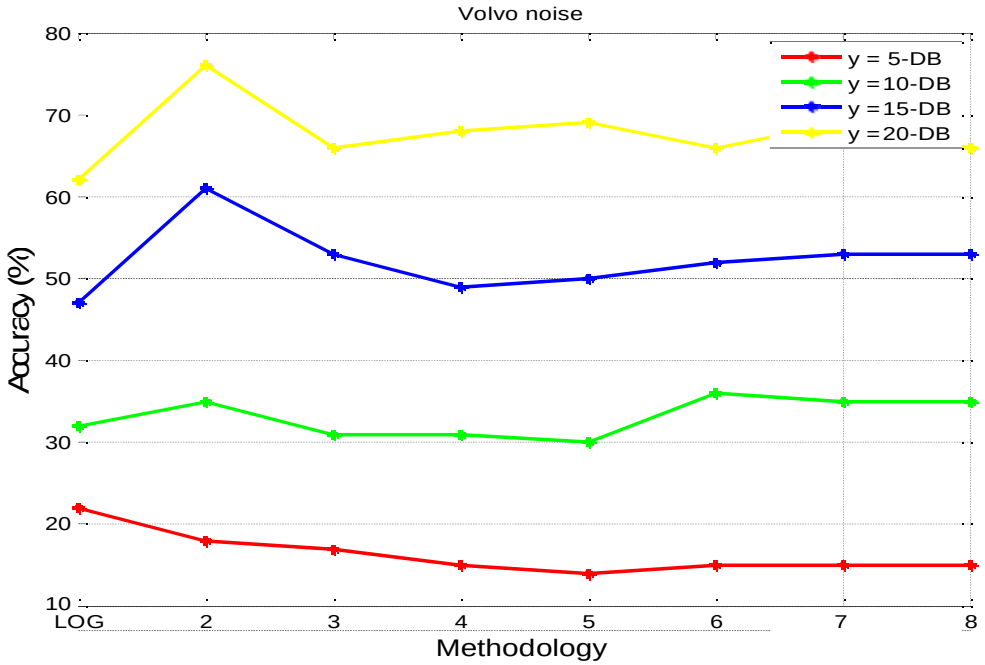

Figure.4.4(c)

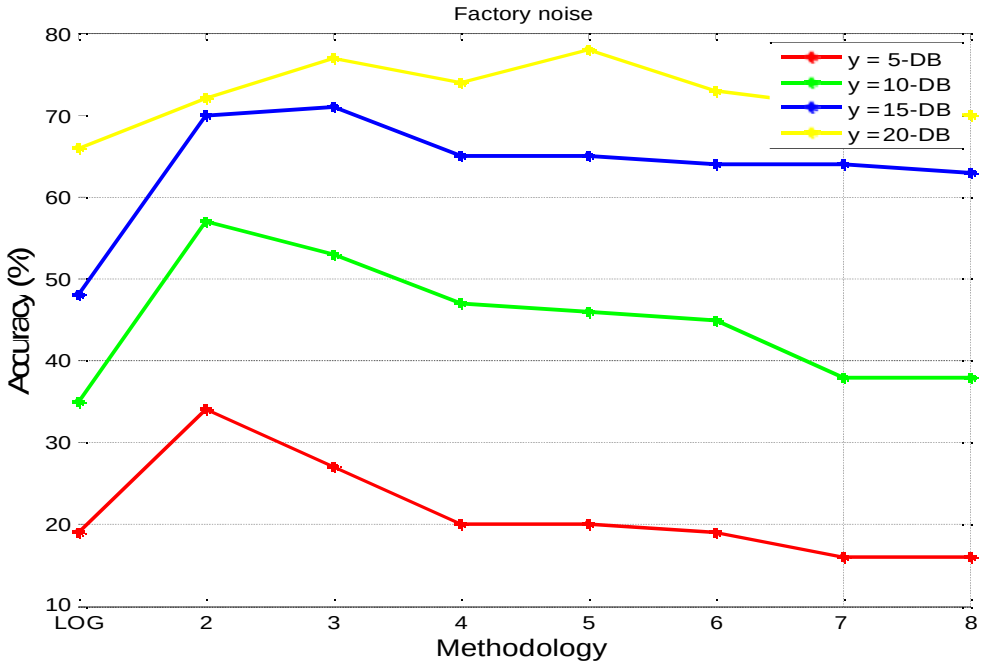

Figure 4.4(a)-(d). Accuracy plot on Vox-forge 2015 Database tested on SVM at different SNR

Chapter -5

CONCLUSIONS AND SCOPE FOR FUTURE WORK

5.1 Conclusion

In the present work, the various features have been tested on gaussian mixture model and support vector machine. The methodology introduced for feature extraction is based on root-compression. Another methodology where information is extracted from standard MFCC is used. It is shown that proposed methodology increases the identification accuracy many folds when the signal is degraded by noise. Results in terms of accuracy (%) are far better than those obtained by standard methodology that uses MFCC features on GMM.

Also, these models are validated on NIST-2003 and Vox-forge database at different values of SNR for different types of noise. Also, the results have been verified with two classification methods, GMM and SVM.

The following conclusions can be drawn from this work:

1) The proposed system has performed far better than the standard procedure available for speaker identification using GMM on MFCC.
2) For a signal dominant system with a SNR in range of 15-20 DB cubic root compression based system have performed better than other values of root.
3) For a noise dominant system performance in terms of accuracy (%), with a SNR in range of 5-10 DB, is better for square root based compression.

5.2 Future Scope of work

The challenges remaining in this work are directed below:

- For future work, some optimisation algorithm like genetic algorithm can be implemented to find the values of parameter in SVM.
- Also, the value of root can be optimised using some optimisation technique.
- In addition to it, wavelet transformation can be employed at the feature extraction stage for increasing the robustness of speaker identification.

BIBLIOGRAPHY

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, 2004.

[2] D. a. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Ieee Transactions On Speech And Audio Processing*, vol. 3, no. 1. pp. 72–83, 1995.

[3] I. Pollack, J. M. Pickett, and W. H. Sumby, "On the Identification of Speakers by Voice," *J. Acoust. Soc. Am.*, vol. 26, no. 3, pp. 403–406, 1954.

[4] J. N. Shearme and J. N. Holmes. An experiment concerning the recognition of voices. Language and Speech, vol. 2, no. 3, pp.123–131, 1959.

[5] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," *J. Acoust. Soc. Am.*, vol. 215, no. 3, pp. 214–215, 1963.

[6] J. Bradbury, "Linear predictive coding," *Florida Inst. Technol.*, ., vol. 2, no. 3, pp.104-198,2000.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech.," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–52, 1990.

[8] I. Introduction, "Comparison of Parametric Representations for," no. 4, 1980.

[9] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," *2011 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 4784–4787, 2011.

[10] N. Hermansky, H., Morgan, "RASTA processing of speech.," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 587–589, 1994.

[11] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM composition in noisy environments," *Computer Speech & Language* 10.2 (1996), pp. 107–116, 1996.

[12] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," *ODYSSEY-2001 - Speak. Recognit. Work.*, pp. 213–218, 2001.

[13] K. S. Rao and S. Sarkar, "Robust Speaker Recognition in Noisy Environments," Springer, vol. 15, no. 5, pp. 13–28, 2014.

[14] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," vol. 19, no. 4, pp. 788–798, 2011.

[15] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[16] Y. Shao, "Sequential organization in computational auditory scene analysis," *Comput. Eng.*, 2007.

[17] M. A. Abd El-Fattah, M. I. Dessouky, A. M. Abbas, S. M. Diab, E. S. M. El-Rabaie,

W. Al-Nuaimy, S. A. Alshebeili, and F. E. Abd El-Samie, “Speech enhancement with an adaptive Wiener filter,” *Int. J. Speech Technol.*, vol. 17, no. 1, pp. 53–64, 2014.

[18] Mallat, S. and Hwang, W. L., "Singularity detection and processing with wavelets," *IEEE Transactions on Information Theory,* vol. 38, pp. 617-643, 1992 .

[19] S. M. Govindan, P. Duraisamy, and X. Yuan, “Adaptive wavelet shrinkage for noise robust speaker recognition,” *Digit. Signal Process.*, vol. 33, pp. 180–190, 2014.

[20] X. Zhao, S. Member, Y. Shao, and D. Wang, “CASA-Based Robust Speaker Identification,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 5, pp. 1608–1616, 2012.

[21] X. Zhao and D. Wang, “ANALYZING NOISE ROBUSTNESS OF MFCC AND GFCC FEATURES IN SPEAKER IDENTIFICATION,” *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 7204–7208, 2013.

[22] J. M. Lovekin, R. E. Yantorno, K. R. Krishnamachari, D. S. Benincasa, and S. J. Wenndt, “Developing usable speech criteria for speaker identification\ntechnology,” *2001 IEEE Int. Conf. Acoust. Speech, Signal Process. Proc. (Cat. No.01CH37221)*, vol. 1, pp. 421–424 vol.1, 2001.

[23] D. Reynolds, “Robust Speaker Identification *,” pp. 185–188, 1992.

[24] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[25] I. Rosenblatt, “Analysis of a Four-Layer Series-Coupled Perceptron. II,” 1962.

[26] G. Hinton, G. Hinton, T. Sejnowski, and T. Sejnowski, *Learning and relearning in Boltzmann machines*, vol. 1. 1986.

[27] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A Training Algorithm for Optimal Margin Classifiers,” *Proc. Fifth Annu. ACM Work. Comput. Learn. Theory*, pp. 144–152, 1992.

[28] A. Ganapathiraju, J. Hamaker, and J. Picone, “Hybrid SVM / HMM Architectures for Speech Recognition,” *Icslp*, vol. 4, no. 1, pp. 504–507, 2000.

[29] J. C. Platt, “Probabilities for SV Machines,” *Adv. large margin Classif.*, pp. 61–74, 2000.

[30] M. Yankayi, “FEATURE EXTRACTION MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC).”

[31] P. Motl, “Feature Extraction in Speech Coding and Recognition,” *Fit.Vutbr.Cz*, pp. 1–50, 2003.

[32] J. Kiefer, “Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to The Annals of Mathematical Statistics. ® www.jstor.org,” *Statistics (Ber).*, vol. 37, no. 3, pp. 688–697, 1966.

[33] D. a. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[34] A. P. A. Dempster, N. M. N. Laird, and D. D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[35] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification," vol. 1, no. 1, pp. 1–16, 2010.

[36] C. Chang and C. Lin, "LIBSVM : A Library for Support Vector Machines," pp. 1–39, 2013.

[37] [Online]. (2003) *The NIST year 2003 speaker recognition evaluation plan*. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2003/2003-spkrec-evalplan-v2.2.pdf

[38] [Online]. (2015). *VoxForge speech corpus*. Available: http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/

LIST OF PUBLICATIONS

[1] Vishu Sharma and Saurabh Bhardwaj, Root-based compression based Speaker Identification, *Communicated* to Computational Intelligence, Wiley-Blackwell, United States, (ISSN: 0824-7935, listed in SCI Expanded).

Vishu_Thesis

ORIGINALITY REPORT

7%	4%	5%	2%
SIMILARIT Y INDEX	INT ERNET SOURCES	PUBLICAT IONS	ST UDENT PAPERS

PRIMARY SOURCES

1	www.fysel.ntnu.no Int ernet Source	2%
2	Submitted to University of Queensland Student Paper	1%
3	formulas.ultrafractal.com Int ernet Source	1%
4	ijcscn.com Int ernet Source	<1%
5	Submitted to Visvesvaraya Technological University Student Paper	<1%
6	Advances in Intelligent Systems and Computing, 2013. Publication	<1%
7	www.iam.unibe.ch Int ernet Source	<1%
8	Fundamentals of Music Processing, 2015. Publication	<1%
9	www.ijecscse.org Int ernet Source	<1%