SPEAKER IDENTIFICATION OF DISGUISED VOICES USING MFCC STATISTICAL MOMENT AND SVM CLASSIFIER

A Dissertation Submitted in Partial Fulfillment of the Requirement for the Award of the Degree of

MASTER OF ENGINEERING

In Wireless Communication

Submitted By

Harleen Kaur

801563005

Under Supervision of

Dr. Ashutosh Singh

Assistant Professor

ELECTRONICS AND COMMUNICATION ENGINEERING DEPARTMENT

THAPAR UNIVERSITY, PATIALA, PUNJAB

JULY, 2017

DECLARATION

I, **Harleen Kaur**, hereby declare that the dissertation entitled **"Speaker Identification of disguised voices using MFCC statistical moments and SVM classifier"** is an authentic record of my study carried out towards the partial fulfilment as requirement for the award of degree of Master of Engineering in Electronics and Communication at Thapar University, Patiala, under the supervision of **Dr. Ashutosh Singh**, **Assistant Professor**, Electronics and Communication Engineering Department. The matter presented in this dissertation has not been submitted to any other University/Institute for the award of any other degree.

Date 08-Sep-2017

Harleen Kaur

801563005

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

Date 08-Sep-2017

Singh

Dr. Ashutosh Singh

Assistant Professor

ACKNOWLEDGEMENT

With deep sense of gratitude I express my sincere thanks to my esteemed and worthy supervisor, **Dr. Ashutosh Singh, Assistant Professor**, Department of Electronics and Communication Engineering, Thapar University, Patiala for his valuable guidance in carrying out work under his effective supervision, encouragement, enlightenment and cooperation. Most of the novel ideas and solutions found in this dissertation are the result of our numerous stimulating discussions.

I shall be failing in my duties if I do not express my deep sense of gratitude towards **Dr. Alpana Agarwal,** Professor and Head of the Department of Electronics and Communication Engineering, Thapar University, Patiala who has been a constant source of inspiration for me throughout this work, and for the providing us with adequate infrastructure in carrying the work.

I am also thankful to **Dr. Hem Dutt Joshi**, Assistant Professor and P.G. Coordinator, Electronics and Communication Engineering Department, for the motivation and inspiration that triggered me for this work.

I am greatly indebted to all my friends who constantly encouraged me and also would like to thank the entire faculty and staff members of Electronics and Communication Engineering Department for their unyielding encouragement.

At last but not the least my gratitude towards my parents, who always supported me in doing the things my way and whose everlasting desires, selfless sacrifice, encouragement, affectionate blessings and help made it possible for me to complete my degree.

Place: TU, Patiala **Harleen Kaur**

Date: Roll no. 801563005

ABSTRACT

Speaker Recognition is the most powerful scheme in which security is the major concern but misclassification rate often increases when speakers attempt to disguise their voices. This project mainly focuses on electronic voice disguise method that involves modification, deviation or twisting of the original voice of a speaker electronically via voice changing software's. With sophisticated algorithms, electronic methods achieve better results than non-electronic methods as they provide a vast range of choices for voice disguise. Voice disguise has been expanding in several illegal applications like annoying phone calls, ransom calls, and bomb threats etc. It has adverse effects on determining the veracity of audio evidence. It is utmost important in audio forensics to spot whether the so-called voice of a speaker is original or disguised.The proposed system presents an efficient way to identify a speaker using MATLAB tool. The system recognizes a speaker and makes a decision on the basis of information present in his/her voice that may be an original voice or disguised voice.The first and foremost task is to find the disguised voices from the original voice and then the next step to find its identity speaker. Several related studies have been accounted on Automatic Speaker Recognition (ASR) systems but not many studies on recognizing speaker from their disguised speech. Here acoustic information of each speaker is passed through various stages i.e., preprocessing, feature extraction, pattern matching, decision making. To carry out this research, MFCC based statistical moments involving mean and correlation coefficients are used as acoustic features in feature extraction stage and SVM classifier are used in feature matching stage. These features and classification technique provide useful and precise results for the identification of hidden speaker.

TABLE OF CONTENTS

LISTS OF FIGURES

LIST OF ABBREVIATION

ASR:	Automatic Speaker Recognition
SVS:	Speaker Verification System
SIS:	Speaker Identification System
NIST:	National Institute of Standards and Technology
MFCC:	Mel Frequency Cepstral Coefficients
SVM:	Support Vector Machine
LPC:	Linear Predictive Coding
LPCC:	Linear Predictive Cepstral Coefficients
DCT:	Discrete Cosine Transform
GMM:	Gaussian Mixture Model
VQ:	Vector Quantization
ANN:	Artificial Neural Network
HMM:	Hidden Markov Model
DTW:	Dynamic Time Warping
KNN:	K-Nearest Neighbours
DFT:	Discrete Fourier Transform
NB:	Naive Bayes
LDA:	Linear Discriminant Analysis

CHAPTER 1

INTRODUCTION

NOWADAYS, it is quite impossible to think of life without gadgets like desktops, laptops, tablets, phones, music etc. These gadgets not only enhance the capabilities of human but also simplify the life in so many ways. Societies' dependence on widely accessible services like telephones, internet, mobile phones and tape recorders at times results in the mishandling, such as kidnapping, blackmail threats, unknown calls, bomb threats, harassment calls, etc. In order to avoid these security breaches, one should be familiar with the concepts by which we can look after such happenings. There is need to develop a robust speaker recognition system that produces accurate results even if any disguised voices are encountered. One such method is speaker recognition process which has been worked upon in this thesis. In order to analyze speaker recognition model, one needs to understand the notion behind speech production process.

1.1 SPEECH PRODUCTION SYSTEM

Speech is defined as an articulated sound generated when the air is expelled out from the lungs and causes acoustical excitation of the vocal tract. Every human makes a variety of sounds whose range is related with the physical configuration and shape of the vocal tract system. The speech production process starts from human's mind when a message is formulated which is further propagated to the listener via speech [1]. For voiced sound production, the vocal chords of a speaker vibrate. For example /a/, /e/, /i/ etc. Unvoiced sounds are formed if the vocal cords are held open and air rushes via lungs and then through the vocal tract it finally comes out at the lips. Examples include the unvoiced fricatives sounds /s/, /f/ and /sh/. Figure 1.1 shows the different organs that generate the different types of speech sounds and its associated parameters. A vocal tract is non-uniform tube that begins at the vocal chords and ends at the mouth. It provides an acoustic transmission for sounds created within the vocal tract. It expels various sounds via oral cavity or nasal cavity through the vocal tract that consists of glottis, tongue, velum, teeth, and lips. In speech processing, basic speech sounds are classified in two distinct categories which are further dependent on the role of the vocal cords: voiced sound and unvoiced sound.

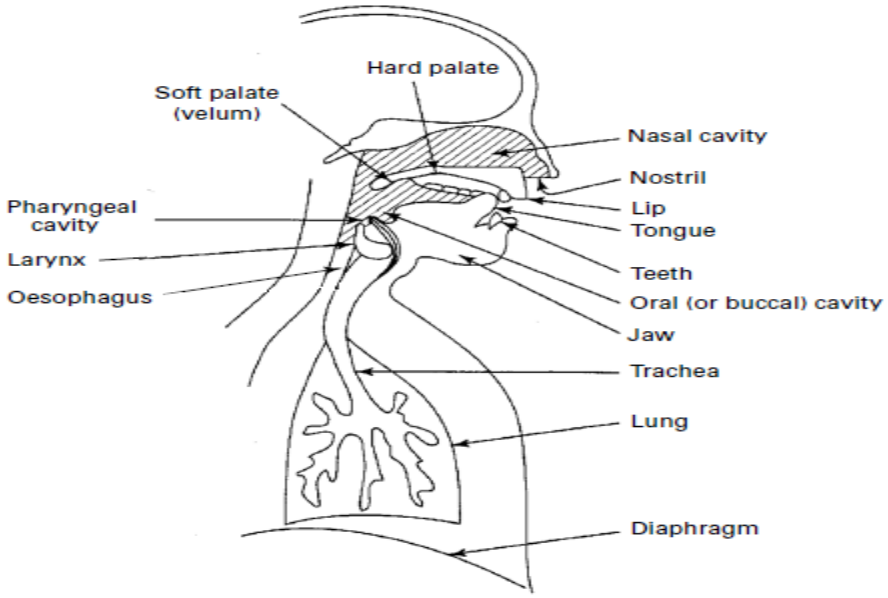


Figure 1.1 Schematic representation of human speech production organ [1].

The nasal tract begins at one end by the velum and ends at the nostrils. When the velum is lowered, nasal tract acoustically coupled to the vocal tract. The soft palate controls the isolation between nasal cavity and the pharynx. There is a pathway at the lower part of the pharynx that stops food from reaching to the larynx and disconnects the oesophagus acoustically from the human vocal tract. The pharynx connects to the oral cavity with larynx. The main function of pharynx is to transfer the food from mouth to pharynx and then via oesophagus to the stomach. The other essential part of the vocal tract is oral cavity which consists of palate, tongue, lips, cheeks, and teeth. The size, shape and acoustic depend on the movement of these oral cavity segments [2].

1.2 SPEECH TAXANOMY

Speech is the efficient mode of communication which brings the world together in so many ways. The speech production process starts from speaker's mind when a person formulates a message to express an idea and feeling which is propagated to the listener via speech [3]. The major areas of research in the field of speech communication include speech recognition, speech coding, speaker recognition, speech enhancement and speech synthesis. One of the most intriguing applications for forensic department from quite a long time is recognition field of speech as illustrated in Figure 1.2.

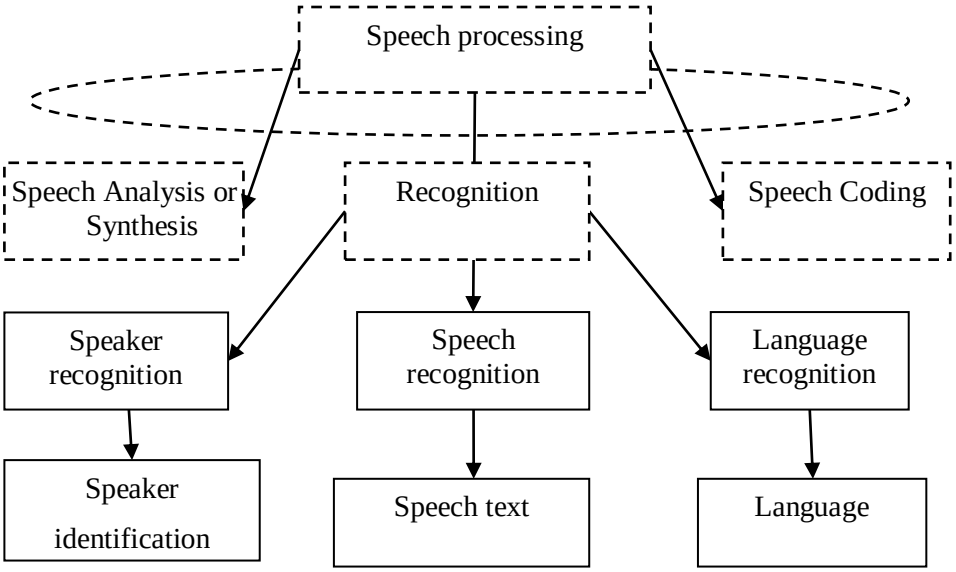


Figure 1.2 Block diagram representation of recognition field [3].

For in-depth study of speech processing field, one should understand three distinct scheme of recognition that is given below.

- **Speech Recognition**

This technique aims at identifying the spoken words automatically. Its task is to automatically convert the human speech into a format or text which can be easily readable by the system, hence known as automatic speech recognition (ASR) system. There are many applications of speech recognition such as speech based biometrics, voice transcription, calendar entry and update, stock quotes, telephone directory assistance etc.

- **Language Recognition**

This system has capability to understand speech in various native languages spoken by the speaker. It enables common man to exploit the benefits of information technology by providing services like automatic voice translation into foreign languages, voice dictation systems. This system keeps an elderly, physically challenged specially blinds closer to the technology.

- **Speaker Recognition**

This system extracts the speaker-specific information from his/her speech in order to identify or verify the speaker [3]. Speech can be considered as an excellent tool to prevent unauthorised access. Nowadays offenders believe that no one would identify them and they will remain unidentified, however it is no longer true. Thus, making the speaker recognition system need of the hour.

1.3 GENERIC SPEAKER RECOGNITION SYSTEM

Speaker Recognition System (SRS) extracts the speaker specific information based on their voices in order to identify or verify an unknown speaker. To identify an offender, a lot of attention is now being paid on *speaker recognition* field. Here human speech is used as a verification tool to verify an unknown speaker. The speaker-specific information is generated due to the complex changes happening at various levels of the speech production. Figure 1.3 shows the generic block diagram of speaker recognition system. It consists of the following sections:

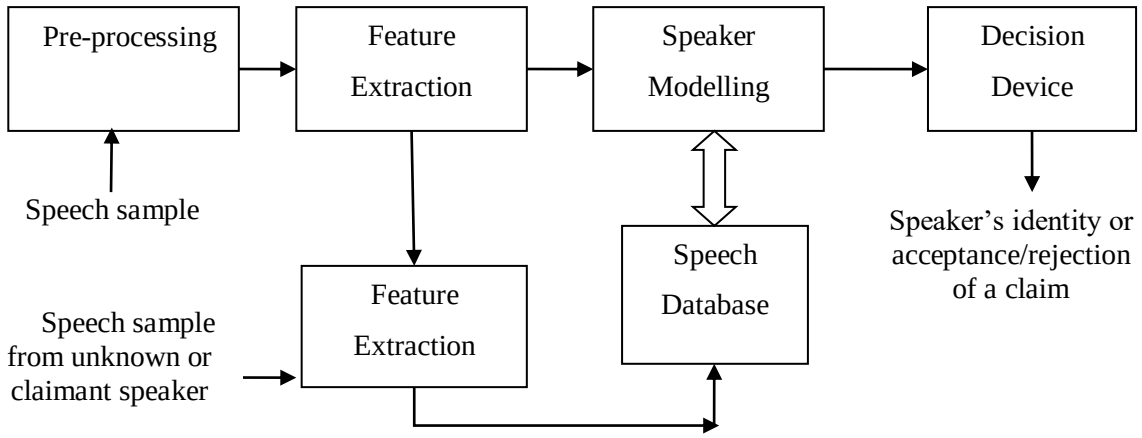


Figure 1.3 Block diagram of generic speaker recognition system [4]

- **Front-end processing**

This stage usually involves two major sections: pre-processing and feature extraction. Initially, the speech is divided into a number of frames with the help of a 20 msec. window whose frame rate is 10 msec. A pre-processing stage removes silence as well as noise from each speech frames. Then, spectral features are calculated from these frames via spectral analysis. Hence this stage removes unwanted noise and enhances the speech signal; furthermore it also reduces the channel distortion effects. Channel compensation module eliminates the channel effects and finally the discrete time acoustic signal is mapped to a sequence of features. These features represent the acoustics properties of speaker that is need to classify an unknown speaker from a known set of speakers. There are many types of feature extraction techniques such as linear prediction coefficients, linear prediction cepstral coefficients, Mel frequency cepstral coefficients etc. After extracting these features, it is further used in speaker modelling that pattern matching and decision.

- **Speaker modeling**

The above mentioned features are used in speaker modelling. There exist many modelling methods to analyze the speaker recognition systems like template matching, nearest neighbour, neural networks, hidden Markov model etc. but its selection is mainly dependent on the nature of speech input, training and testing activities, and storage and computation efficiencies.

- **Speaker Database**

All the speaker models are stored here.

- **Decision Device**

This device helps in decision making by choosing the most appropriate model by comparing the estimated features vectors to the speaker models in which all the speech samples and their feature vectors of corresponding speaker are present. Finally, the decision module analyzes the similarity score(s) for decision making.

1.4 TYPES OF SPEAKER RECOGNITION

The two major application areas of speaker recognition systems that are used to discriminate people from their voices are: speaker identification and speaker verification. In Speaker Identification System (SIS), a given speech is compared with a set of known voices. The task is to find the identity of an unknown speaker that best matches with a set of known labelled speaker models. There are two types of speaker identification system: closed and open set. In closed set SIS, an unknown speaker is assumed to be from known set of speakers whereas in *open-set SIS,* test speaker is not from the predefined model of known speakers. If a speaker is not matched to any other known speakers then no-match output is displayed.

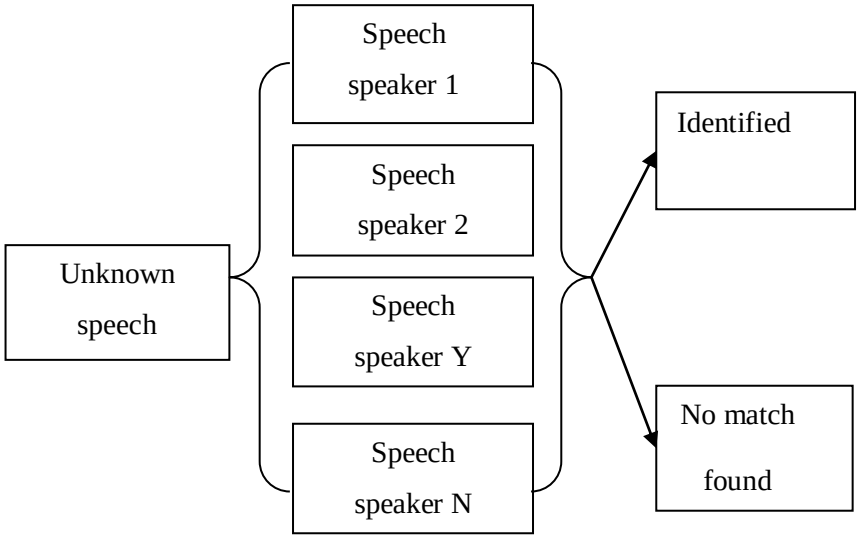


Figure 1.4 Schematic representation of speaker identification system [5].

This mode can be used for forensic purposes, in which an offender's identity is revealed among several known suspects such as voice evidence can be used to identify the perpetrators. Although there are numerous applications of speaker identification but verification field has been the backbone of almost all the speaker recognition applications.

Speaker Verification System (SVS) determines whether a speaker is who he or she claims to be. This process is also known as a true-false binary decision problem which provides control access to services like voice authentication, speaker authentication or speaker detection, database access verification, and security control for some confidential data security[3]. The claimed speaker is accepted or rejected by comparing the appropriate matching score with a predefined threshold. In such systems, there is no need of the determining the identity of the speaker, only verification is necessary. A biometric system is the most important application of speaker verification systems (SVS). Speaker verification system is another type of open-set speaker identification problem which has one target speaker. The performance of speaker identification system decreases as the number of speakers increases[6]. On the other hand, the performance of speaker verification system is independent of the speakers' size as it requires only one trial of comparison.

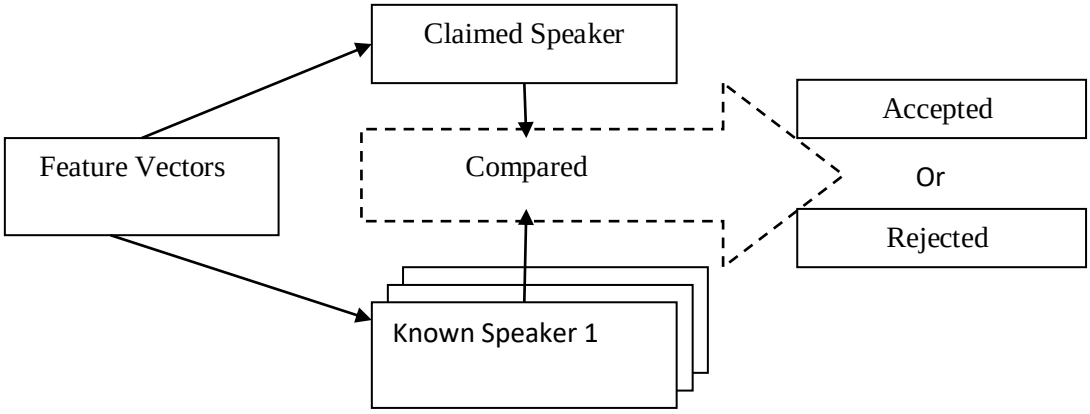


Figure 1.5 Schematic representation of speaker verification system [5].

On the basis of speech modalities, the task of speaker recognition is arranged into two classes: text-dependent speaker recognition and text-independent speaker recognition. In *text-dependent* SRS, system must know all the words or sentences spoken by the speaker. This spoken text is used to train and test the SR system with same customized code. In *text independent* SRS, speaker is not restricted to speak the specific sentences or words. Here system does not have any idea about the speech utterance's content. This system is used for applications where less control over user input is required and their main focus is on verifying the speaker.

1.5 VOICE DISGUISE

Several studies related to disguise voice speaker recognition began in the 1970s. Voice disguise is the process in which a speaker deliberately alters his/her voice in order to obscure one's identity. To create confusion for the human ear or automatic system, offenders attempt to modify or distort their speech by exploiting activities like whispering, falsetto, and foreign accent, change in the rate of speaking, imitation, pinching of a nostril and object in mouth. Hence, the identification rate of speaker recognition system rapidly falls due to voice disguise. Various factors that are responsible for voice alterations are channel variations during voice transmission, emotions, illnesses, electronic scrambling, cold, background noise etc. Voice disguise is often encountered in Indian forensics, where the offender always creates problem in order to conceal one's identity. Such mischievous activities impose severe impact on identification rate. There are two methods of voice disguise: electronic and non-electronic. Electronic voice disguise technique modifies the frequency spectral properties of an original voice by any electronic means like audio editing or voice changing software's. An electronic disguised voice is frequently used by radio stations to obscure one's identity. On the other hand non-electronic disguise disturbs the speech production system by altering the voice tone of a speaker [7] [8]. The imitation can also be performed by changing one's voice to sound like another target person's voice by a trained offender without using any electronic means. The non-electronic disguised voices include whispered speech, raised pitch, creaky voice, and lowered pitch. General non electronic schemes consist of pulling the cheek, pinching of the nostrils, covering the mouth with handkerchief, clenching the jaw, tongue twisting, bite block, etc.

1.6 MOTIVATION

A lot of research on speaker recognition has been going on from last two-three decades and thus making the system need of the hour. For real-time identification of a person, various types of biometric systems are introduced such as face recognition system, fingerprint recognition system, and finger geometry system, hand geometry system, iris recognition system, vein and signature recognition system but the most popular one is speaker recognition system. Many research has shown a dramatic decrease in identification performance of speaker recognition systems when disguised voices are encountered. Voice disguise imposes serious threats to security as it can easily mislead humans and automatic speaker verification (ASV) systems. Research in the domain of speech recognition is always under scrutiny for the accuracy of the systems, raising questions on the parameters to be used for detection, duration and circumstances under which the system performs efficiently. As

long as these questions exist there is a scope of research in this field. This thesis deals in the area of speech processing that investigates the speakers using disguised voice. The disguised voice is frequently used by the criminals to commit crimes [9], [10], [11]. Therefore, it is necessary to find out whether an offender's voice is disguise or not. But there are some issues that are still need to be solved such as disguised or original voice identification, speaker identification and verification even if the disguise voice is encountered. But as stated in [57] that "speaker identification system is essentially unable to accurately recognize the identity of a user when a test sample of user's disguised voice is compared to a reference based on his/her usual speaking mode." Hence, the key role of forensic speaker recognition is to carry out investigation to overturn those circumstances for large and valuable subsets of disguise voice form.

Speaker recognition has attained tremendous results in controlled situations as stated in the NIST annual speaker recognition evaluation. However, real world situations are different from laboratory ones. The performance of the SRS degrades, if there is variation in the speech signal characteristics. It arises from the speakers as well as from the recording environment and communication channels. There are still many problems that need to be solved to address the issue of identifying disguised voices more systematically [12]. The main goal of this thesis to identify the unknown speaker whose voice is disguised. When the original speech sample is disguised via electronic or non- electronic methods, then their spectral based features and distributions are altered. This project uses electronic voice disguise method instead of non-electronic methods. A mathematical analysis of the MFCC based acoustic features is implemented. Voice disguise changes the distributions of a feature vector of an original voice. Therefore, acoustic features helps in differentiating the disguised speech from original speech.

1.7 THESIS OUTLINE

The rest of the thesis is organized in three more chapters. The details are given below:

Chapter 2 starts with the basic terminologies of speech and its production model. Here, various structures of the human organs that produce the speech sounds are also discussed. The mathematical expression and description of voice disguise along with its effects on original speech is well explained. A comparative analysis of speaker recognition along with its applications has been carried out. After reviewing the complete literature, a problem statement is illustrated.

Chapter 3 describes the MFCC based feature extraction technique. Here, first the advantageous role of MFCC is discussed. Then the need of MFCC technique in speaker recognition is highlighted for distinguishing various speech signals. Then a comparative analysis between MFCC statistical moments and existing MFCC techniques has been performed. A brief introduction of support vector machine is presented. This chapter also well explained how a SVM classifier distinguishes a disguised voice from the original voice.

Chapter 4 describes the solution to the problem and display their results that are obtained by applying the proposed recognition algorithm. The algorithm uses a MFCC based statistical moments as a acoustic features and a SVM classifier for classification of a hidden speaker among several known speaker.

Chapter 5 concludes the complete report and describes the future scope for the dissertation.

CHAPTER 2

LITERATURE SURVEY

THE literature review provides a useful structure for almost all the essential techniques for analysis of speaker recognition that have been developed so far. Due to several interventions from the past few decades, many of the researchers have been fascinated in this technology. It evaluates existing literature or available material in order to get knowledge about the topic. The reason for doing so relates to ongoing study. The literature survey may fix a controversy and establish the need for further research.

2.1 SPEAKER RECOGNITION AND FORENSIC IMPLICATIONS

The speaker recognition system aims to extract the speaker-specific content from their speech signal that acts as a biometric tool to recognize an unknown speaker. It can be used for authentication, security control for confidential data, surveillance, and a number of related activities. This technology makes our everyday lives more secure and has becomes an integral part of forensic departments. The speaker recognition's task is segmented into two stages: training stage and testing stage, as described in Figure 2.1.

- In the training stage, user enrolls by providing voice samples to the system. A speech model is formed with the extraction of speaker-specific details from these voice samples of the enrolled speaker.
- In the testing stage, system compares the user's voice with the speech model(s) of the earlier enrolled user(s) to make a decision.

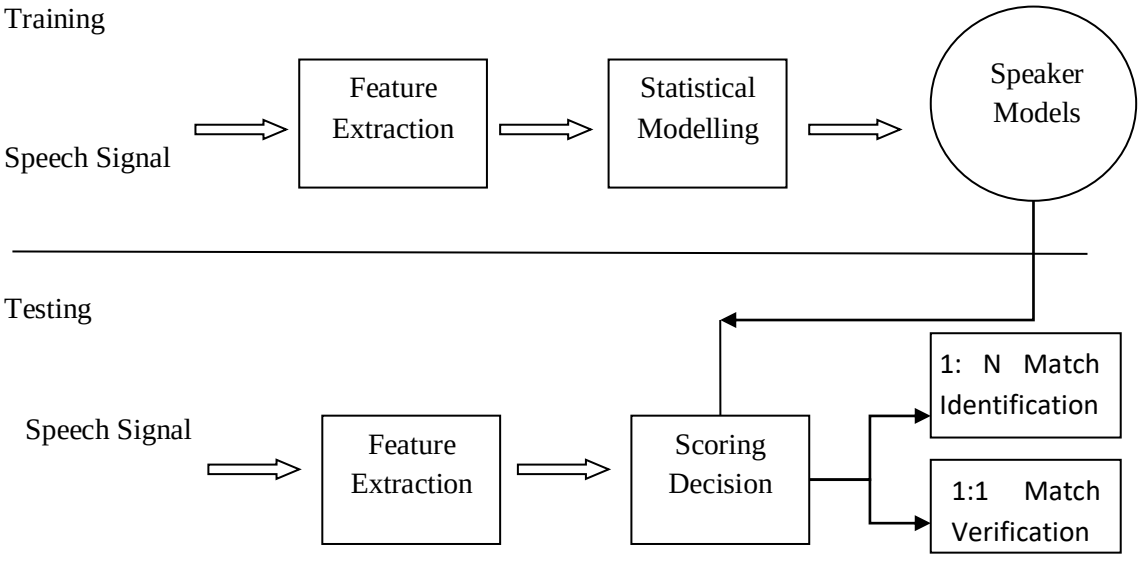


Figure 2.1 Two different phases of speaker recognition system [13]

Research in the field of speaker recognition commence in the early 1970's. **Kersta** [14] introduces a term named voiceprint identification that allows reliable detection of the original speaker by analyzing their speech spectrograms. At the time technology was not adequately developed to aid the process. It leads several people to believe that it is just as trustworthy as the details of the fingertips but the results are limited even with human expert's interpretation. A number of research projects have been developed with the advancement of computer technologies in the post-1960s. The major drawback of voiceprint algorithm is that the spectrograms from same person produce huge intra-speaker variations, because of the fact that no speakers can produce identical speech utterances, hence it is not a convenient technique for aural-perceptual methods, and thus it moved to the visual domain.

Stevens *et al.* [15] accomplish the task of speaker identification by comparing the error rates of speech utterances both aurally via headphones and visually from spectrum. The study shows that there is a huge variation in the detection scores for individual speakers and there are some speakers whose rates are much more difficult to identify than others.

Bolt *et al.* [16] use a spectrographic pattern to search for the correct person among different. This paper also presents a comparison between voiceprint and fingerprint techniques.

Tosi *et al.* [17] simulate a project on voice identification using acoustic spectrograph technique and make decision to identify the unknown speaker. They tested [5] ideas in this paper and study the effect of five variables on speaker identification. The five variables are i.e. number of known or unknown speakers present in the set, closed or open set test, the context of speech materials, speech transmission scheme, contemporary voice input or non contemporary voice input. They observed two kinds of error: false identification and false rejection. The errors rate is less for closed sets isolated words and contemporary voice input as compared to non contemporary. The result for non-contemporary voice samples is more useful to forensic scenario as it provides direct evidence of the negative effect caused on identification of voice.

Endres *et al.* [18] discuss the major troubles faced during identification of an unknown speaker. They put forth some important question to solve these troubles. The first question is that the formants or pitch of a speaker is changing with the time or not. They investigate the phonemes articulated by a speaker remain invariable during his/her whole life or it depends upon parameters such as age and heath. Another major question of them is whether the original voices spectrogram has different formants as compared with disguised voices or not. As in the case of the imitators who try to adjust the pitch of their voice to that of the target

person by comparing the spectrogram of familiar speaker's voices with their own spectrogram. The results have shown that imitators are unable to achieve exact frequency position as compared with the target speaker frequency position. Hence, the sound and mean of the pitch frequency is not only the factor that helps in identifying the imitators. Various other factors like intonation, speech dynamics, loudness, classic phrases and dialects play a predominant role to achieve imitation but they are difficult to characterize or define in spectrograms.

Hollien *et al.* [19] [20] developed a speaker identification model based on long-term statistical measures of speech. They use three different speaking situations: normal speech, stressful speech and random speech. Then the group of listeners who is well-known to the speakers recognizes well and performs considerably better under all circumstances however recognition is by no means always perfect as describes by **Ladefoged *et al.*** [21].The rate of recognition falls severely if utterances are short and belong to relatively large open set. An impact of utterance duration on the recognition performance of familiar speakers has also been discussed in **Koenig *et al.*** [22]. They conduct 2000 forensic comparisons using the spectrographic or voiceprint technique, under actual forensic scenario. They all are FBI examiners who identify the error rates of spectrographic voice identification.

Kunzel [23] discusses the importance of the SR technology in forensics scenario and also clarify the fundamental differences between forensic and other SR applications. They put forth the detail descriptions of three different approaches to analyze forensic speaker identification namely, auditory based speaker recognition; visual based inspection of speech spectrograms, semi-automatic and computer-aided speaker recognition system but system is restricted by real world conditions.

2.2 FEATURE EXTRACTION

This specific module revolves around the features of speech signal associated with the frames obtained as a result from the Pre-processor. Each received frames is processed by applying numerous mathematical signal computations like logarithms, Fourier transforms etc. as well as their variants and combinations. These mathematical functions are representative of the various speech features. The module executes the functions and computations involved for calculating each of the mentioned features. The extracted features are either having fixed dimension or multidimensional. A feature vector comprising all of these features is extracted for each frame. It improves the storage and processing efficiency of the speech signals. There exist many feature extraction techniques such as Linear Prediction Coefficients (LPC) or

Linear Prediction Cepstral Coefficients (LPCC), Me1 Cepstral Coefficients and the Discrete Wavelet Coefficients.

Picone [24] presents voice modelling procedure for analyzing speaker recognition system. In this paper, techniques like speech spectral shaping, feature extraction, parametric mapping and statistical signal modelling are presented. **Guyon** [25] defines the role of feature extraction technique that act as an efficient process for determining the information about the signal while discarding signals the other unwanted signal like noise. They improve prediction rate of the system and provide fast and cost effective search methods.

Atal *et al.* [26] introduce a technique named linear prediction in the 1970's and becomes the dominant algorithm during the early 1980's. Both Fourier transform and linear prediction are well-known techniques and are widely used in speech processing areas. In linear predictive coding (LPC), system predicts the present speech sample from the linear combination of its past samples. The methodology behind the usage of LPC is to reduce the error between original and approximate speech signal at a fixed interval of time. He evaluates the performance of LPC based ASR system using same speech content recorded in different environment such in a quiet room, from dialled up telephones and via a suction cup tap. To reduce the complexity for speech analysis, an all pole models for speech production system is considered.

Oppenheim *et al.* [27] [28] define cepstrum as the inverse Fourier transform of the log magnitude spectrum of a signal. They introduce homomorphism signal processing technique that aims at separating the excitation signal from the vocal tract shape. This analysis splits the speech into two components: excitation source signal and a vocal tract impulse response. These components provide information regarding pitch and vocal tract parameters.

Rabiner *et al.* [3] defines a Linear Predictive Cepstral Coefficients (LPCC) technique for extracting spectral features. It is a combination of LPC method along with the de-correlating nature of cepstrum. However **Li *et al.*** [29, 30] explain a major problem of linear Prediction Cepstral Coefficients features that can work well only in a noise-free environment. Its spectral envelope shows a huge spectral distortion in noisy conditions.

Shaughnessy [31] introduces a critical band filter bank technique that contains bank of FIR band-pass filters. These filters are represented linearly along the *Mel* or *Bark* scale. *Bark scale* are a critical band rate scale whereas *Mel* scale are perceptual frequency scale as discussed in below equations:

$$Bark = 13\,atan\left(\frac{0.76f}{1000}\right) + 3.5\,atan\left(\frac{f^2}{(7500)^2}\right) \quad (2.1)$$

$$melfreq = 2595\,log\,10(1 + \frac{f}{700}) \quad (2.2)$$

The selected bandwidths should be equivalent to a critical bandwidth for subsequent centre frequency. Mathematically, critical bandwidth is expressed as follows:

$$Critical\ BW = 25 + 75\left[1 + 1.4\frac{f}{(1000)^2}\right]^{0.69} \quad (2.3)$$

He also define Mel cepstral coefficients mfcc^n [m] as an alternative approach for extracting acoustic features. He also explains the concept of Cepstral analysis and its use in the Mel Frequency Cepstral Coefficients (MFCC).

Davis *et al.* [32] formulated a novel form of cepstrum illustration that has come to be widely used and known as the MFCC. They perform a frequency analysis using filter bank technique that follows *Mel* scale. In this scale, filters are arranged linearly from 100 to l000 Hz and logarithmically above 1000 Hz. Each filter has a linear phase so as to make the zero group delay for all filters in a filer bank. Usually the resulted signals generated from the filters will be time synchronized. The equation for a linear phase filter operation is described as follows:

$$si(n) = \sum_{-\frac{NFBi-1}{2}}^{\frac{NFBi-1}{2}} \alpha FBi(j)s(n+j) \quad (2.4)$$

where αFBi(j) denotes the jth coefficient for the ith critical band filter. Here, firstly the power vector for each speech frame is obtained that is further joint with other components to create a signal observation vector. Next, the discrete cosine transform (DCT) of the log magnitude of the Mel scale filter bank outputs is performed for each frame to obtain mfcc^n[m], i.e.

$$mfcc_{n^{\wedge}}[m] = \frac{1}{R}\sum_{r=1}^{R} log(MFn^{\wedge}[r])\,cos\left[\frac{2\pi}{R}\left(r + \frac{1}{2}\right)m\right] \quad (2.5)$$

2.3 FEATURE CLASSIFICATION

The algorithm used in classification stage of speaker recognition depends upon the technique adopted. There are two types of training algorithms used in classification techniques; supervised and non-supervised. In supervised algorithm, the data provided to the

speaker model include information regarding class by means of a label. On the other hand, unsupervised algorithms do not need such kind of information for the training voices. Some of the unsupervised methods involve nearest neighbour (NN), Gaussian mixture models (GMM), vector quantization (VQ), and hidden Markov models (HMM), supervised vector machine (SVM). In this project SVM classification is used for modelling the different speakers. The task of speaker classification can be categorized into two types: text-dependent and text-independent.

- *Text-dependent speaker recognition*

The text-dependent system has prior information of the script to be spoken by a speaker and it is assumed that a speaker will cooperatively articulate this text. It uses DTW and HMM to model the speech features in speaker recognition tasks.

Fururi [6] discusses the concept of text-dependent and text-independent methods of speaker recognition and its various types of available processing techniques. In [33], he explains a simple and effective technique of pattern matching using Dynamic Time Warping (DTW) that has been successfully used as a non-statistical tool in several text-dependent tasks of speaker recognition. In this technique, each spoken word is depicted by a set of feature vectors, and utterances of the same word is normalized by aligning the analyzed feature set of a test speech sample to the template set of feature vector via a DTW algorithm. Then decision is made by calculating the distance between the test input and the template. In dynamic time warping (DTW)**,** strings of features matched from speech under analysis are shifted in time compressed and expanded to match to stored templates of those features from each candidate speaker. Early text-dependent speaker recognition also used template matching techniques for text-dependent speaker recognition. But nowadays text-independent speaker recognition applications are focused and techniques like template matching are no longer used.

Naik *et al.* [34], **Rosenberg *et al.*** [35], **Zheng** [36] compares the performance of dynamic time warping (DTW) and hidden Markov model (HMM). Their results show huge improvement in recognition rates while using HMM as compared with DTW. HMM matches the statistical similarities between a speaker and candidates.

- *Text-independent Speaker Recognition*

The text independent system has no prior information about the text to be spoken by a speaker, like using an extemporaneous speech. This is more complex but a flexible technique. Markel et al. uses the long-term statistical approach of analyzing various spectral features,

such as the mean and variance of spectral features over a series of utterances. In this algorithm a text-independent approach is used. He well describes the matching of the feature tendencies of one speaker with the other candidates.

Li et al. [37], **Matsui *et al.*** [38], **and Rosenberg *et al.*** [39] present the second most used classifier based on vector quantization (VQ) method. This method includes VQ codebooks which consist of a small amount of feature vectors, which are further used as an effective way of obtaining speaker-specific features. It carries features from one speaker, and matches this against the stored codebook from candidates. A codebook is generated by clustering the feature vectors of each and every speaker. In the recognition step, an input speech is vector-quantized with the help of codebook of each reference speaker and the VQ distortion accumulated over the entire input utterance is used for making the recognition possible.

Reynolds *et al.* [40] present a speaker recognition system based on Gaussian mixture model (GMM). This speaker model is first evaluated on informal speech experiences various issues while training and performance in accordance with model order, number of training input and size of the population. Several compensation techniques to avoid spectral variability are used to develop sturdiness for telephone voice signal. The results also demonstrate that the GMM outperforms the long-term average and VQ speaker models. For identification, it uses maximum likelihood criteria and for verification system, a likelihood ratio hypothesis test has been followed. The major restrictive factor in this paper is transmission degradations like channel and noise variability.

Oglesby *et al.* [41] describes a new approach to speaker recognition where a neural classifier is used to separate the different speakers. They perform computation and the results shows the neural classifier achieves at least comparable performance to a conventional approach based on vector quantization and codebooks. Detailed comparisons in terms of speaker identification error rates, levels of training, and text dependency are presented.

Campbell *et al.* [42] described an algorithm for modelling high-level speech features using SVM. This technique transforms the feature vectors of one class into a high dimensional space and then split classes with a hyper plane.

Clarkson [43] present a support vector machines (SVMs) approach to pattern classification which has recently become the most robust classification technique for the purpose of recognition. This is one of the most successful techniques in the areas of image identification and face detection. Here results show that SVMs outperform the other nonlinear classifiers like artificial neural networks (ANN) and k-mean nearest neighbours

(KNN). Results of several vowel and phonetic classification using SVM show much better performance than with the Gaussian mixture classifiers.

2.4 EFFECT OF VOICE DISGUISE ON SPEAKER RECOGNITION

Several studies related to disguise voice speaker recognition began in the 1970s. Voice disguise is defined as the process in which a person make deliberate attempt to alter his/her voice in order to obscure one's identity. Speakers alters their voices by doing activities like whispering, falsetto, and foreign accent, change in the rate of speaking , imitation, pinching of a nostril and object in mouth. Voice disguise is often encountered in Indian forensics, where the offender always creates problem in order to conceal one's identity. Such mischievous activities impose severe impact on identification rate.

Reich *et al.* [44] perform spectrographic speaker-identification to examine the effects of certain vocal disguises via listening. A group of listeners are selected who decides whether two sentences is spoken by the same speaker or by different one. They achieve moderately high degree of accuracy in discriminating speakers when the speaker's voice is undisguised however addition of a disguised voice drastically degrades the listener performance.

Fururi [45] highlights many advances and progress achieved from speaker recognition technology along with various problems for which accurate results remain to be found. The problem mainly arises when there is variation in speech due to factors like transmission channel, emotions and recording environments, age, health issues, speaking manner, speaking rate and level. This paper helps to solve problems like distortion due to channels variations and other environmental surroundings.

Rodman [12] researches on speaker identification or verification of disguised voices. He briefly explains the concept of disguising along with its creation of databases and how it is tested on conventional systems of speaker recognition using database of disguised-voices. He investigates several classification methods to model the vocal tract.

Tan [9], **Zhang *et al.*** [46], **Perrot *et al.*** [11] and **Künzel *et al.*** [1] revealed the effect of several voice disguises methods with raised and lowered pitch on speaker recognition by both human and machine including hand over mouth, pinched nostril, high pitch and low pitch, and conducted experiments using Gaussian Mixture Model (GMM), Vector Quantization (VQ) and Support Vector Machine (SVM) classifiers for automatic speaker verification. However, no detailed solution is given.

Zhang [46] matches features like phoneme duration, fundamental frequency, long term average spectrum (LTAS), and intensity, with original speech. He compares the effect of low-

pitched voice disguise on automatic speaker recognition (ASR) with high-pitched voice disguise. **Perrot *et al.*** [11] present the effect of voice disguise on automatic speaker recognition and, also discuss a statistical approach, to identify disguises under restrained recording situations.

Künzel *et al.* [1] present the effect on speaker recognition performance of three commonly used non-electronic methods i.e. pinching of the nostril, increased pitch and lowered pitch. They use likelihood ratio (LR) based forensic speaker recognition system designed by ATVS-UPM.

Jin *et al.* [48] investigated voice disguise for speaker re-identification. However, concrete solutions for speaker recognition against voice disguise or detection of disguised voice have not been reported.

Wang *et al.* [49] **and Wu *et al.*** [50, 51] identify disguised voices from original voices. Speech database like cross-corpus and cross-disguise-method are used in testing. In all test conditions, the rate of detection is 90%. However, there is a need to identify hidden speaker using disguised voice.

2.5 PROBLEM STATMENT

Although the rapid advancement in the field of speaker recognition technology is occurring but there are still various issues to be solved. The problem arises when the disguised voices are encountered for the purpose of identification. Due to this problem, performance of speaker recognition (SR) systems is severely degraded. It has adverse threats to security where a person intends to alter his/her voice in order to deceive the listener. The amount of degradation varies with different disguising factors. When a speaker deliberately makes anonymous calls, ransom calls and threatening calls, one wants to know about the offender's identity. Hence, firstly we have to check whether a given voice is disguised or not before putting into an ASR system. Thus, the first and foremost step of such a system is identification of disguised voices but several studies show that the concrete solutions against voice disguise detection are still very insufficient and provide a scope for potentials improvements.

Up to now, very few efforts have been accounted on disguised voice identification. Several researches have been ongoing from the past few decades to study the effects of voice disguise on ASV systems but they do not offer any robust or complete solutions to expose genuine identity of the speaker. Here, the proposed system is accomplished in two tasks. The first task is to identify disguised or genuine speech. Then the second task is to spot the hidden

speaker. Thus, concrete countermeasures should be taken to eradicate the effects of voice disguise in order to verify the speakers. The proposed system is evaluated via voice conversion algorithms or other audio editor's software. In this project, MFCC statistical moments and support vector machine classifiers are used to separate disguised voices from original voices in order to identify a hidden speaker. These works provide significant contributions to the department of audio forensics.

CHAPTER 3

METHODOLOGY

VOICE disguise is the process in which a speaker deliberately alters his/her voice in order to obscure identity. To create confusion for the human ear or automatic system, speakers try to change and distort their voices. There are two types of methods present for disguising a voice i.e. non-electronic and electronic [7]. In this research, electronic disguise method is used which is described in detail as follows.

3.1 ELECTRONIC VOICE DISGUISE

As discussed in previous section 2.3, an electronic voice disguise aims at modifying the speech parameters such as pitch, formants and rate. Voice re-sampling is an effective technique that is used to modify the pitch of a continuous-time speech signal x(n). Here stretching or compressing alters the pitch of the signal [48]. The output of this technique will be a re-sampled signal denoted as x'(n). Let $X(\omega)$ denote the frequency spectrum of $x(n)$, and $X'(\omega)$ the frequency spectrum of $x'(n)$. Voice disguise can be defined as a voice re-sampling method with time-scale adjustment. Suppose the original speech signal $x(n)$ having duration D and pitch P, be re-sampled by a factor $1/\alpha$ to obtain the another signal $x'(n)$. The resulted re-sampled signal $x'(n)$ obtained has duration D' and pitch P'. The duration D of the original signal changes when its pitch P changes. This modification in duration D often results in change of the speed of the voice signal. The following mathematical equation explains the relationship between the original and the resulted re-sampled signal.

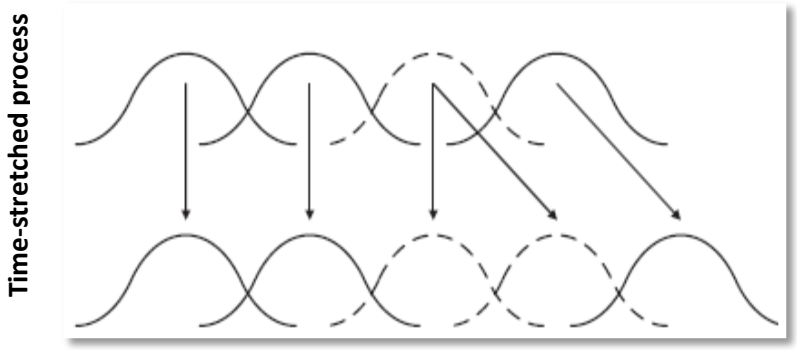
$$X'(w) = \frac{1}{\alpha} X\left(\frac{w}{\alpha}\right) \tag{3.1}$$

$$P' = \alpha P \tag{3.2}$$

$$D = \frac{D'}{\alpha} \tag{3.3}$$

The resulted re-sampled signal either becomes too fast or too slow when compared to the original signal $x(n)$. To adjust the duration back to D from D' time-scale modification is used [49]. This technique is broadly used in applications like speech synthesis, speech compression, foreign language learning, etc. The idea behind Synchronized Over-Lap Add (SOLA) algorithm is well illustrated in Figure 3.1.

a)



b)

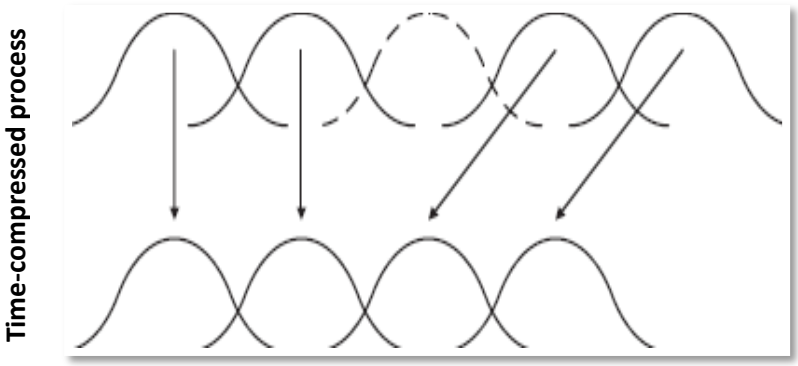


Figure 3.1 Description of SOLA algorithm (a) Time-stretched signal (b) Time-compressed signal [49]

In this technique, firstly the original signal is divided into several short frames. In order to obtain time-stretched or time-compressed speech output, some of their speech frames are repeated or discarded and leaving all other frames unchanged. The frequency spectral properties and pitch of the original signal remains unaffected during time-scale modification. This technique shifts the duration D'' and speed of the voices. The duration D' of the re-sampled signal is adjusted back to the duration D by a factor of α. Voice disguise leads to frequency spectral modification that leaves telltale footprints on the MFCC features. The relationship between time-scale modified signal x*(n)* and original signal in terms of duration D'' and pitch P'' is given as:

$$D" = \alpha \ D' = D \tag{3.4}$$

$$P'' = P' = \alpha \, P \tag{3.5}$$

3.2 MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCCs)

MFCCs represent short term speech features extracted from their spectrum. It is often used for the analysis of acoustic signal. The scheme of the MFCCs is the well-known and accepted. It relates the variation in critical bandwidth of the human ear with regards to mel frequency scale in which filters are arranged linearly at frequencies less than 1000 Hz and

logarithmically above 1000 Hz frequencies. First the signal is segmented into frames and then a hamming window is multiplied with each frame to avoid discontinuities. Thereafter, discrete Fourier transform for each frame is calculated and then log of amplitude spectrum is measured. Finally discrete cosine transform (DCT) is implemented to smooth the speech spectra that further helps in generating cepstral feature vectors for each frame [50]. The essential steps that help in generation of MFCCs are as follows:

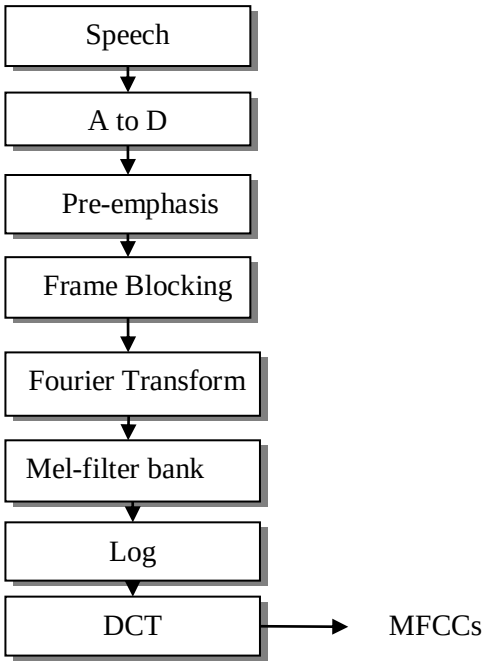


Figure 3.2 Procedure to generate MFCCs [51], [50]

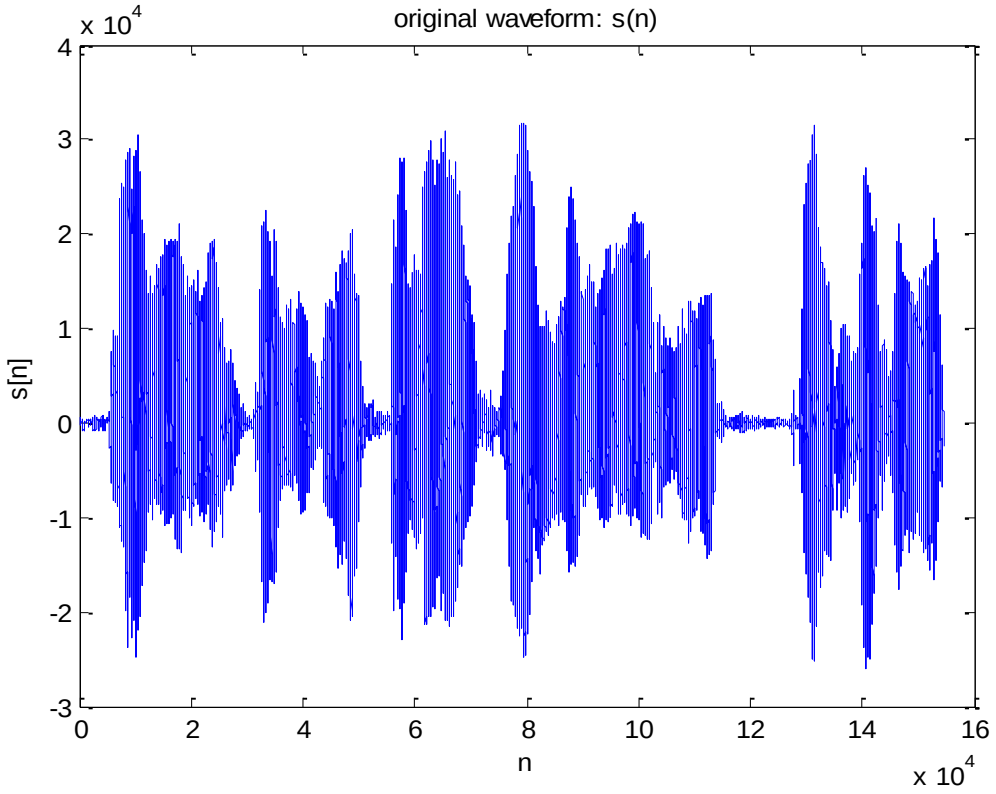
MFCC first converts an analog speech waveform to digital speech. To minimize the aliasing effects in speech, its rate of sampling is greater than or equal to 10000 Hz. The following steps are used to generate MFCC feature vectors:

- *Pre-emphasis*

This stage amplifies the specific portion of the spectrum where hearing becomes sensitive. So, it boosts up only the high frequency region of the spectrum as sufficient energy is present at the lower frequency region. Thus, pre-emphasis plays an important role in spectral analysis. Before spectral investigation, it equalizes the natural slope, thereby improving the accuracy of analysis. Figure 3.3 shows a time domain speech representation of some sentence before and after pre-emphasis.The pre-emphasis filter is a high-pass filter of first-order. For input s[n] and pre-emphasis coefficient a, whose range varies from 0.9 to 1.0, the time domain equation of the filter is expressed as:

$$y[n] = s[n] - \alpha s[n-1] \tag{3.6}$$

a)



b)

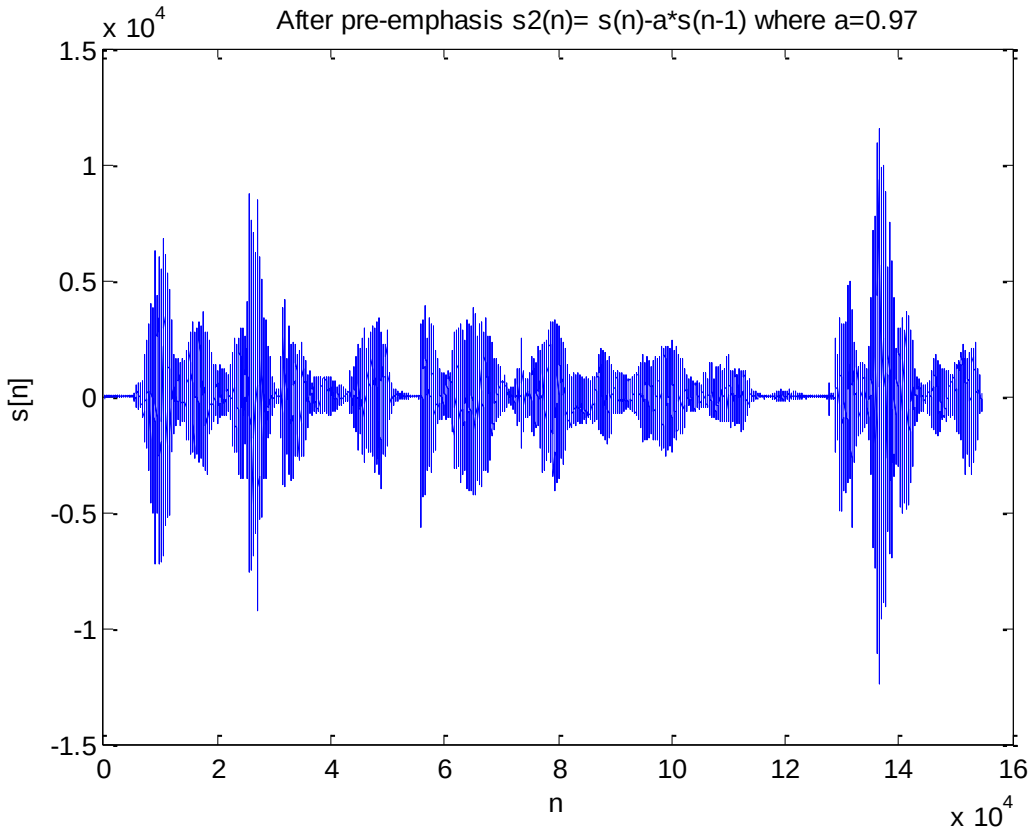


Figure 3.3 Speech waveform a) Before pre-emphasis and b) After pre-emphasis.

- *Framing*

A speech signal is constantly varying so for simplification it is assumed that the signal doesn't change much on short time scales. Hence, speech signal is segmented into several frames of duration 20-40ms as shown in Figure 3.4. If the size of the frame is much smaller, then it requires enough samples to obtain a robust spectral estimate but if its size is longer than usual, then signal varies greatly throughout the frame.

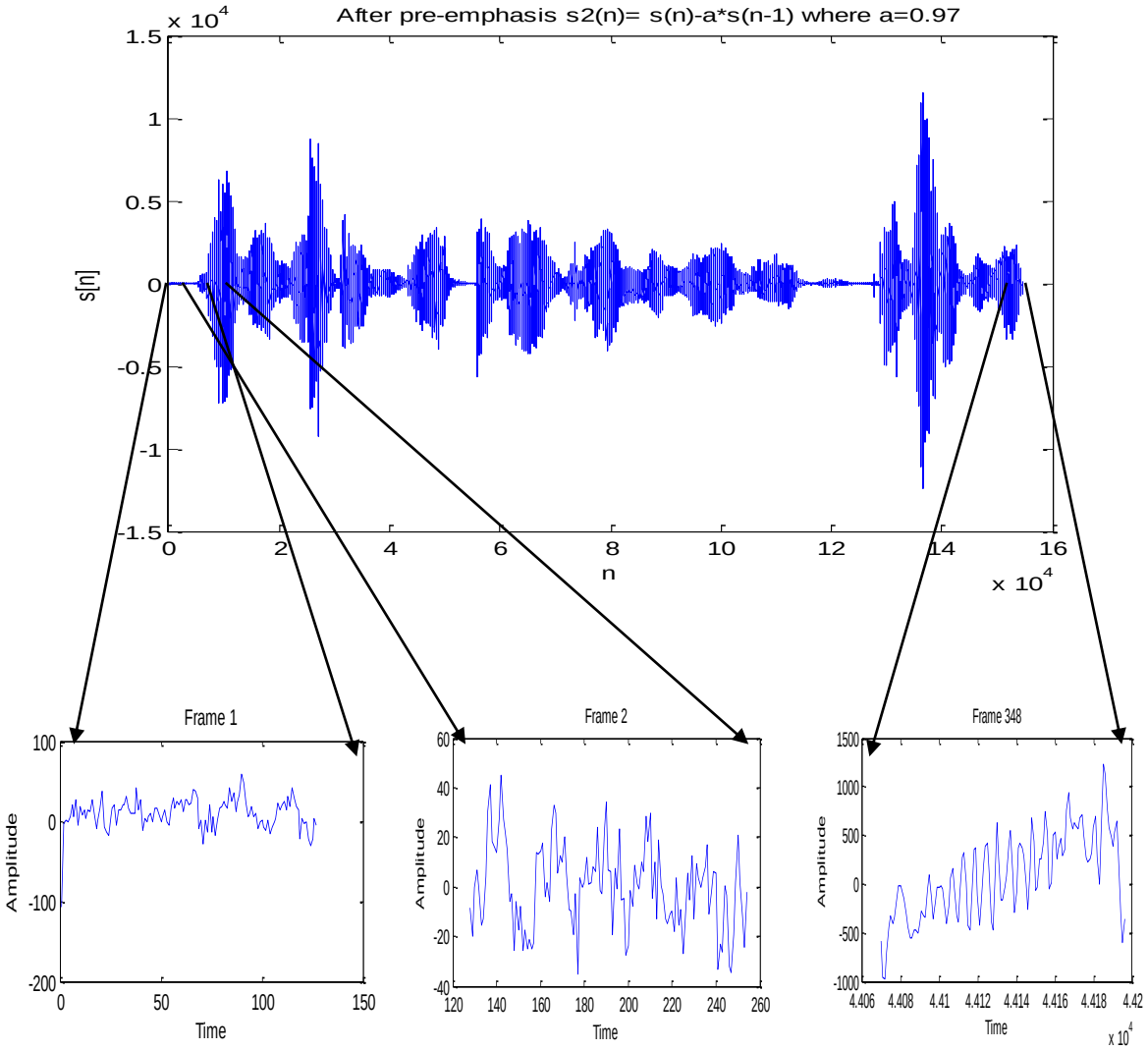


Figure 3.4 Framing of speech signal.

- *Windowing*

Windowing carefully exploits the window to make the signal equivalent equal to zero at the beginning as well as end. It reduces the spectral distortion and signal discontinuities. The speech obtained from each window frame depends on frame size and frame shift as shown in figure 4.6. Frame size is defined as the length of the frame in milliseconds (usually 25 ms) whereas frame shift is the number of milliseconds linking the left ends of consecutive

windows. Mathematically, a resultant signal is extracted by multiplying the original signal s[n], with the window w[n], at time n.

$$x[n] = s[n].w[n] \tag{3.7}$$

As we know the rectangular window is the simplest window which abruptly slices of the signal at its edges. However, the abrupt nature causes discontinuities and creates trouble in Fourier analysis of signal. This is why a Hamming window is used for extracting MFCC features, which shifts the signal values toward zero at the boundaries to avoid discontinuities. Mathematically, the rectangular and hamming windows are represented by:

$$wR[n] = \begin{cases} 1 & 0 < L < n-1 \\ 0 & otherwise \end{cases} \tag{3.8}$$

$$wH[n] = \begin{cases} 0.54 - 0.46\,cos\left(\frac{2\pi n}{L}\right) & 0 < L < n-1 \\ 0 & otherwise \end{cases} \tag{3.9}$$

- *Fast Fourier Transform (FFT)*

In this stage, the spectral specific details for our windowed signal are extracted. There is a need to know the amount of energy present in the signal at different frequency regions. This stage performs inter-domain transformation for each N frame samples of every frame from time to frequency. Thus, Discrete Fourier Transform (DFT) technique is used for an array of N samples i.e. {xn}. The magnitude spectrum of speech sample is depicted in figure 4.6. Fast Fourier Transform or FFT is an efficient and simple algorithm than DFT. The only problem with DFT is that it only works for those values of N that are powers of two. Mathematically, DFT can be defined as:

$$X_i(k) = \sum_{n=1}^{N} x_i(n) e^{-j2\pi kn/N} \qquad 1 \le k \le K \tag{3.10}$$

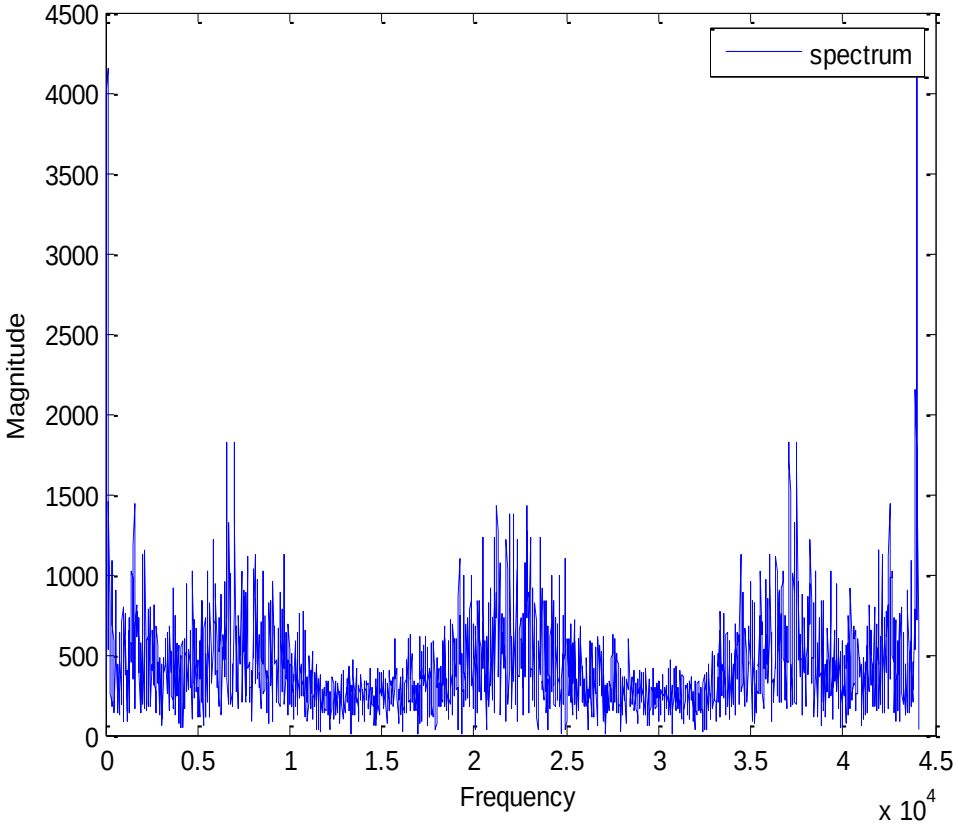


Figure 3.5 Magnitude spectrum of the windowed speech signal.

- *Mel frequency warping*

Human perception does not accept a scale that is linear, both for the speech signals or for tones that are pure. Every voice signal has a frequency f and a pitch P which is subjective in nature, on a scale known as 'Mel' scale. Thus, this graph shows the relationship among original and perceived pitch frequency. It follows linear scale when the frequency is below 1000 Hz otherwise follows logarithmic scale when the frequency becomes greater than 1000 Hz. Since the logarithm of the linear frequency is proportional to human perception level, Mel scale tries to replicates the same effects of human's auditory system.

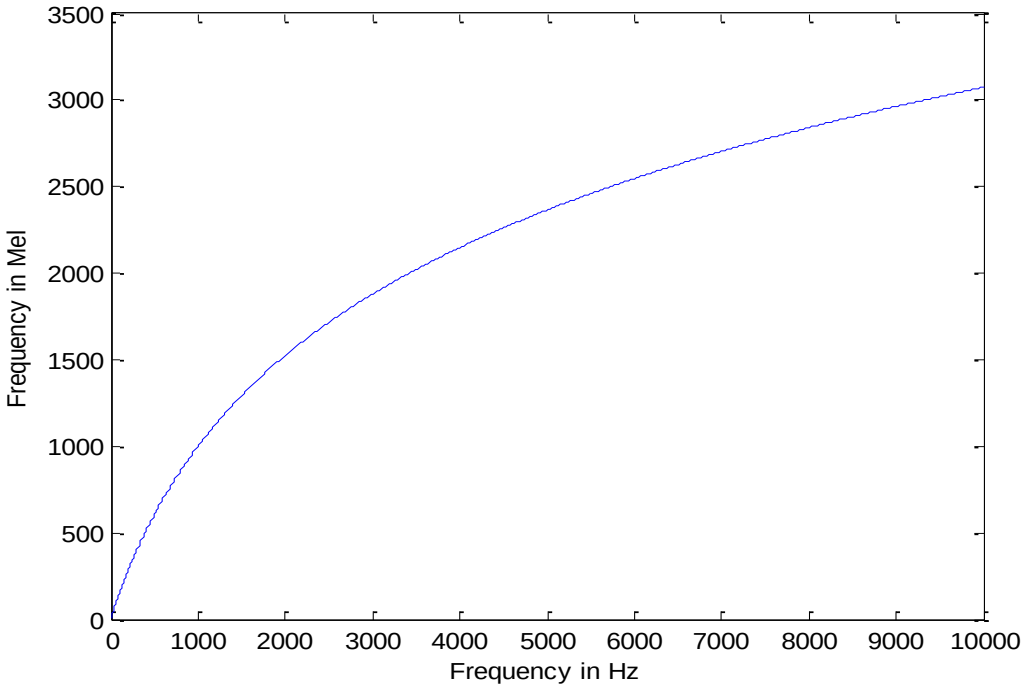


Figure 3.6 Conversion normal frequencies to Mel frequency.

- *Mel Filter banks*

The first filter bank begins at the first frequency point. When the first filter reaches to peak, second frequency point begins and then the filter will be come back at zero at the third frequency point. Then the next filter bank will start from the second frequency point. Similarly this filter achieves peak value at the third frequency point, and in the same way it returns to zero at the fourth etc. They intersect at the position of the boundary of each filter. A simple formula for obtaining these is as follows:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \ \ and \ k > f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \ \ \leq k \leq f(m) \\ \dfrac{f(m-1) - k}{f(m-1) - f(m)} & f(m) \leq k \leq f(m-1) \end{cases} \qquad (3.11)$$

In order to obtain Mel filter banks, ten filters are spaced linearly at frequencies less than 1000 Hz and rest of the filters is positioned logarithmically above 1000 Hz frequencies. These filters gather energy from each band of filters.

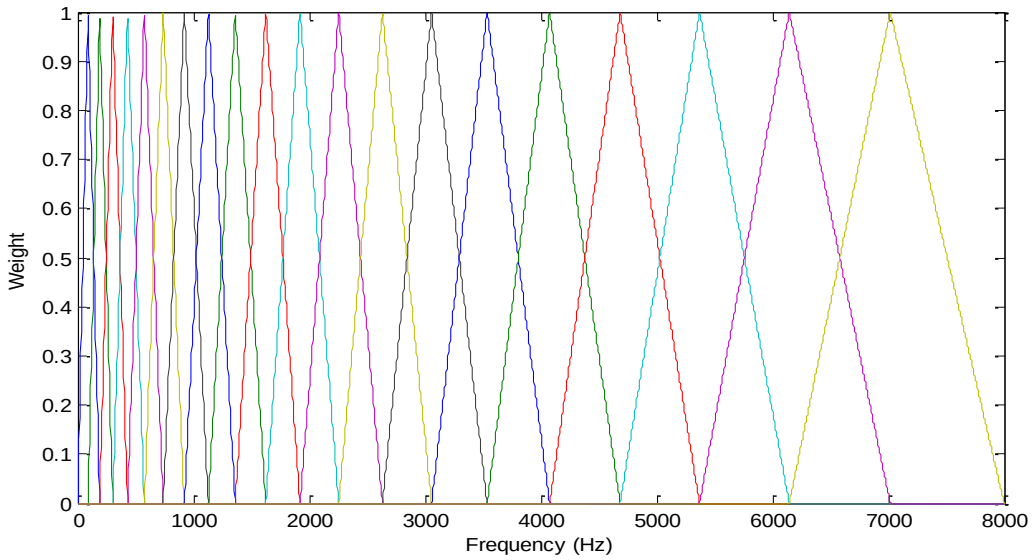


Figure 3.7 Plot of Mel filter-banks.

The above figure contains 20 triangular band pass filters. The filters are placed at regular interval along the Mel scale frequency, which is expressed as:

$$mel(f) = 2595 \times log_{10}\left(1 + \frac{f}{100}\right) \tag{3.12}$$

The purpose for triangular band pass filter is it diminish the size of the features involved and smooth the magnitude spectrum such that the harmonics are flattened to achieve the envelope of the spectrum with harmonics [9].

- *Periodogram*

After Fourier analysis, the power spectrum for each frame is calculated. It is encouraged by an organ named cochlea which is present in the human ear. Cochlea vibrates at different spots depending upon the incoming sounds. Depending on the position in the human cochlea, different nerves notify the brain about frequencies of the incoming sounds. This periodogram perform a similar task of identifying frequencies present in the frame.

$$P_i(k) = \frac{1}{N} |X_i(k)|^2 \tag{3.13}$$

- *Logarithmic compression*

At that point, the log of all the values of Mel filter banks is computed. The reaction of a human to a signal is logarithmic as they have less responsive nature. They can't predict slight variation in amplitude of the voice. Here, the advantage of utilizing log is that it extricates the best component to assess the less sensitive input and diminishes the dynamic range by processing the square of the magnitude of each filter bank output.

- *Cepstrum with DCT*

Finally, the logarithmic form of Mel spectrum is reverted back to time. The response of this will create MFCC. MFCC represents the speech spectra in cepstral domain. Figure 3.8 presents the spectrum plot of MFCC. For frame analysis, only first 12 cepstral coefficients are considered. These cepstral coefficients tend to be uncorrelated but this is not true for the spectrum. Spectral coefficients are correlated at different frequency bands.

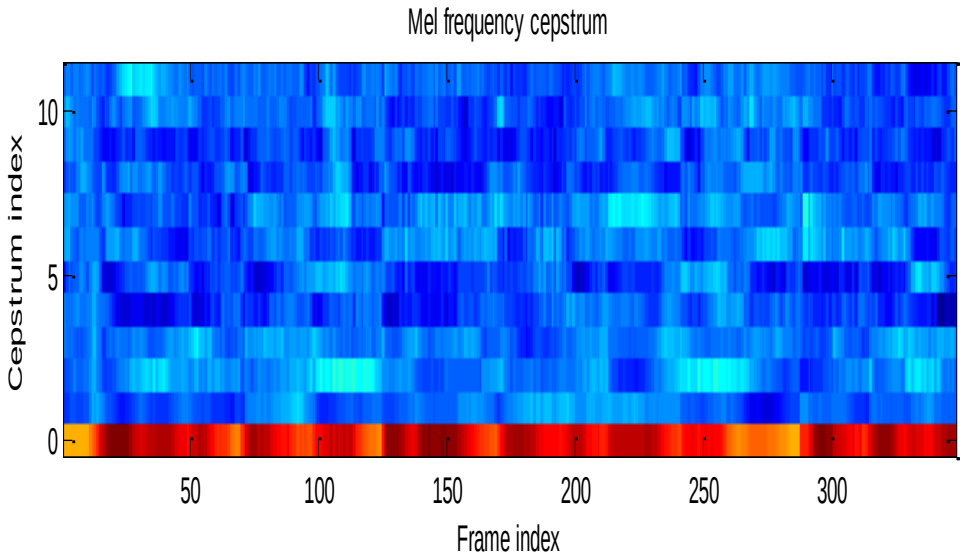


Figure 3.8 Pictorial representations of 12 MFCCs values vs. speech frames

Here, Discrete Cosine Transform (DCT) is performed to transform it into time domain. Mathematically, MFCC can be expressed as:

$$c[n] = \sum_{n=0}^{N} \log\left(\left| \sum_{n=0}^{N} x[n] e^{-\frac{j2\pi kn}{N}} \right|^2 \right) \cos\left(\frac{\mathrm{k(n-0.5)}\pi}{\mathrm{N}}\right) \quad (3.14)$$

- *Deltas and Double Deltas*

The features of MFCC depict the power spectral values of only one frame, but speech contain some details in the dynamics. Here energy correlates with the identity of speaker which is a very valuable cue for speaker recognition. Therefore an energy feature is added as the cepstral coefficients that are not capable of capturing energy [51]. The energy of a frame can be defined as the sum over point in time of the power of the each frame samples; hence the energy of signal x in a window from $t1$ to $t2$ time sample is expressed as:

$$Energy = \sum_{t=t1}^{t2} x^2(t) \quad (3.15)$$

The speech signal varies from frame to frame. This variation, like a formant's slope at its transitions, or to stop burst from a stop closure, offer a helpful cue for phone recognition. This is why these features are added which basically depicts a change in cepstral features vs. time. This is achieved by adding a delta feature, and a double delta features. All the 13 delta features signifies the modification among frames in the subsequent cepstral/energy feature and each 13 double delta features symbolizes the change between frames of delta features.

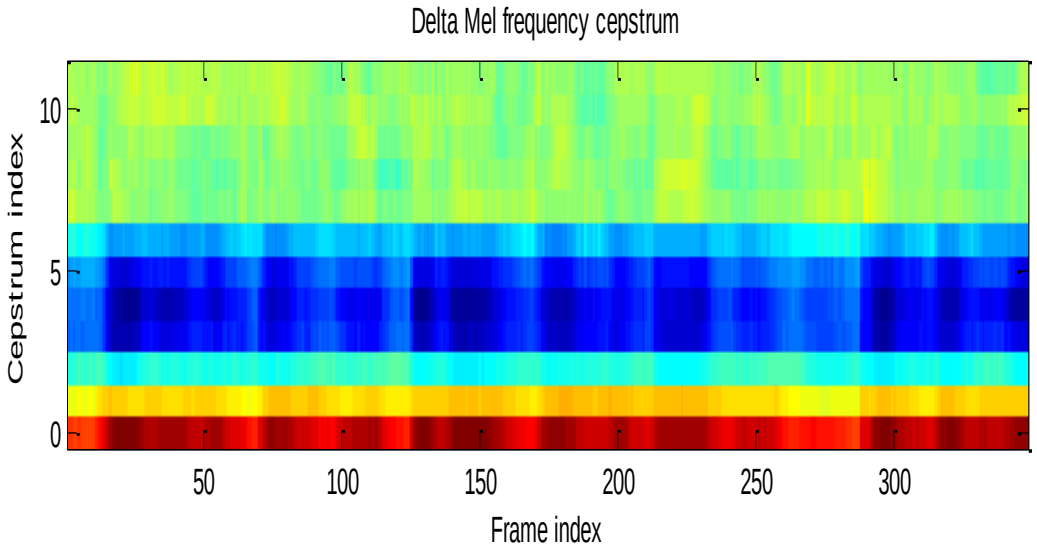


Figure 3.9 Delta representation of MFCC

The delta features and double delta can be simply extracted by computing the difference among frames. The delta features $d(t)$ at time t for some cepstral value $c(t)$ can be expressed as:

$$d[n] = \frac{\sum_{n=1}^{N} n\big(c(t+n) + c(t-n)\big)}{2\sum_{n=1}^{N} n^2} \tag{3.16}$$

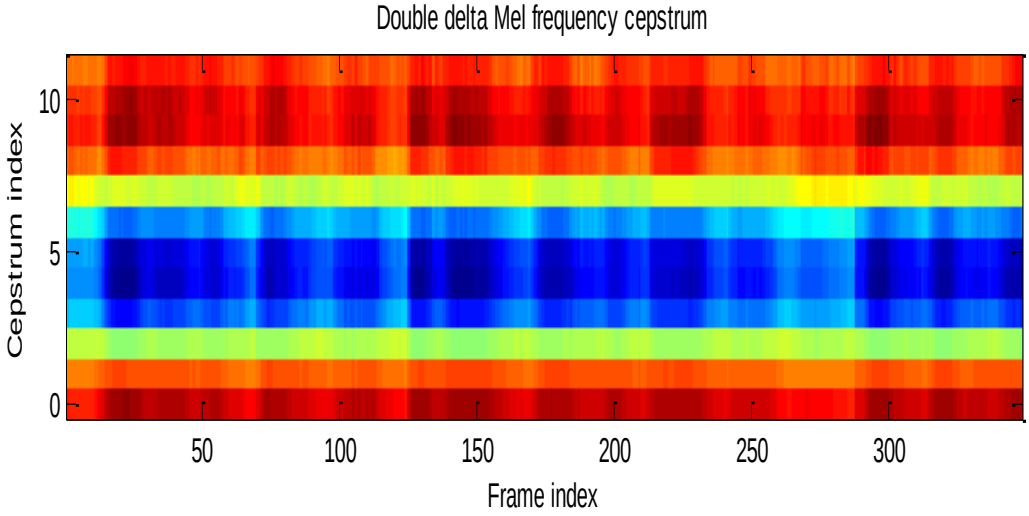


Figure 3.10 Double Delta representation of MFCC

The major advantage of Mel frequency cepstral coefficients (MFCC) is that these cepstral coefficients have a tendency to be uncorrelated, which makes the system easier. MFCC features of each frame represent only the power spectral envelope of its subsequent frame. The speech specific details in the dynamics are extracted from derivative coefficients. Each delta features signify the variation among frames whereas each double delta features symbolizes the variation among the frames in the subsequent delta features attributes.

- *Acoustic Features*

As the number of MFCC features changes with the duration, statistical moments extract acoustic vectors with the similar duration [7]. Here *x(n)* is a speech signal with *N* frames whose MFCC vector is denoted as *vij* where *j* represents the feature component and *i* depicts the frame number, and it can be expressed as:

$$V_j = \{v1j, v2j, \dots\dots, vNj\};\ \ where\ j = 1,2, \dots\dots L \tag{3.17}$$

In this project two types of statistical moments are used. Firstly the mean *Ej* of each MFCC features *Vj* is extracted and then the correlation coefficients $CR_{jj'}$ among different MFCC features V_j and V_j, are considered. The procedure is described in Equation (3.18)and Equation (3.19) respectively.

$$E_j = E(V_j);\ \ j = 1,2, \dots\dots L \tag{3.18}$$

$$CR_{jj'} = \frac{cov(V_j, V_{j'})}{\sqrt{var(V_j)}\sqrt{var(V_{j'})}};\ \ 1 \le j < j' \le L \tag{3.19}$$

The resulting mean values E_j and correlation coefficients $CR_{jj'}$ are united to create the statistical moments W_{MFCC} of MFCC vectors as described in below Equation:

$$W_{MFCC} = (E_1, E_2 \dots\dots E_L, CR_{12}, CR_{13}, \dots\dots CR_{L-1L}) \tag{3.20}$$

Similarly, the statistical moments of delta MFCC $W_{\Delta MFCC}$ and double delta MFCC $W_{\Delta\Delta MFCC}$ are extracted. Finally, by combining W_{MFCC} , $W_{\Delta MFCC}$ and $W_{\Delta\Delta MFCC}$ an acoustic feature *W* of *x(n)* is created. *W* can be expressed as follows:

$$W = [W_{MFCC}, W_{\Delta MFCC}, W_{\Delta\Delta MFCC}] \tag{3.21}$$

3.3 FEATURE CLASSIFICATION

The problem of pattern recognition in speaker recognition has always been a great deal of interest for several past decades. Basically, the pattern recognition maps items of interest into a number of classes. For speaker recognition systems, each class depicts unique speaker. The

above presented topic well explains the MFCC procedure in which a speech signal is transformed to a set of vectors. Then after this step's completion, a speaker model should be created that is further used for speaker classification tasks. Here different classifiers like support vector machine, decision tree, naive bayes and linear discriminant analysis are used for making decision regarding relationship between the input test voice and the speaker(s) voice contained in model.

- *Support Vector Machine (SVM)*

SVM classifiers are based on supervised learning that requires training prior to classification. SVM receives a sequence of data and estimate each data inputs and further categorized it into two possible classes. With the help of an SVM, one can easily resolve the problems faced in pattern recognition tasks. SVM is constructed from the hyper plane in the linearly independent case. This hyper plane segments the classes and maximizes the minimum distance among classes [53], [54], [55]. A pictorial representation of an optimized decision-making SVM classifier is shown in Figure 3.11. A basic approach of a SVM is data classification. SVM defines a hyper plane that split all data points of first class from the other class. The ideal hyper plane for an SVM will have the highest margin among the two classes. Here margin means the maximal distance across the portion parallel to the hyper plane.

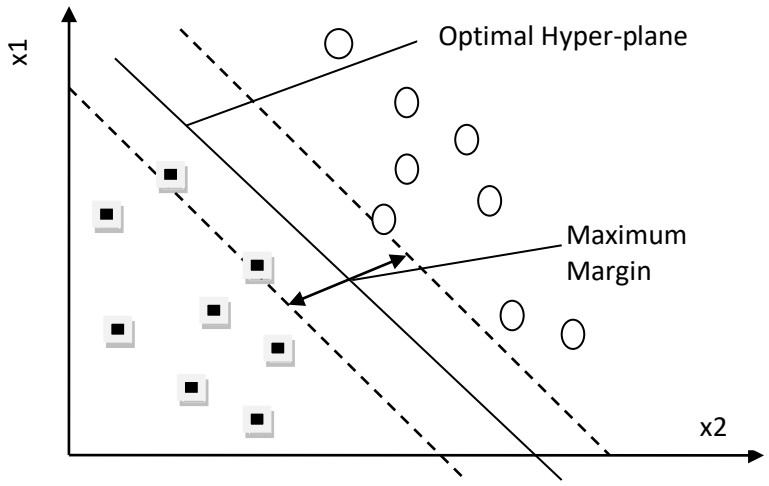


Figure 3.11 Description of SVM algorithms [53].

- *Decision Tree*

A decision tree classifier is a tree like structure, where all inner nodes represent a test on attributes, each branch denotes a result of the test, and leaf nodes correspond to classes or their distributions [55]. The illustration for decision tree classifier is a binary tree. These

classifiers learn fast and also make fast predictions. These predictions are more often accurate for a large range of problems.

- *Naive Bayes (NB)*

Naïve Bayesian classifiers assume that each input attributes is independent; hence it is named *naïve* [55]. It is composed of two kinds of probabilities which can be directly calculated from training data: probability of class and conditional probability of class given some value. Once these calculated, probabilities are used further for making predictions for the new data via Bayes Theorem. The technique is more effective for a broad range of problems as compared with other.

3.4 PROPOSED METHODOLOGY

The proposed identification system uses MFCC statistical moments and SVM classifiers for disguised voice modelling. A voice can be disguised with the help of multiple disguising factors. The consequence of each disguising factor on MFCC is different. Here we select ±4, ±5, ±6, ±7, ±8 disguising factors in audacity software for pitch alterations. The extracted feature vectors are useful in disguised voice identification that is disguised with different factors. Hence, all the original voices and the voices disguised via each factor are used to train SVM classifiers. In an ideal case, the SVM classifier fully separates the original and disguise voices.

When an original voice is placed into an SVM classifier, the result will be an original voice but if a disguised voice with R semitones is put into that classifier then the result of the original and disguised voice will be different. Hence, such a technique helps in discriminating disguised voices from original voices. In this work, the concept of disguise voice identification is merged with the speaker recognition.

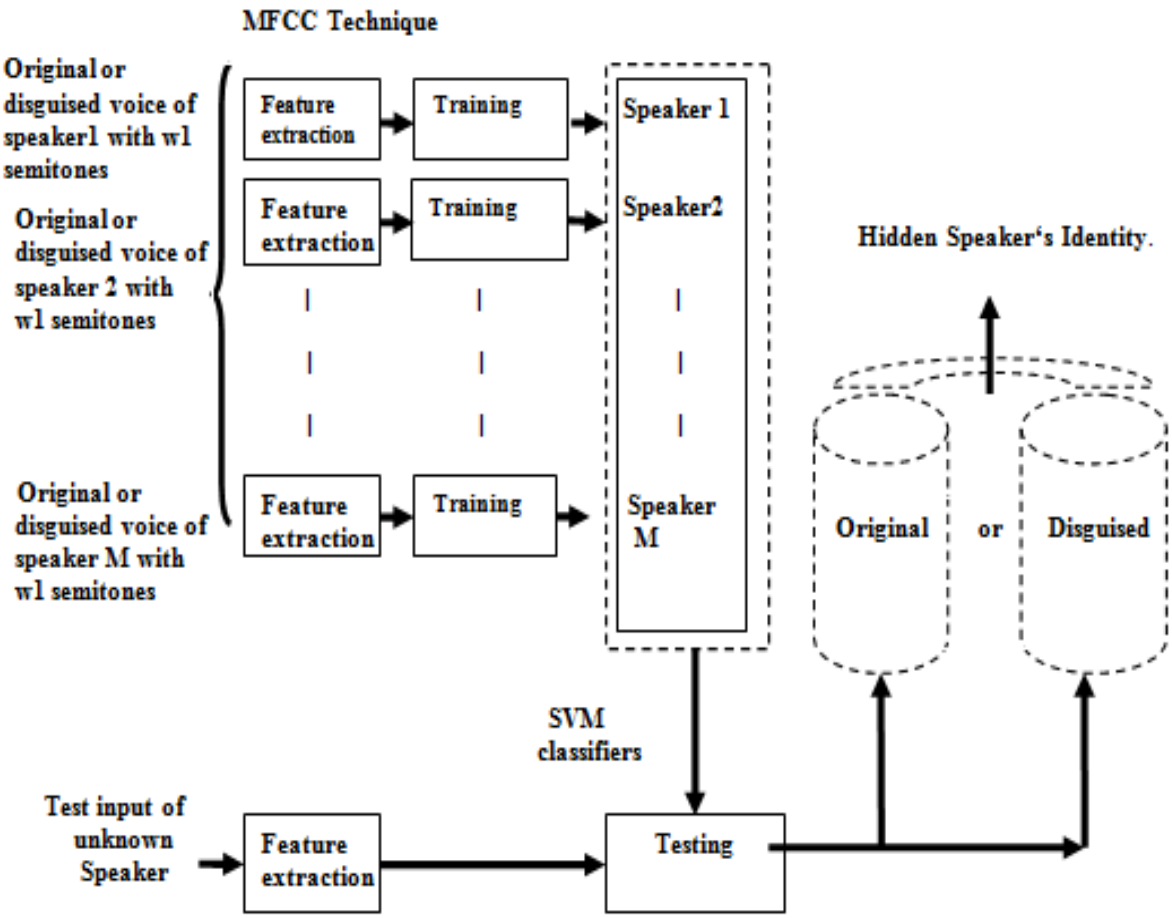


Figure 3.12 Description of proposed methodology

CHAPTER 4

RESULTS AND DISCUSSION

THIS chapter displays the whole process and results generated from the proposed recognition algorithm. As discussed in chapter 3, the first and foremost step for generating the speaker recognition model is MFCC feature extraction process that acts as a foundation for further improvement of the speaker identification scheme.

4.1 SPEECH DATABASE

Speech recordings were collected from the students of Thapar University, Patiala with the help of audacity software [56]. Database of about 40 students were used for training which consists of 20 male and 20 female students. The speech recording was text and language independent. They were allowed to speak for more than 2s. The recordings were made at 16,000 Hz sampling rate and 16 bit quantization. All these words are recorded with microphone using audacity software through which the recorded signals are save in **'.wav'** format.

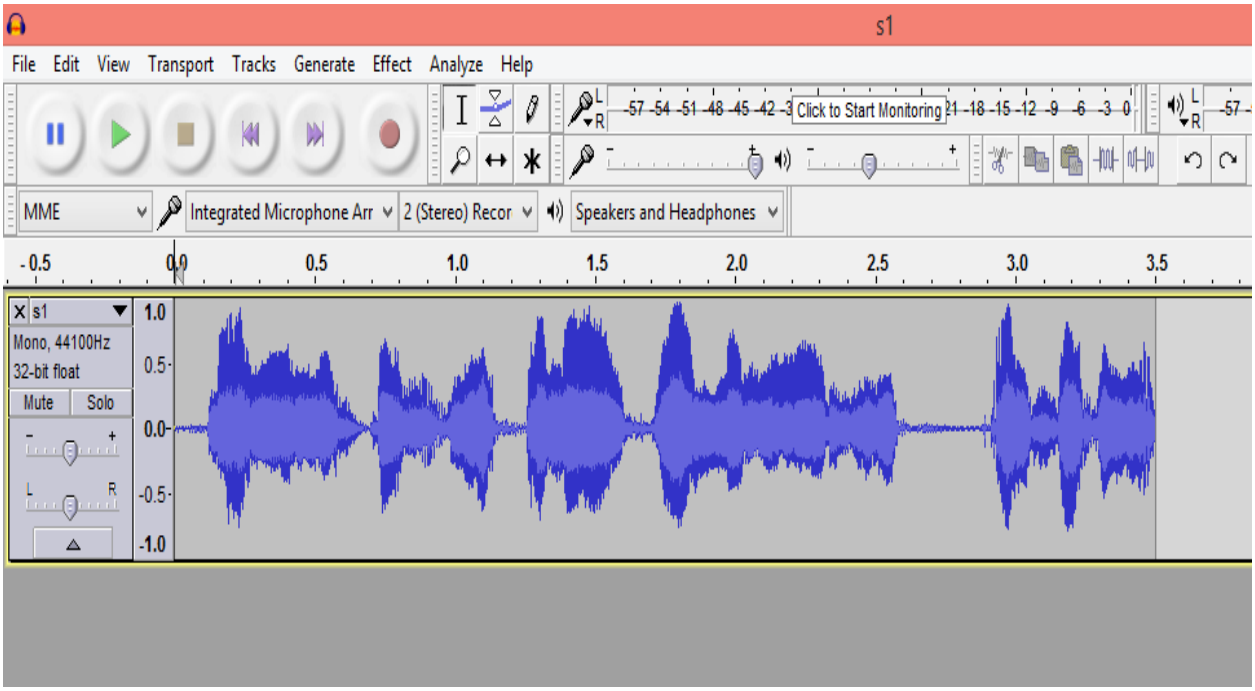


Figure 4.1. Recording of a speech sample 1

In this work, audacity tool is used for disguising a voice. This software records original voices and then disguised them by adjusting appropriate semitones of voice. In phonetics, voice pitch is always measured by 12-semitones-division, implying that pitch can be raised or lowered by 11 semitones at most. Here, a K semitone is used as a disguising factor to

represent the extent of voice disguise. For the original voice signal $x(n)$ and the disguised voice signal $x'(n)$, the changed value K of the voice pitch measured in semitones can also be used as a disguising factor [15]:

$$\alpha = \frac{2K}{12};\ K = \pm 1, \pm 2, \pm 3 \ldots \pm 11 \tag{4.1}$$

$$P'' = \frac{2K}{12} P;\ K = \pm 1, \pm 2, \pm 3 \ldots \pm 11 \tag{4.2}$$

Similarly, a positive K denotes raising the pitch P by K semitones from $x(n)$, while a negative K denotes lowering P by K semitones. In practice, algorithms for voice disguise used in audio editing software's and voice changing software's are much more sophisticated. Since algorithm optimization may be considered for computational complexity and good disguising performance the concrete approaches in various disguising tools may differ from each other.

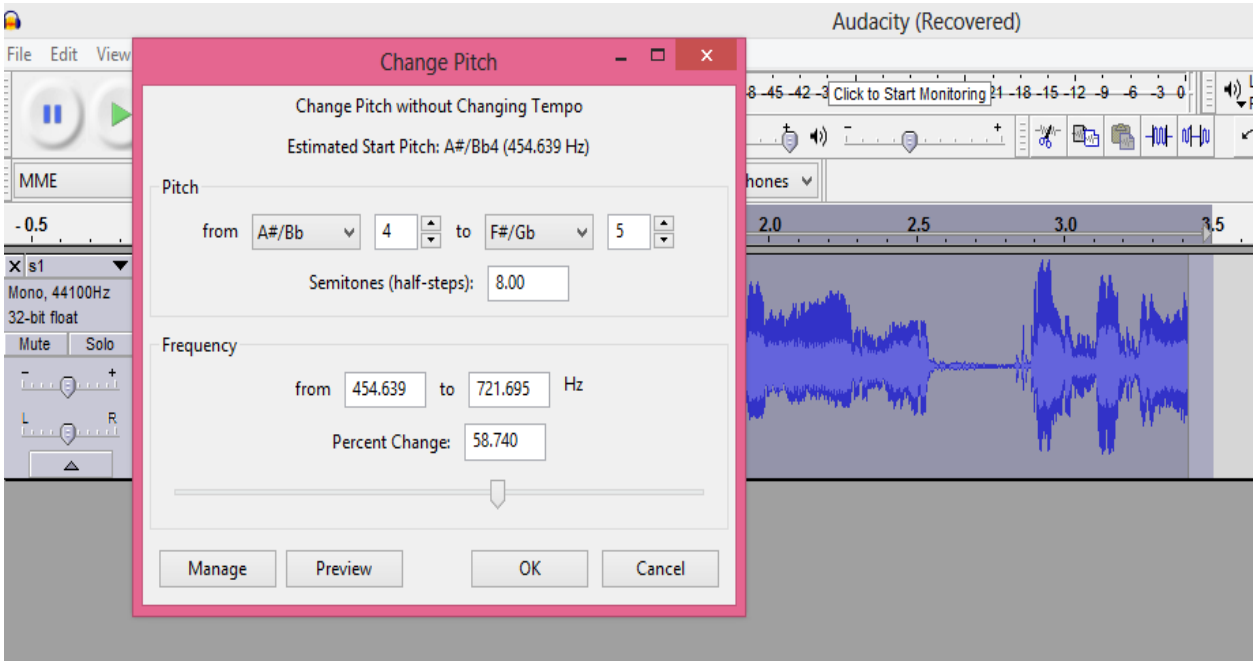


Figure 4.2 Voice disguising using Audacity

4.2 EXPERIMENTAL RESULTS

Mel Frequency Cepstral Coefficients (MFCC) account physiological activities of human ear perception which has a linear scale up to 1000 Hz and after that it follows logarithmic scale. Hence it transforms frequency into Mel domain with the help of a number of Mel filters. Then its absolute value is considered and a log function is applied, which is further used to convert back into time domain via Discrete Cosine Transform (DCT). For each speaker, 12 MFCC coefficients are taken into consideration which entirely represents

information about the vocal tract filter and it clearly separates the information about the glottal source. For visualization point of view, only few feature vectors and their MFCCs are shown.

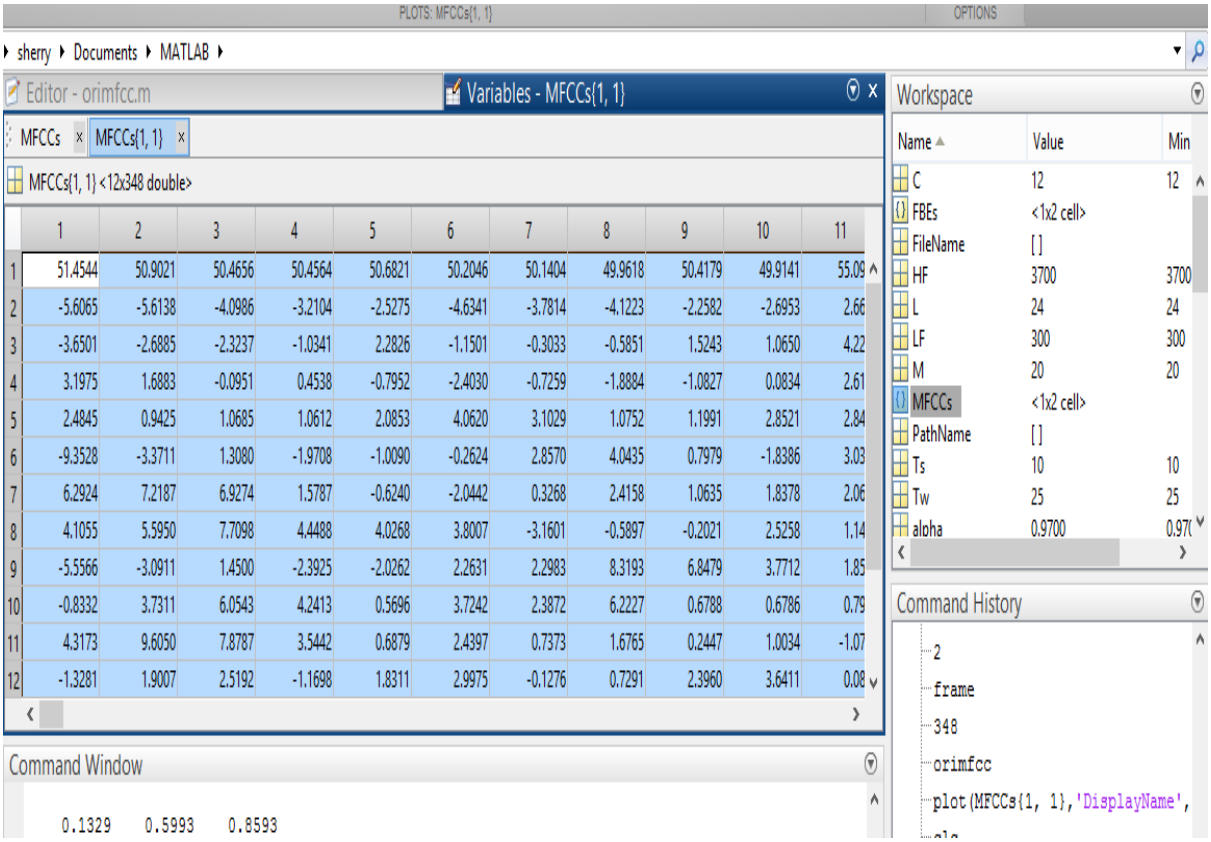


Figure 4.3 Workspace descriptions of 12 MFCCs of a speech sample.

Figure 4.3 Represents the MFCC feature vectors for each frame. Here each column represents a feature vector and the elements of each column are the subsequent MFCCs. The first 12 DCT coefficients are chosen and therefore each column will have 12 elements.

Figure 4.4, 4.5, 4.6 depicts the 2D plot of MFCC, delta and double MFCC values of both original and disguised speech of a single speaker. Delta MFCCs is computed by calculating the difference between frames of MFCC as depicted in Equation 4.3. The similar approach is used to find double delta values of MFCCs.

$$d[n] = \frac{\sum_{n=1}^{N} n\big(c(t+n) + c(t-n)\big)}{2\sum_{n=1}^{N} n^2} \tag{4.3}$$

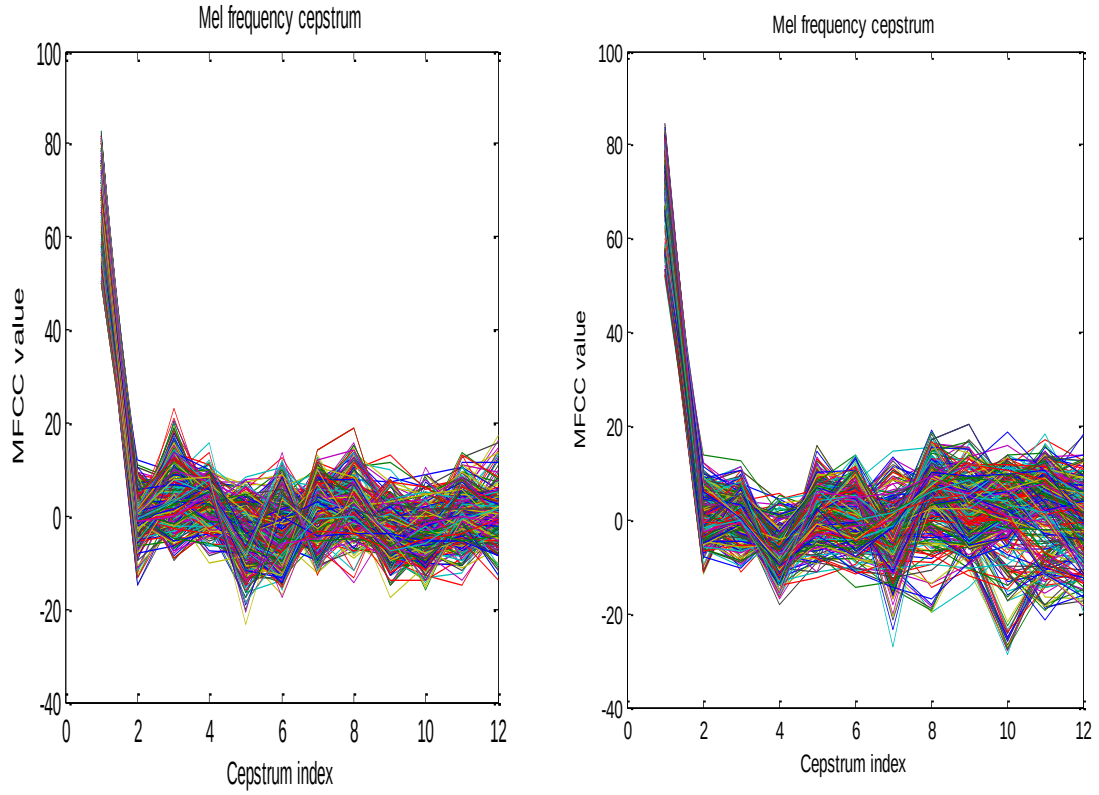


Figure 4.4 MFCC representation (a) Plot of MFCC values of original signal (b) Plot of MFCC values of disguised voices

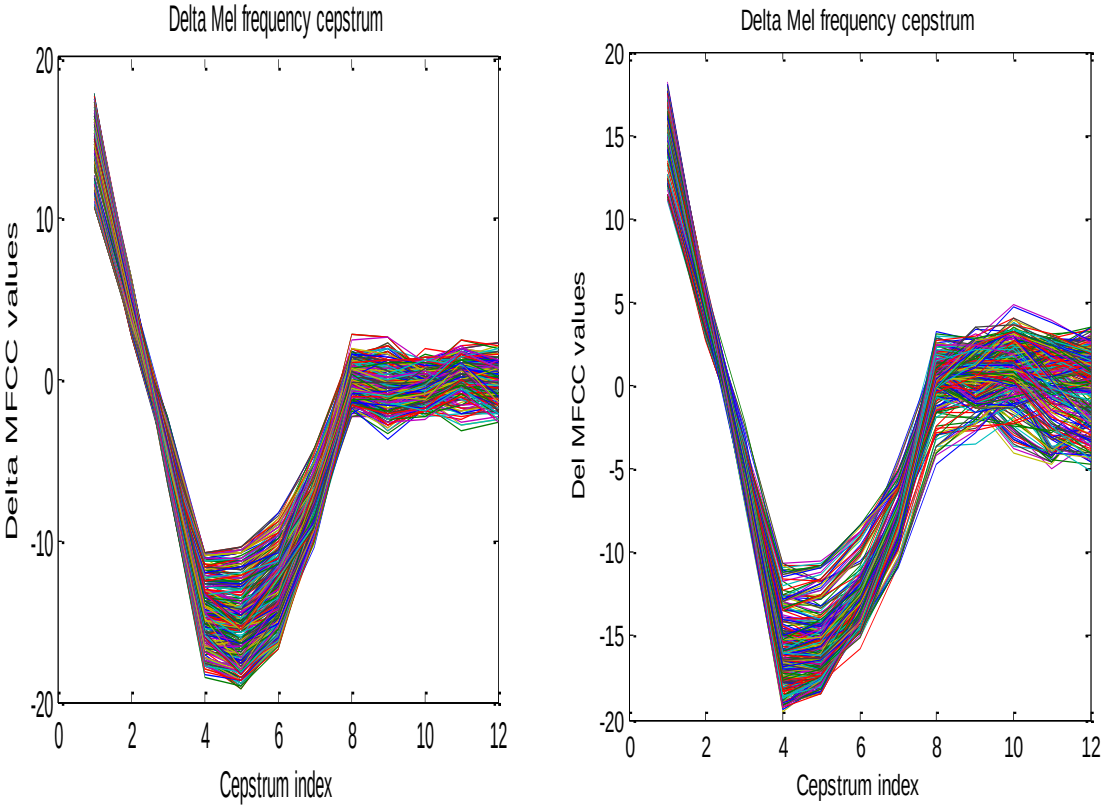


Figure 4.5 Delta MFCC representation (a) Plot of delta MFCC values of original signal (b) Plot of delta MFCC values of disguised voices

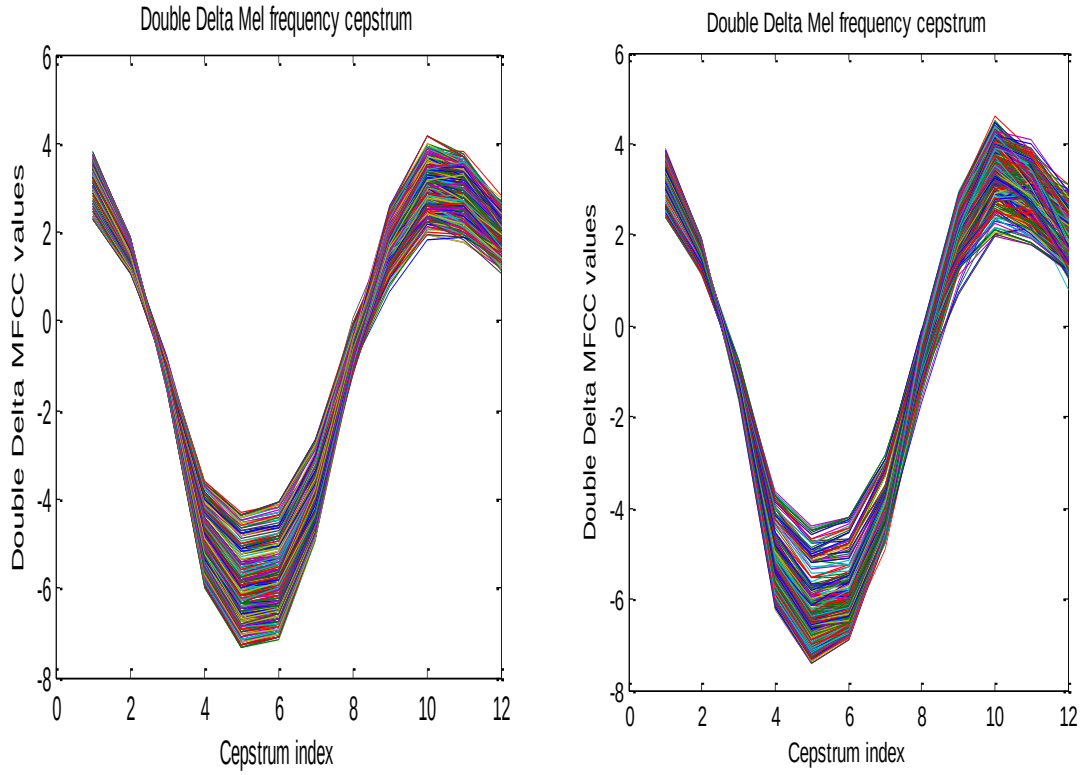


Figure 4.6 (a) Plot of double delta MFCC values of original signal (b) Plot of double delta MFCC values of disguised voices.

After implementing these steps for each speaker, acoustical feature set can be obtained for each speaker by calculating mean and correlation coefficients of MFCCs, delta and double delta MFCCs. Figure 4.7 , 4.8 represent the mean values of MFCC of 48 speakers' original and disguised voice respectively and Figure 4.9 , 4.10 represent the mean values of MFCC for 48 speakers' original as well as disguised speech respectively. Every individual voice has a unique acoustical features. It helps in identifying an unfamiliar speaker from a known set of speakers.

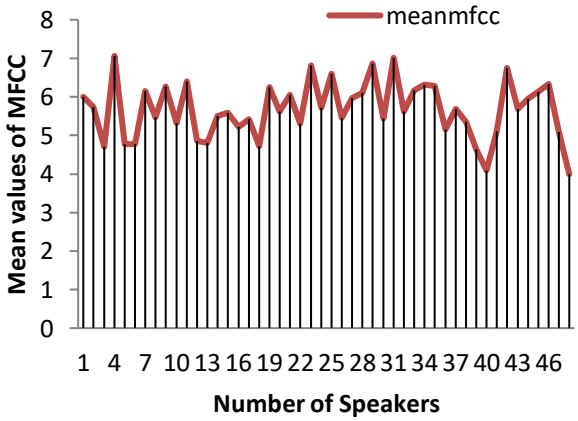


Figure 4.7 Mean values of MFCC of 48 speakers' original voice

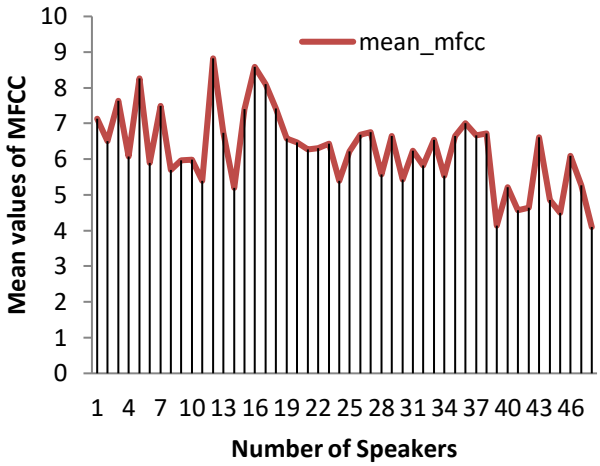


Figure 4.8 Mean values of MFCC of 48 speakers' disguise voice

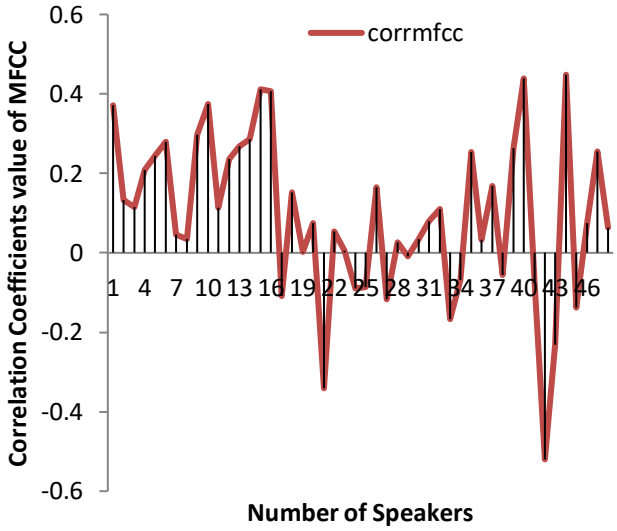


Figure 4.9 Plot of Correlation coefficients of MFCC features of original signal

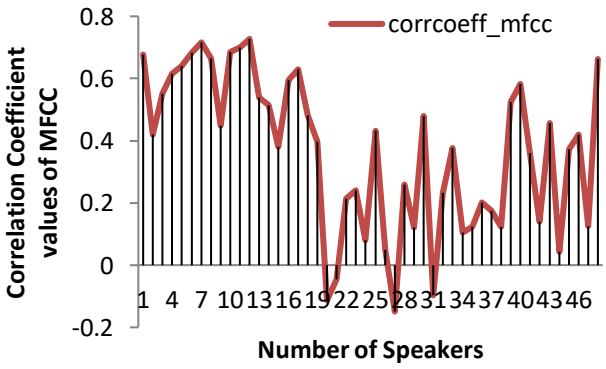


Figure 4.10 Plot of Correlation coefficients of MFCC features of disguised signal

MFCC statistical moments are extracted from the original and disguised speech so as to obtain acoustic features. The Figure 4.11 shows an acoustic feature for 10 speaker. These features helps in identifying a disguised voice from the original voice. A statistical investigation is implemented to easily analyze the effect of voice disguise on MFCC.

meanmfcc	meandelmfcc	meandoubledeltam	corrmfcc	corrdelmfcc	corrdddelmfcc	class
6.00150248	-3.275772843	-0.958273863	0.371308783	0.730563194	0.926449656	original
5.732972119	-2.945793765	-0.851412304	0.132945928	0.59929542	0.859323536	original
4.707533191	-3.035746699	-0.871799714	0.114776419	0.423237934	0.787928833	original
7.063243802	-3.036437323	-0.934589049	0.209302013	0.681101842	0.829441081	original
4.764940075	-3.084002115	-0.897822223	0.243317558	0.632921092	0.832010416	original
4.773777115	-3.150724059	-0.876016553	0.279074317	0.62004282	0.698808532	original
6.151983193	-2.890337593	-0.859864723	0.045963431	0.609164644	0.792017327	original
5.473633028	-2.925451448	-0.859354501	0.035112795	0.563219522	0.873900041	original
6.262282358	-3.270706868	-0.954843778	0.296714387	0.651317382	0.90733005	original
5.323117327	-2.943267368	-0.862293633	0.374003727	0.689320194	0.820369157	original
7.138912554	-3.160842906	-0.924053783	0.677037199	0.853081711	0.829391304	Disguise
6.507788597	-2.927457075	-0.858933722	0.420139101	0.639711589	0.839311801	Disguise
7.637168851	-2.668834414	-0.811411682	0.551889189	0.715529331	0.728971803	Disguise
6.079340016	-3.074947813	-0.892240976	0.614356287	0.734516156	0.820814168	Disguise
8.259670784	-2.561467149	-0.789165853	0.639538387	0.788772544	0.894071028	Disguise
5.895223042	-3.066949587	-0.864543011	0.683367715	0.820451534	0.901282648	Disguise
7.484377606	-2.738103864	-0.829272582	0.715464111	0.738517577	0.886475468	Disguise
5.698002	-2.82703465	-0.809275912	0.663721303	0.723809396	0.870731378	Disguise
5.974353508	-3.120567168	-0.898559804	0.447991984	0.687576318	0.825590837	Disguise
5.992983531	-2.700821476	-0.785249284	0.685190361	0.786418403	0.850007609	Disguise

Figure 4.11 Set of Acoustic features.

The standard identification system of disguised voices consists of a training stage and a testing stage. For training the speech, firstly a speech database is created in which an original voice module and a disguised voice module are presented. The above measured features of the original set of voices along with the features from disguised voice set are used as the training components to train different classifiers in turn, to classify whether a testing voice is disguised or original and identify a speaker even if his voice is disguised.

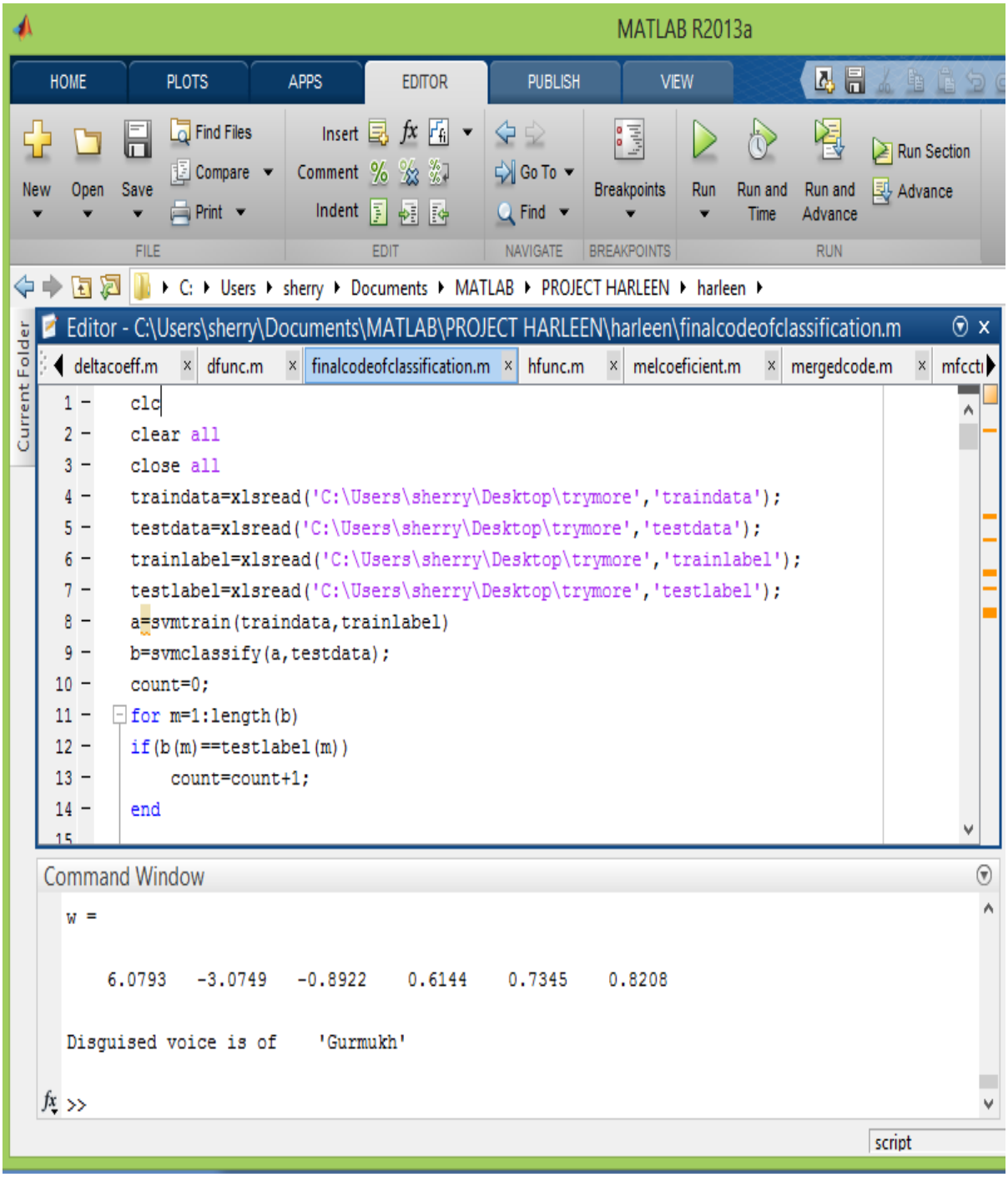


Figure 4.12 Identification of speaker whose voice is disguised.

Based on training values, classifiers predict the desired values of the test data when only the test data attributes are known. Then a feature vectors are calculated for each speech sample in the training model. After these steps, the next procedure is to extract the features or attributes from the testing module. In an identification of a particular speaker as shown in Figure 4.12, several speech attributes of different classes are compared. If these are matched to new input then speaker's identity is displayed on system's screen and if speaker is not found in the database then no match is displayed.

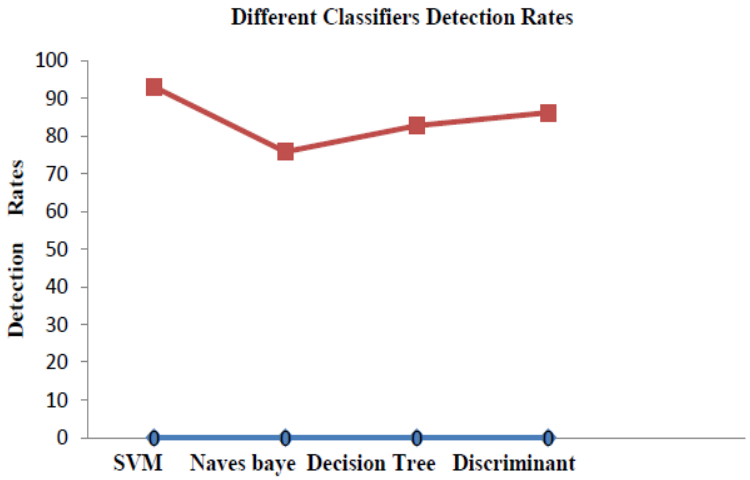


Figure 4.13 Comparison of classifiers vs. detection rate for +8 disguising factor and

Figure 4.13 shows the plot of different classifiers detection rates. SVM have significantly higher detection rates than all other classifiers. Our results on SVM may be conflicting with some other evaluation relating SVM. Here superiority of SVM is showed over other learning algorithms as it has achieved an excellent accuracy of 93.1034 percent.

CHAPTER 5

Concluding Remarks and Future Scope

Speaker Identification of disguised voice in itself is a novel application. This task is an attempt to create a system that automatically identifies a hidden speaker lying under the disguised speech signals. This will be accomplished by mining MFCC statistical moments, i.e., the mean and correlation coefficients values of MFCC features, delta MFCC vectors and double delta MFCC vectors. A statistical analysis of these features signifies that the distributions of the feature vectors of original voices are changed due to voice disguise. It can be used as acoustic features that help to separate disguised voice from original voice. The proposed method has potential to differentiate the voices disguised by a certain factor from original in an automatic way. By giving appropriate tags to original or disguised voice sets, hence leading to identification of speaker even if his/her voice is disguised. The path paved the road to a number of researches in future. It would lead to the identification of a speaker even if his/her voice is disguised.

Several researches have been ongoing from the past few decades to study the effects of voice disguise on ASV systems but they do not offer any robust or complete solutions to expose genuine identity of the speaker. Furthermore, it would result in the enhancement of the set of data to achieve better diversity for the purpose of training and testing of speakers. As the performance of the proposed algorithm is analyzed only for disguising factor of +8, to widen the system capabilities various disguising factors can be used. System may be widened by including different disguising factors and voice changing software's such as Cool Edit, Pratt, and RTISI etc. which makes suitable modification in acoustic features. The study and analysis of the acoustic features may result in a better system having higher accuracy rate. Furthermore, its use is increasing in various forensic applications to identify whether a suspected voice is disguised or not. The method can be modified by using various disguising factors. The algorithm can further be extended to provide application in confidential data security.

REFERENCES

[1] Paliwal K K. *Advances in Speech, Hearing and Language Processing*. Brisbane: Griffith University, 1990, 1-78.

[2] Flanagan J. *Speech Analysis and Perception*. New York: Springer-Verlag Berlin Heidelberg, 1972, 9-21.

[3] Rabiner LR and Schafer R W. *Introduction to Digital Speech Processing.* USA: Foundations and Trends in Signal Processing, 2007, 1–194.

[4] Benesty J, Sondhi MM and Huang Y, *Springer Handbook of speech processing*, Canada: Springer-Verlag Berlin Heidelberg, 2008, 1-1161.

[5] Reynolds DA (1995). Speaker identification and verification systems using Gaussian mixture speaker models, *Journal Speech Communication*, 17(1-2), 91-108.

[6] Furui S (1997). Recent advances in speaker recognition, *Pattern Recognition Letters*, 18(9), 859–872.

[7] Wu H, Wang Y and Huang J (2014). Identification of electronic disguised voices, *IEEE Transaction on Information Forensics and Security*, 9(3), 489–500.

[8] Wang Y *et al.* (2013). Blind detection of electronic voice transformation with natural disguise, *International Conference on Digital Forensics and Watermarking* [11th: Vancouver, Canada: 2012], 336-343.

[9] Tan T (2010). The effect of voice disguise on automatic speaker recognition, *Congress on Image and Signal Processing* [3rd: Yantai, China: 2010], 3538-3541.

[10] Reich AR and Duke JE (1979). Effects of selected vocal disguises by listening, *Journal of the Acoustical Society of America*, 66(4), 1023-1028.

[11] Perrot P and Chollet C (2008). The question of voice disguise, *Journal of the Acoustical Society of America*, 123(5), 9681-9685.

[12] Rodman RD (2003), "Speaker Recognition of disguised voices: A program for research".

[13] Kunzel HJ (1994). Current approaches to forensic speaker recognition, *ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification, Martigny*, 135-141.

[14] Kersta LG (1962). Voiceprint identification, *Journal of the Acoustical Society of America*, 47(2), 597-612.

[15] Stevens KN (1968). Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material, *Journal of the Acoustical Society of America*, 44(6), 1596-1607.

[16] Bolt RH (1969). Identification of a speaker by speech spectrogram, *Journal of the Acoustical Society of America*, 54(2), 531–534.

[17] Tosi O *et al.* (1972). Voice Identification through acoustic spectrography, *Journal of the Acoustical Society of America*, 51(6), 2030-2043.

[18] Endres W, Bambach W, and Flosser G (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation, *Journal of the Acoustical Society of America*, 49(6), 1824–1848.

[19] Hollien H, Majewski W and Doherty ET (1982). Perceptual identification of voices under normal, stress, and disguise speaking conditions, *Journal of Phonetics*, 10(2), 139-148.

[20] Hollien H and Majewski W (1977). Speaker identification by long-term spectra under normal and distorted speech conditions, *Journal of the Acoustical Society of America*, 62(4), 975-980.

[21] Ladefoged P and Ladefoged J (1980). The ability of listeners to identify voices, *University of California Working Papers in Phonetics*, 49(1), 43–51.

[22] Koenig BE (1986). Spectrographic voice identification: A forensic survey, *Journal of the Acoustical Society of America*, 79(6), 2088-2090.

[23] Kunzel HJ (1984).Current approaches to forensic speaker recognition, *ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*, Martigny, 135-141.

[24] Picone JW (1993). Signal modeling techniques in speech recognition. *IEEE Proceeding of Signal and Processing*, 81(9), 1215-1247.

[25] Guyon I and Elisseeff A (2003)**.** An Introduction to Variable and Feature Selection**,** *Journal of Machine Learning Research*, 3, 1157-1182.

[26] Atal BS and Hanauer SL (1971). Speech analysis and synthesis by linear prediction of the speech wave, *Journal of the Acoustical Society of America,* 50(2), 637-655.

[27] Oppenheim AV (1969). A speech analysis-synthesis system based on homomorphic filtering, *Journal of the Acoustical Society of America*, 45(2), 293–309.

[28] Oppenheim AV and Schafer RW (1968). Homomorphic analysis of speech, *IEEE Transaction on Audio and Electroacoustics*, 16(2), 221-226.

[29] Li Q, Soong FK, and Siohan O (2000). A high-performance auditory feature for robust speech recognition, *International Conference on Spoken Language Processing* [6th: Bell Labs, USA: 2000], 51–54.

[30] Li Q, Soong FK, and Olivier S (2001). An auditory system-based feature for robust speech recognition, *European Conference on Speech Communication and Technology*, [7th: Taipei, Taiwan: 2001], 619–622.

[31] Shaughnessy DO. *Speech Communication: Human and Machine*. New York: Institute of Electrical and Electronics Engineers, 173-227, 2000.

[32] Davis SB and Mermelstein P (1980). Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. *IEEE Transaction of Acoustical, Speech and Signal Processing*, 28(4), 357-366.

[33] Furui S (1981). Cepstral analysis technique for automatic speaker verification, *IEEE Transaction of Acoustical Speech and Signal Processing*, 29(2), 254–272.

[34] Naik J, Netsch M and Doddington G (1989). Speaker verification over a long distance telephone lines, *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* [4th: Glasgow, UK: 1989], 324-527.

[35] Rosenberg A, Lee C and Gokcen S (1991). Connected word talker verification using whole word hidden Markov models, *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* [Toronto, Canada: 1991], 381-384.

[36] Zheng Y, Yuan B (1988). Text-dependent speaker identification using circular hidden Markov models, *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* [New York, USA: 1988], 580- 582.

[37] Li K and Wrench Jr (2003). An approach to text-independent speaker recognition with short utterances, *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* [Boston, USA: 2003], 555- 558.

[38] Matsui, Furui S (2002). A text-independent speaker recognition method robust against utterance variations, *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* [Toronto, Canada: 1991], 377-380.

[39] Rosenberg A, Soong F (1987). Evaluation of a vector quantization talker recognition system in text independent and text dependent modes, *Computer Speech and Language*, 22, 143-157.

[40] Reynolds DA. A Gaussian mixture modelling approach to text-independent speaker identification. PhD Thesis, Georgia Institute of Technology, September, 1992.

[41] Oglesby J, Mason JS (2002). Speaker recognition with a neural classifier, *IEEE International Conference* [1st: London, UK: 1989], 306-309, 2002.

[42] Campbell WM *et al.* (2007). Speaker verification using support vector machines and high-level features, *IEEE Transaction on Audio, Speech, and Language Processing*, 15(7), 2085 – 2094.

[43] Clarkson P and Moreno PJ (2002). On the use of support vector machine for phonetic classification, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, [Phoenix, AZ, USA: 1999], 585-588.

[44] Reich AR and Duke JE (1979). Effects of selected vocal disguises by listening, *Journal of the Acoustical Society of America*, 66(4), 1023-1028.

[45] Zhu X, Beauregard G, and Wyse L (2007). Real-time signal estimation from modified short-time Fourier transform magnitude spectra, *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*,15(5), 1645–1653.

[46] C. Zhang (2012). Acoustic Analysis of disguised voices with raised and lowered pitch, IEEE *International Conference on Chinese Spoken Language Processing* [8th: Kowloon, China: 2012], 353-357.

[47] Jin Q *et al.* (2009). Voice converging speaker de-identification by voice transformation, I*nternational Conference on Acoustics, Speech, and Signal Processing*, [Pittsburgh, USA: 2009], 3909–3912.

[48] Laroche J (2002). Time and pitch scale modification of audio signals. Kahrs M., Brandenburg K. (eds), *Applications of Digital Signal Processing to Audio and Acoustics*. New York, NY, USA: Springer-Verlag, 2002, 279–309.

[49] Salim Roucos and Alexander M. Wilgus (1985). High quality time scale modification for speech, *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*,[Tampa, FL, USA: 1985], 493-496.

[50] MFCC Feature Extraction, Available at: http://www2.cmpe.boun.edu.tr/courses cmpe/362/spring2014/files/projects/MFCC%20Feature%20Extraction.pdf/ (Accessed on 10 December 2016)

[51] MFCC Algorithm, Available at http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/ (Accessed on 19 November 2016)

[52] W. M. Campbell *et al.* (2007). Speaker verification using support vector machines and high-level features, *IEEE Transaction on Audio, Speech, and Language Processing*, 15(7), 2085 – 2094.

[53] P. Clarkson and P. J. Moreno (1999). On the use of support vector machine for phonetic classification, *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, [Phoenix, USA : 1999], pp. 585-588

[54] Gopi ES, *Digital Speech Processing Using Matlab*. New York: Springer Heidelberg, 2014, 10-23.

[55] Brownlee J. Machine learning algorithms, Available at: http://machinelearning mastery.com/machine-learning-algorithms-mini-course/ (Accessed on 12 May 2017)

[56] Mazzoni D and Dannenberg R. Audacity Free Audio Editor and Recorder. Available at: http://audacity.sourceforge.net/ (Accessed on 11 January 2017).

Harleen_07-07-2017

ORIGINALITY REPORT

%14	%6	%10	%4
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Wu, Haojun, Yong Wang, and Jiwu Huang. "Identification of Electronic Disguised Voices", IEEE Transactions on Information Forensics and Security, 2014. Publication	%4
2	Submitted to Thapar University, Patiala Student Paper	%1
3	Wang, Yong, Haojun Wu, and Jiwu Huang. "Verification of hidden speaker behind transformation disguised voices", Digital Signal Processing, 2015. Publication	%1
4	citeseerx.ist.psu.edu Internet Source	<%1
5	Submitted to King's College Student Paper	<%1
6	www.diva-portal.org Internet Source	<%1
7	www.ling.gu.se	

	Internet Source	<%1
8	www.ijritcc.org Internet Source	<%1
9	www.scholarpedia.org Internet Source	<%1
10	es.scribd.com Internet Source	<%1
11	Submitted to University of Malaya Student Paper	<%1
12	www.dfki.de Internet Source	<%1
13	Submitted to University of Pune Student Paper	<%1
14	Submitted to King Saud University Student Paper	<%1
15	researchbank.rmit.edu.au Internet Source	<%1
16	Wu, Haojun, Yong Wang, and Jiwu Huang. "Blind detection of electronic disguised voice", 2013 IEEE International Conference on Acoustics Speech and Signal Processing, 2013. Publication	<%1
17	www.academypublisher.com	

	Internet Source	<%1
18	Zhang, Cuiling. "Acoustic analysis of disguised voices with raised and lowered pitch", 2012 8th International Symposium on Chinese Spoken Language Processing, 2012. Publication	<%1
19	www.ee.iitb.ac.in Internet Source	<%1
20	Submitted to Jawaharlal Nehru Technological University Student Paper	<%1
21	Ronald W. Schafer. "Homomorphic Systems and Cepstrum Analysis of Speech", Springer Handbook of Speech Processing, 2008 Publication	<%1
22	courses.cs.ut.ee Internet Source	<%1
23	www.csc.ncsu.edu Internet Source	<%1
24	www.ee.cuhk.edu.hk Internet Source	<%1
25	Reynolds, D.A.. "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, 199508 Publication	<%1

26	Lecture Notes in Computer Science, 2007. Publication	<%1
27	www.ece.utk.edu Internet Source	<%1
28	Aarabi, . "Single-Microphone Speech Processing", Phase-Based Speech Processing, 2005. Publication	<%1
29	aprodeus.narod.ru Internet Source	<%1
30	Submitted to iGroup Student Paper	<%1
31	Saquib, Zia; Salam, Nirmala; Nair, Rekha and Pandey, Nipun. "Voiceprint Recognition Systems for Remote Authentication-A Survey", International Journal of Hybrid Information Technology, 2011. Publication	<%1
32	Cemal Hanilci. "Principal component based classification for text-independent speaker identification", 2009 Fifth International Conference on Soft Computing Computing with Words and Perceptions in System Analysis Decision and Control, 09/2009 Publication	<%1

33	documents.mx Internet Source	<%1
34	www.ll.mit.edu Internet Source	<%1
35	Submitted to Marquette University Student Paper	<%1
36	Studies in Computational Intelligence, 2014. Publication	<%1
37	Furui, S.. "Recent advances in speaker recognition", Pattern Recognition Letters, 199709 Publication	<%1
38	Submitted to University of Nottingham Student Paper	<%1
39	"Digital Forensics and Watermarking", Springer Nature, 2017 Publication	<%1
40	www.ai.rug.nl Internet Source	<%1
41	Submitted to Higher Education Commission Pakistan Student Paper	<%1
42	www.annexpublishers.co Internet Source	<%1

43	Bahoura, Mohammed. "FPGA Implementation of Blue Whale Calls Classifier Using High-Level Programming Tool", Electronics, 2016. Publication	<%1
44	Visalakshi, R., and P. Dhanalakshmi. "Performance of speaker identification using CSM and TM", International Journal of Speech Technology, 2016. Publication	<%1
45	Lecture Notes in Computer Science, 2013. Publication	<%1
46	www.dline.info Internet Source	<%1
47	etheses.whiterose.ac.uk Internet Source	<%1
48	www.ijesi.org Internet Source	<%1
49	www.europment.org Internet Source	<%1
50	Campbell, W.M.. "Support vector machines for speaker and language recognition", Computer Speech & Language, 200604/07 Publication	<%1
51	Patrick Perrot. "Voice Disguise and Automatic Detection: Review and Perspectives", Lecture	<%1

Notes in Computer Science, 2007
Publication

52 kom.aau.dk
Internet Source

<%1

EXCLUDE QUOTES ON
EXCLUDE BIBLIOGRAPHY ON
EXCLUDE MATCHES < 10 WORDS