Improving Performance of Speaker Recognition Using Prosodic Features

THESIS

Submitted to
Babasaheb Bhimrao Ambedkar University
(A Central University)
Lucknow

BABASAHEB
BHIMRAO
AMBEDKAR
UNIVERSITY
LUCKNOW
प्रज्ञा शील करुणा
ESTABLISHED 1996

For the Degree of

Doctor of Philosophy

In

INFORMATION TECHNOLOGY

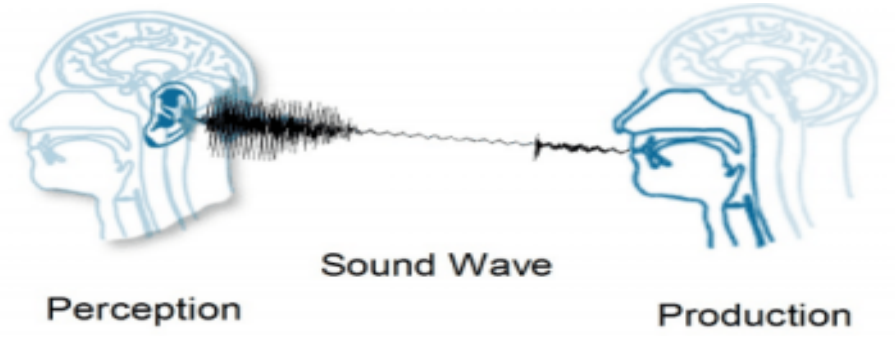
By:

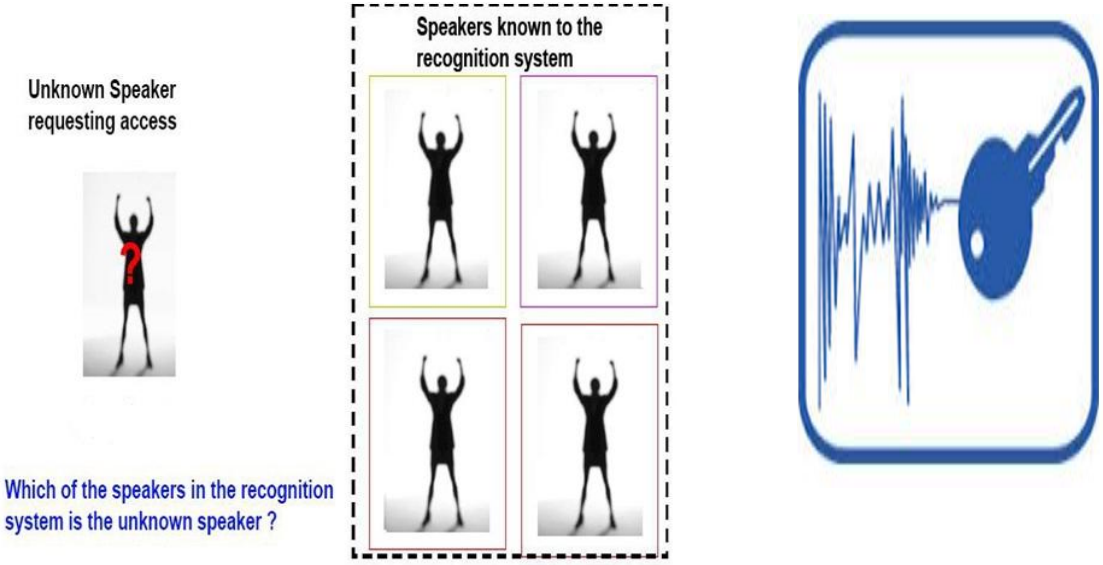
Nilu Singh

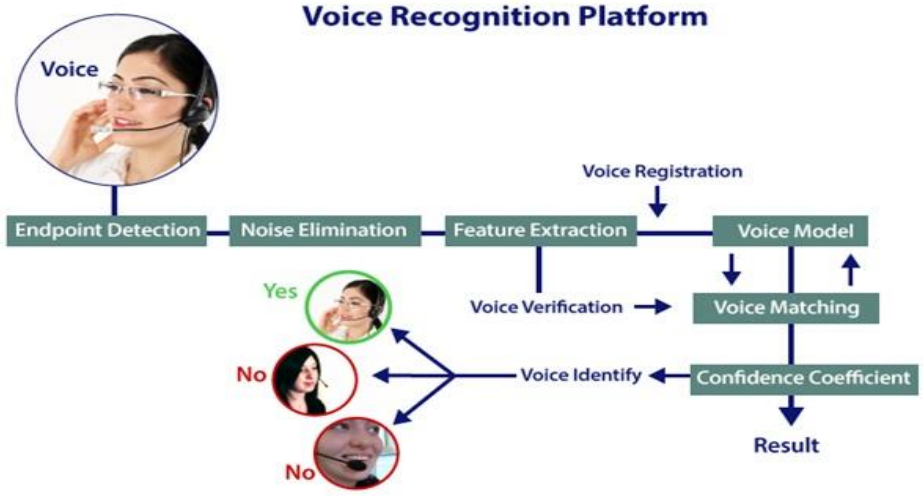
(Enrollment No. 1293/15)

DEPARTMENT OF INFORMATION TECHNOLOGY
BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
(A CENTRAL UNIVERSITY)
VIDYA VIHAR, RAEBARELI ROAD, LUCKNOW-226 025 (U.P.), INDIA

2018


Sound Wave
Perception
Production


Unknown Speaker
requesting access
Speakers known to the
recognition system
Which of the speakers in the recognition
system is the unknown speaker ?


Voice Recognition Platform
Voice
Voice Registration
Endpoint Detection
Noise Elimination
Feature Extraction
Voice Model
Yes
No
No
Voice Verification
Voice Matching
Voice Identify
Confidence Coefficient
Result

MY VOICE IS MY PASSWORD

DECLARATION

I, Nilu Singh, solemnly declare that this thesis of research on **'Improving Performance of Speaker Recognition Using Prosodic Features'** is my original work. The study has been conducted under the guidance of Prof. R. A. Khan, at Department of Information Technology, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow. It is further declared that to the best of my knowledge and belief it has not been submitted earlier for the award of any degree and also undertaken that the thesis is essentially free from all kinds of plagiarism.

Dated: ……

(Nilu Singh)

Research Scholar
Department of Information Technology,
Babasaheb Bhimrao Ambedkar University,
(A Central University)
Lucknow-226025, India

CERTIFICATE

This is to certify that the thesis entitled **"Improving Performance of Speaker Recognition Using Prosodic Features"** submitted by **Ms. Nilu Singh** is an original research work and has not been previously submitted in part or full for the award of any other degree or diploma to this or any other University.

This thesis submitted to Babasaheb Bhimrao Ambedkar University Lucknow satisfies all the requirements as stipulated in the *Doctor of Philosophy (Ph.D.)* regulations-1999 as amended in 2013 and it is fit for submission and evaluation for the award of the degree of Doctor of Philosophy of the University.

Co-Supervisor **Supervisor**

Dated: ……. **Head of the Department**

ACKNOWLEDGEMENTS

Too Many hands make light work.

—John Heywood

As a researcher, I am honoured to conduct my doctoral research under the supervision of Prof. R. A. Khan and Dr. Alka who have been a great mentor during all steps of the research. For their non-stopping support and interest in my work, I extend my most honest gratitude. I would like to thank to both for everything from choosing research topic to writing research papers and preparing presentations. Their guidance and support throughout my research work are invaluable.

I am very thankful to Dr. Shalini Chandra for her endless help and guidance at each and every step during the research. I also express my heartfelt thanks to Dr. Pawan K. Chaurasia for providing moral support, encouragement, and consultations during the course of the study. I am also thankful to all the faculty members, research scholars, and lab / office staff of the department for their time-to-time encouragements and continuous support.

Last but not the least; I am heartily grateful to my parents for always believing in me and encouraging me to follow my dreams. I am thankful to my loving son Shaurya for his love and support. I have neglected some of the duties towards them and appreciate his tolerance even in the worst condition. I could not have achieved any of this without the support and encouragement that they have always given me.

As a Research Scholar, I am privileged enough to complete my studies in such a nourishing environment; made available by the university resources.

Nilu Singh

ABSTRACT

Automatic Speaker Recognition is a process of recognizing a speaker using his/her voice. Speaker / voice recognition is a biometric sensory system which uses human voice for recognition process. It is different from speech recognition. In speech recognition words are recognized as they are articulated. In other words, speech recognition identifies what you are saying and speaker recognition identifies who is speaking. There are many studies about the speaker recognition technology in early 1960s; it was Lawrence Kersta of bell lab who developed the first speaker recognition system which was based on spectrographic voice verification. The shape and size of vocal tract varies from one speaker to another; showing the differences in resonance frequencies. Now days most of the speaker recognition systems are based on spectral information i.e. these system use spectral information extracted from the speech signal segments of size 10-30ms. But cepstral based system performance degrades if the received speech signal have any noise. Though many technologies have been developed but still several research issues remain, which are still as a challenge.

Till date a lot of work has been done in automatic speaker recognition field, but still many realistic problems need to be solved. In the research, the researcher has focused on Speaker Identification System (SIS) and has tried to examine areas of possible improvements in the field of speaker recognition. In addition, research work has formulated general view of the techniques used in practice to design a speaker recognition system. As this is evident from the literature that the prosodic features are more robust to noise, the primary goal of the research is to improve the performance of speaker recognition system by modeling prosodic features. In addition, the researcher has focused on text-independent speaker identification system using Gaussian Mixture Model (GMM) in the presence of environmental noise.

In order to achieve the goal, the researcher has proposed a speaker recognition framework for the development of speaker recognition system. Proposed framework has mainly six phases named as speech acquisition phase, features extraction phase, speaker modeling phase, pattern matching phase, decision phase and performance evaluation phase. The proposed framework is implemented by using prosodic features. The major

reason behind using prosodic features for speaker identification is that these features improve system performance and consistency. The prosodic features are robust against noise and channel effect. Training and testing databases have been created using enrolled speaker's voice. Experiments are performed on the created voice databases of male and female utterances.

To keep in mind session variability in a particular duration, same speaker's voice is acquired again. 57 speakers were enrolled (both male and female) for carrying out experiment. The range of speaker's age lies between 22 to 45 years. Each speaker was given to read different content for 2 to 3 minutes. Speakers were allowed to speak in their own reading style. Reading material included articles from newspaper and books. Every speaker was allowed to read the randomly selected articles in Hindi & English both. The voice has been recorded in a normal lab environment with a sampling frequency of 8000 using a mono channel. At the time of voice recording in the lab, electric equipment were switched on.

Each and every phase of the proposed framework has been implemented and tested using MATLAB as well as Praat software. Speaker models are created by using Gaussian mixture modeling technique, and stored for training and testing purpose. After creation of speaker models matching is performed and on the basis of match score decision is made that either speaker is accepted or rejected. In addition, system performance is evaluated on the basis of equal error rate metric.

Recognition results are evaluated by MATLAB as well as by Praat software. For testing the proposed system, 1561 utterances of 57 speakers are used. As the system performance improves by maintaining session variability, hence for maintain the same the voice of the speakers has been recorded twice throughout the experiment. In the research, prosodic (49 dimensional) features has been extracted from the speakers database. During experiment extracted pitch, duration and energy related features are broken into segments of 30-45 seconds. 128 -512 Gaussians components have been created for these training segments for each speaker. By using Legendre polynomial expansions we have estimated pitch and energy contours for each segment. One feature vector for each segment has been calculated then it has been modeled by using GMM.

The proposed approach is compared with another recognition system using MFCC and it is found that the recognition rate of the proposed speech recognition system is noticeably higher. Since, no approach can be accepted unless it passes the validation test. Hence, in order to prove the acceptability of the proposed system, it has been also validated. For validation of the proposed approach an experimental tryout has been carried out. In the research Student's t- Test is used for validating the system. For the purpose, null as well as alternate hypothesis have been formulated. Since the rejection or acceptance of a null hypothesis is based upon either (0.05) alpha (α) or (0.01) alpha (α) level of significance for one tailed or two tailed test; (0.05) alpha (α) level of significance for a two tailed test is taken for rejection of the null hypothesis. The results of statistical analysis show that there is significance difference between the previous recognition system (MFCC) and the proposed approach. Now, t- value is calculated to determine for rejecting the null hypothesis and accepting the alternate hypothesis. The t value comes out to be -7.2218 from Student t- Test. As the value exceeds the t critical value of 0.0019 for a two tail test at the 0.01 level for 4 degree of freedom, the null hypothesis H_{01} is strongly rejected and the alternate hypothesis H_{11} is accepted. Hence it is validated that performance of speaker recognition system can be improved by using Prosodic features.

ABBREVIATIONS

ASR	:	Automatic Speaker Recognition
SR	:	Speaker Recognition
SV	:	Speaker Verification
SIS	:	Speaker Identification System
SI	:	Speaker Identification
F_0	:	Fundamental frequency of a speech signal
DCT	:	Discrete Cosine Transform
DDTW	:	Derivative Dynamic Time Warping
DFT	:	Discrete Fourier Transform
DTW	:	Dynamic Time Warping
EER	:	Equal Error Rate
FAR	:	False Acceptance Rate
FFT	:	Fast Fourier Transform
FIR	:	Finite Impulse Response
FRR	:	False Rejection Rate
GDW	:	Gaussian Dynamic Warping
GMM	:	Gaussian Mixture Model
HMM	:	Hidden Markov Model
LPCC	:	Linear Prediction Cepstral Coefficient
MFCC	:	Mel Frequency Cepstrum Coefficient
SVM	:	Support Vector Machine
TSM	:	Time Scale Modification
VQ	:	Vector Quantization
KF	:	Kalman Filter
kNN	:	k-Nearest Neighbour
LDA	:	Linear Discriminant Analysis
LDC	:	Linear Discriminant Classifier
LPC	:	Linear Predictive Coding
PLP	:	Perceptual Linear Predictive

RMS	:	Root Mean Square
SD	:	Standard Deviation
SVM	:	Support Vector Machine
V/UV	:	Voiced/Unvoiced
ZCR	:	Zero Crossing Rate
NIST	:	National Institute of Standards and Technology
ML	:	Maximum Likelihood
EM	:	Expectation Maximization
MAP	:	Maximum A Posteriori
UBM	:	Universal Background Model
SRR	:	Speaker Recognition Rate

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

Chapter-1
Introduction

CHAPTER-1: INTRODUCTION

Men think they can copy Nature as Correctly as I copy Imagination; this they will find Impossible, & all the Copies or Pretended Copies of Nature, from Rembrandt to Reynolds, Prove that Nature becomes to its Victim nothing but Blots and Blurs... Copiers of Nature are Incorrect, while Copiers of Imagination are Correct.

- William Blake

1.1 Introduction

Speech is a natural way to convey information by humans. Speech signal is enriched with information of the individual. Recognizing a person's individuality by his/her voice is known as Automatic Speaker Recognition (ASR). Speaker recognition falls in the category of biometric security systems. Biometric is related to human characteristics or individuality. Biometric verification or realistic authentication is used to recognize an individual through his/her voice's individual characteristic. Voice biometric includes behavioral or physiological measurements of individual. Behavioral biometric is performed by Voice, Signature, Keystrokes, and Typing etc. whereas physiological biometric includes iris, face, retina, fingerprints, ear, DNA etc. Now a days voice biometric is an emerging research area [1-2].

Human speech is a medium for expressing their thoughts during communication. Spoken language is the most natural way for human to transfer information. A speech signal is a complex signal which is packed with several knowledge resources such as acoustic, articulatory, semantics, linguistic and many more [3-4]. During communication, human easily understand information such as emotion, language, and mental status etc. This ability of human to decode information motivated many researchers to understand speech signal production and perception. This idea helps to developing a system which automatically extract and process the built in information in a speech signal. A person's voice is different from another due to the acoustic properties of speech signal. Speaker's voice is unique to an individual due to differences which occur as anatomical differences inherent in the vocal tract and the cultured speaking behaviors of different individuals [3-5].

Speaker recognition is a process of recognizing who is speaking on the basis of information included his/her speech signal/waves. In this digital era speaker recognition is the most useful biometric recognition technique [5]. Now days many organizations like bank, industries, access control systems etc. are using this technology for providing greater security to their vast databases [5-6]. Speaker recognition is broadly classified into speaker identification (1: N matching) and speaker verification (1:1 matching). Identification is considered as more difficult than verification [7]. This is intuitive that performance of speaker identification system affected by the number of registered speakers increases (the probability of incorrect decision increases). While the performance of speaker verification system is not affected by increase in voice database size since only two speakers are compared.

In last few years, requirement for authentication has been increased with the increasing digital world of information. It has already been proved that a biometrics authentication technique increases security levels. Speaker identification is the process of identifying an utterance from the known set of speakers while speaker verification is the process of accepting or rejecting the claimed identity. Speaker verification systems are the real example of biometric authentication systems. Further, it can be classified as text-dependent and text-independent [8]. The text-dependent systems are based on same utterance spoken by speaker in both cases i.e. training and testing while in text-independent systems it is not required to utter the same sentence/words during training and testing [7]. It is accepted that text-dependent systems provide more accurate results as both the content and voice can be compared that is speaker utters exact the sentence which he/she uttered during training. While text-independent recognition systems, may use either the same utterance or different for every verification/identification session.

Gaussian mixture model (GMM) is one of the popular approaches used to speaker modeling for speaker identification. GMM is used as two distinct ways for identification system; firstly, when training database principle is the maximum likelihood (ML) and parameter estimation is performed by using expectation maximization (EM) algorithm; and secondly when the training database principle is maximum a posteriori (MAP). In this case, GMM as a universal background model

(UBM), is created for training database of speakers and these models are trained by the UBM (using registered speakers specific data) [9-11].

Speaker recognition basically has two categories; speaker identification and speaker verification. Speaker verification is used for those applications where speech is used as the key to authorize the identity claim of a speaker. Speaker identification is used to decide that a given utterance comes from a certain registered speaker. Speaker verification has larger usability than speaker identification. The basic purpose of speaker identification is crime investigation. It is used to decide which of the suspected speaker's voice match with the registered speakers. With the increase in voice database, difficulty of speaker identification increases. Speaker verification is independent of voice database population since it works only on binary decision that is acceptance or rejection. Speaker recognition system performance (recognition accuracy) is most affected by intersession variability (variability over time) and spectra of a speakers speech signal [1] [12].

1.2 Speaker Recognition

It is well accepted that in this electronic era people interact using voice with the help of electronic devices. Human voice is a signal which contains several information related to human characteristics, such as emotion, words, language, speaker identity etc.. To identify a human being by their voice, required speech features are selected from speech signal with available feature extraction techniques [13]. It is the process of recognizing a person on the basis of his/her voice. Automatic speaker recognition system is categorized as speaker identification and speaker verification. Speaker verification system decision is binary i.e. 0 or 1 (accept or reject) as this justifies an identity claimed by the speaker. Speaker identification decision requires N matching and then the decision is made about acceptance or rejection [14]. Further it is distinguished as text-dependent and text-independent speaker recognition. In the text-dependent system, the recognition of speaker's identity is based on the specific words or phrases. In the case of text-independent recognition, speaker's have no restriction to speak sentence or phrases [5] [15-19]. Figure-1.2(a) presents an overview of speaker recognition system and figure-1.2(b) shows the basic concept of speaker identification and speaker verification methodology.

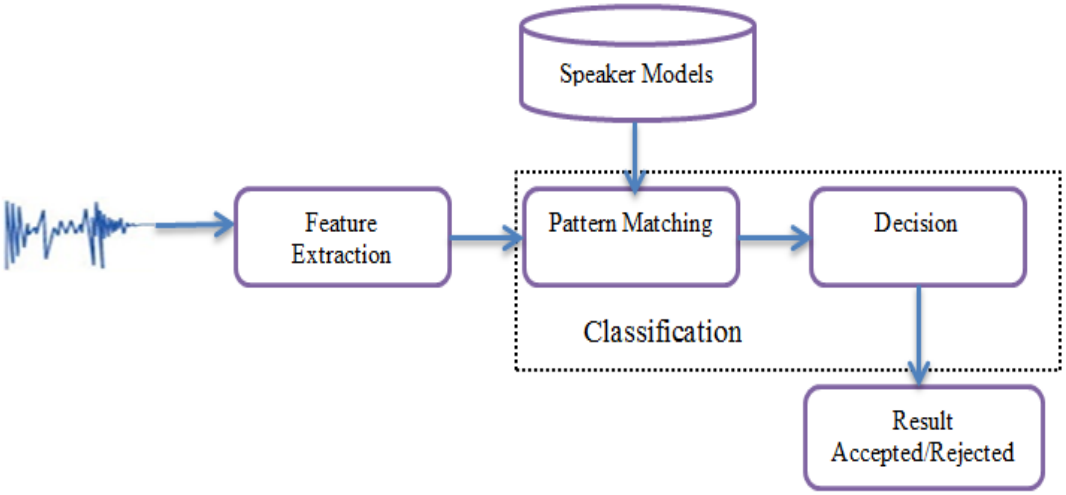


Figure-1.2(a): Basic Speaker Recognition System

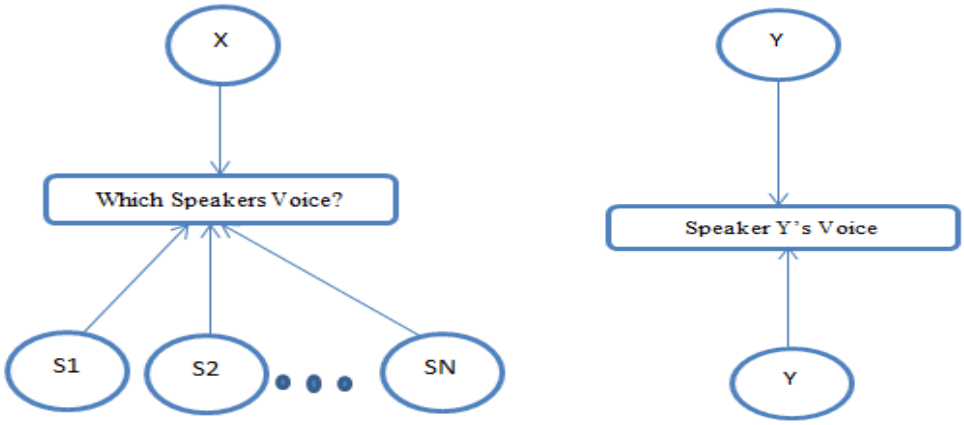


Figure-1.2 (b): Speaker Identification and Speaker Verification

The popularity of speaker recognition system is due to its low cost of implementation. This is because of the easily availability of microphones and the universal telephone network. As in this digital era it is very easy to capture someone's voice and authenticate it by using speaker recognition system. The only cost is due to the software which is used for speaker recognition system. The problem with speaker recognition system is the analysis of speech signal because speaker recognition is embedded in the study of the speech signal. The study of speech signal is about its characteristics which distinguish one speech signal with another [5] [15] [17].

1.2.1 Speaker Identification

Speaker identification system is 1: n matching system. In this, user need not to provide his/her identity to the system. During identification user has to input his/her speech to the ASR system and the system now decides the identity of user on the basis of the match score. In this case system has to perform N comparison (N is the number of stored speaker/user model of voice database). During identification, comparison with each registered model will produce a likelihood score, on the basis of this score higher likely model is identified for the speaker [1] [20].

From the study it is clear that speaker identification is complex than speaker verification. Hence in case of speaker identification, system performance degrades as compared to speaker verification [21]. Figure-1.2.1 shows the process of speaker identification system.

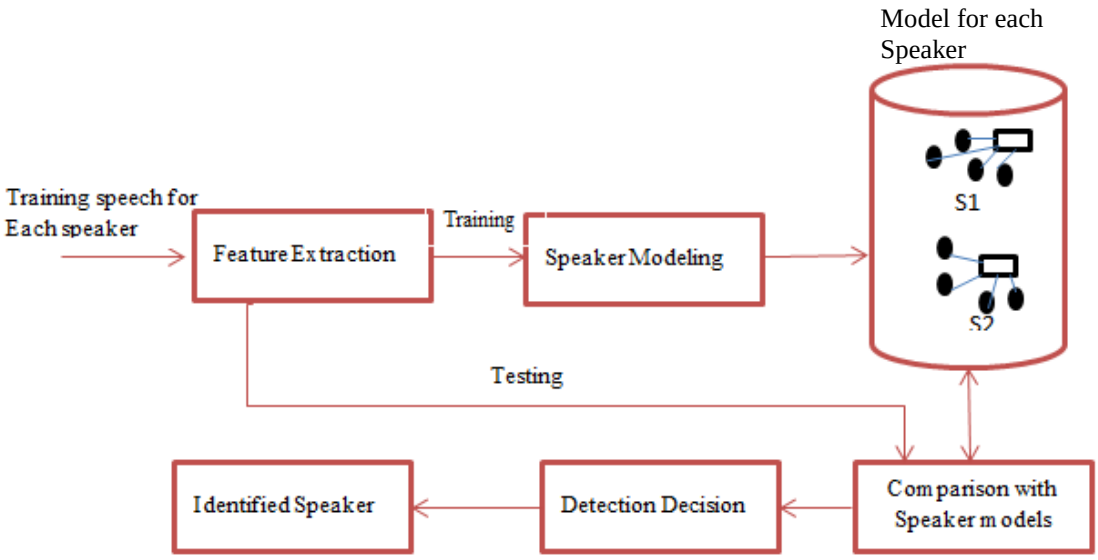


Figure-1.2.1: Process of Speaker Identification system

1.2.2 Speaker Verification

Speaker verification system is 1:1 that is the system either accepts or rejects. In verification process, firstly user need to provide his/her identity and then it is checked by the system and decisions are made accordingly that whether the claimed identity is true (accept) or false (reject) [22]. Speaker verification can be explained easily with the help of an example of Automated Teller Machine (ATM). Before any transaction,

users are first needed to insert the ATM card in the machine. This credit/debit card contains the information about user such as name, signature etc. Now, if the ATM is working on ASR technology then it will check that the card is used by its genuine holder by asking to produce his/her voice. Since the user has already provided his/her identity to the system, only 'yes or no' decision has to be made by the ATM machine i.e. either the card holder is accepted or rejected. This decision is made on the basis of comparing the voice input to the previous voice input provided by the user [4] [7]. Figure-1.2.2 shows the process of speaker verification system.

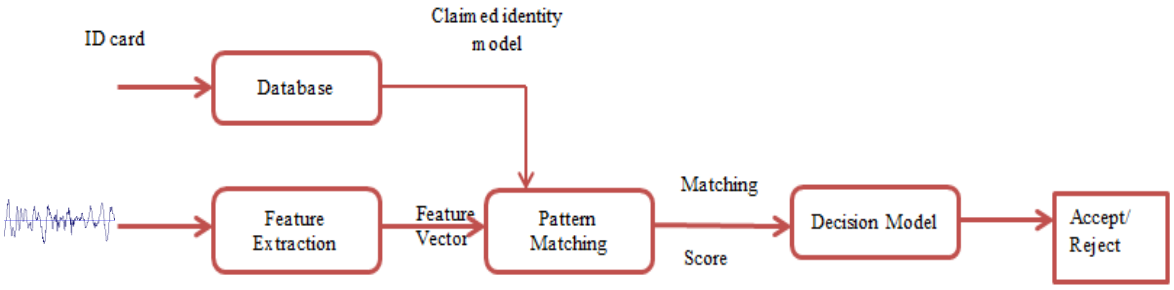


Figure-1.2.2: Process of Speaker Verification System

1.2.3 Open-Set vs. Closed-Set

Speaker identification can be divided into closed-set and open-set speaker identification. Open-set speaker identification for a test utterance or a set of enrolled speakers is a twofold problem. Firstly, it is necessary to find speaker model which have best matches in the given set; Secondly, it must be determined that the best match test utterance is actually produced by the best matched model speaker or some unknown speaker [4] [21] [23] [28]. Figure-1.2.3 shows the classification of speaker recognition system.

The possible error and problems in open set speaker identification can be examined as follows. Suppose that N speakers are enrolled in the system for voice database and M_1, M_2……Mi, are their statistical models [60]. If *O* represents the feature vector sequence extracted from the given utterance, then the open-set identification is given as follows:

$$\text{Max}_{1\le i\le M}\{p(O/Mi)\} > \theta \longrightarrow O \in \begin{cases} Mk,\ k = \arg\max_{1\le i\le M}\{p(O/Mi)\} \\ \text{Unknown speaker model} \end{cases}$$

$$\text{Max}_{1\le i\le M}\{p(O/Mi)\} < \theta \longrightarrow O \notin \begin{cases} Mk,\ k = \arg\max_{1\le i\le M}\{p(O/Mi)\} \\ \text{Unknown speaker model} \end{cases}$$

Where θ is a pre-defined threshold and *O* is assigned to speaker model that yields the maximum likelihood over all other speaker models in the system. If the maximum likelihood score is greater than the threshold θ, it is stated that the voice is originated from a known speaker. For a given *O* three types of error are possible [4] [23-26].

- **False Acceptance (FA):** The system accepts an impostor as one of the registered speaker [4].
- **False Rejection (FR):** The system rejects a true speaker [4].
- **Speaker Confusion (SC):** The system correctly accepts a true speaker but confuses him/her with another enrolled speaker [4].

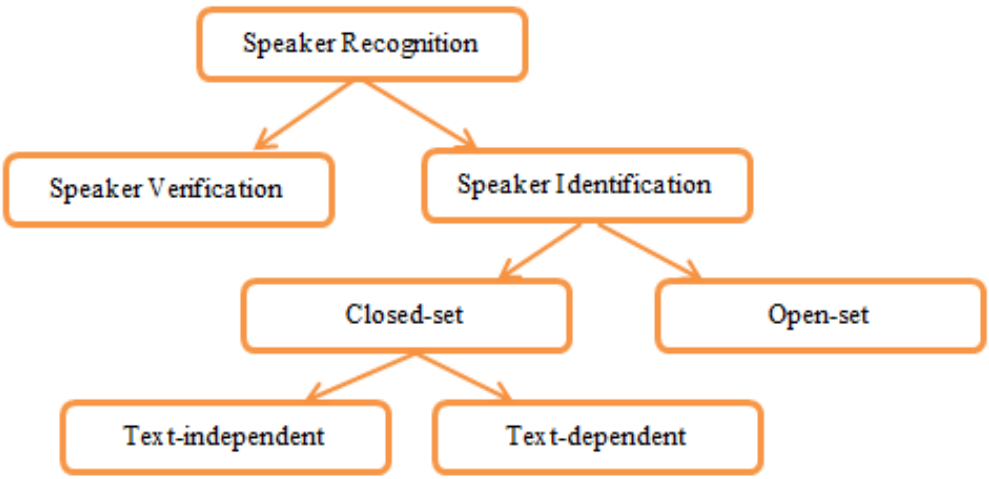


Figure- 1.2.3: Classification of Speaker Recognition

1.3 Current Approaches in Speaker Recognition

Research in speech and speaker recognition by machine has been conducted for more than five decades. Speech and speaker recognition is different in such a manner that

speech recognition is the process of recognizing the linguistic content in spoken utterance while speaker recognition is the process of determining who is speaking based on features of his/her voice [29]. Currently, various commercial text-independent speaker recognition systems exist. These commercial systems are used in many areas due to their low enough error rates. In addition, many researches are making an effort to minimize the equal error rate [30].

Speaker recognition comes from its larger field of pattern recognition. In the last several years, pattern recognition techniques make lots of contribution to speaker recognition research. Recently, many high performance speaker recognition systems build by using joint factor analysis [30] [31-32]. Continuous research is going on to improve error rates. Among all, a major application which needs more improvement and accuracy is speaker identification in forensic applications. Use of speaker recognition in forensics has opened many new fascinating research areas. For example, which voice features are mutual between speakers, which voice features vary by language, health, emotion, age, dialect and some other factors [33-35].

A major area of research in speaker recognition is reducing the impact of noise such as background noise, environmental noise, handset variability etc. This is a big challenge where unknown conditions as security applications are accessed over internet. Therefore noise (environmental conditions) still continues to be a big research area [36-37]. In NIST 2010, SRE, it is found that a system having equal error rates below 2% is best system. For many applications, 2% EER is acceptable and the others require stringent requirements. Recently many researches have been made to minimizing EER [38].

1.4 Extraction of Speech Features

A speech signal has several features such as phonetic, prosodic and acoustic etc. Selection of the required speech features among several features is the important task. By selecting more informative speech features will help to improve system performance. Selection of less useful or not useful features for a particular task is unfavorable to the system performance. Hence, examining the new speech features for a specific task has been always an important and difficult task [48]. The process of extracting speech features from speech signal is not straightforward. Speech signals

do not only carry linguistic messages but it also includes major paralinguistic components, prosody. Speech features stated as prosodic features define prosody of a speech while the features which are not used to describe prosody are called acoustic features of speech. Fundamental frequency is the main component of a speech signal which is further explained as:

- **Fundamental Frequency features:** Fundamental frequency (F_0) features are useful for tonal languages. Tones are related with dynamics of the fundamental frequency. To extract F_0 there are many methods used. One common method is through autocorrelation i.e. autocorrelation of the signal within a frame. In this computation, second highest peak of autocorrelation is represented as fundamental frequency of speech signal. For better accuracy or to make system more robust against noise another technique is required. It is based on observation such as tracking the peak of the autocorrelation across frames or normalizing the autocorrelation according to the analysis window [15] [49-50] [44].

  Measurement unit of F_0 is Hertz (Hz) i.e. number of periods within one second. Fundamental frequency period can be further divided as jitter and shimmer.

  - **Jitter:** It is frequency stability in terms of equality of period's duration. It is computed as average absolute difference between consecutive periods (divided by the average period) [40] [46].
  - **Shimmer:** It is the measurement for amplitude stability of F_0 periods. It is computed as Average absolute difference between the amplitudes of sequential periods (divided by the average amplitude) [40] [46].

Both jitter and shimmer have their observation based thresholds and basically used in speech pathology research. From the fundamental frequency and glottal cycle point of view, there are many phonation types voice such as the following [15]:

- Normal Voice: Vocal cords are in their natural mode [15] [16].
- Creaky Voice, Vocal Fry: Creaky phonation is characteristically related with aperiodic glottal pulses. In such type of voice, degree of aperiodicity in the glottal source is quantified by measurement of the jitter. During creaky phonation, jitter values are higher than other phonation types [14-16].

- Falsetto voice: in such type of voice production vocal folds are stretched longitudinally (thin). Therefore vibrating mass is smaller hence tone is higher [15].
- Breathy voice: It is noticeable as compound phonation type. Such type of voice production has vocal fold vibration which is inefficient due to incomplete closure of the glottis [15].

F_0 rises when the vowel follows an unvoiced volatile and decrease when it follows a voiced explosive.

1.5 Application of Speaker Recognition

In the last few years use of biometric system has become a reality. There are lots of commercial as well as personal applications where biometric is used for security purpose. Speaker verification has gained a huge acceptance in both government and financial sectors for secure authentication [7] [51-52]. Australian Government organization Centrelink, use speaker verification for authentication of recipients using telephone transactions [53]. Possible applications of speaker recognition are forensic investigation, telephone banking, access control, user authentication etc. [176].

Speaker recognition has more potential to other biometric such as face recognition, finger prints, and retina scans. The main advantage of speaker recognition over other biometric is low cost, high acceptance and non-invasive character of speech acquisition. To develop a speaker recognition system, expensive equipment as well as direct participation of speakers is not required. Speaker recognition have potential to eliminate the need of carrying debit card, credit card, remembering password for bank account or any other security locks and many other online services [8] [29] [61]. With the continuous improvement in reliability of speaker recognition technology, its usability has increased. Now days, use of speaker recognition has become a commercial reality and part of consumer's everyday life [52] [230].

The performance of speaker recognition system is vulnerable to change in speaker characteristics such as age, health problems, speaking environment etc. Another disadvantage is that it is possible to play a recorded voice instead of the actual voice

of a speaker [52-55]. The use of speaker recognition are continuously increasing, there are many areas where speaker recognition can be used [23] [198].

- **Access Control:** Controlling access to computer networks
- **Transaction Authentication:** For telephone banking and account access control
- **Law Enforcement:** Used in home parole monitoring and prison call monitoring
- **Speech Data Management:** Used for voice mail or intelligent answering machines. E.g. speech skimming or audio mining applications.
- **Personalization:** Device customization, store and fetch personal setting based on user verification for multi user site or device [51].

All the above mentioned applications require robust speaker recognition techniques. The requirement of robustness in case of speaker recognition system can be explained with the help of an example. In telephone based services a user speaks in many circumstances (in noisy environment or street), use different communication medium (telephone or mobile), differs the distance of microphone etc. Therefore robustness is the key factor for deciding the success of speaker recognition system. In the area, the first international patent was filed in 1983 by Michele Cavazza, Alberto Ciaramella. This invention relates to analysis of speech characteristics of speakers, in particular, to a device for a speaker's verification [56].

Barclays Wealth and Investment Management was the first organization in the world to deploy passive voice security services. The basic aim behind using voice based security was transforming the customer service experience. By using this technique customers are automatically verified as they speak with a service center executive [57]

In August 2014 **GoVivace**, developed speaker identification system by using voice biometric technology. This technology can be used for rapid voice sample matching with thousands or millions of voice recordings. The purpose of implementing this technology is to identify callers in enterprise contact canter settings where security is a major concern. GoVivace's SI technology is also available as an engine. They provide software developer kit (SDK), library as well as the Simple

Object Access Protocol (SOAP) and representational state transfer (REST) Application Programming Interfaces (APIs) to use the software as a service [58]. HSBC is rolling out voice recognition and touch ID services for 15 million customers by the summer in a big step towards biometric banking in the UK [59].

The popularity of voice biometric has risen more in past few years. According to **Opus** research, more than a half billion voiceprint will be in record, alone by 2020. People found more comfortable with biometric authentication [5].

1.6 Strengths and Weaknesses of Speaker Recognition System

With rapid advancements in the area of speaker recognition system, it is ready for use. But it is not a universal solution for security. The main strength of speaker recognition technology is that it depends on speech signal which can be acquired very easily in this digital era [7] [23]**.** The weakness of speaker recognition is that it is easily affected by speaker's health. The variability in channel and microphones affects system performance. Robustness against channel variability is the biggest challenge for the current system. There are efforts made to overcome from such type of weaknesses. For example voice biometric combined with face recognition is one of the solutions to overcome from such weaknesses i.e. increase security levels to make more robust systems [23] [61].

1.7 Identified Issues in the Speaker Recognition System

During the implementation of the speaker recognition system, many problems occur. Some specific one is discussed here such as:

- **Scalability**

Time of identification in speaker recognition system increases with the increase in the number of speakers in the voice database. Hence performance of recognition decreases with respect to increase in speaker models [7] [27].

- **Channel divergence problem**

Channel mismatch problem arises due to differences occur during acquisition of training and testing data [4] [7].

- **Time complexity:**

The efficiency of automatic speaker recognition system is decided by the time taken by the system during testing process. To achieve better accuracy, higher dimensional feature vectors are needed which again add higher computational complexity and increase computation time. Therefore there is a trade-off between computational time complexity and speaker identification rate [4] [12] [62].

- **Performance against noisy speech signal**

Practically it is next to impossible to capture a noise less speech, so a system which is robust against noise is required to develop.

1.8 Motivation for the Research

Research in speaker recognition systems have been continued for many years. This technology nowadays widely used for secure authentication. Speaker recognition is defined as the process of recognizing a person by his/her voice. This methodology allows user identity by their voice. The goal of speaker recognition system is to provide secure authentication in daily life such as telephone banking, access control, information services, security check for confidential areas etc. In the current scenario, it becomes a strong security feature for many confidential areas [4]. The performance of speaker identification system is affected by many factors. Figure- 1.8 shows the characteristics of different types of speech features.

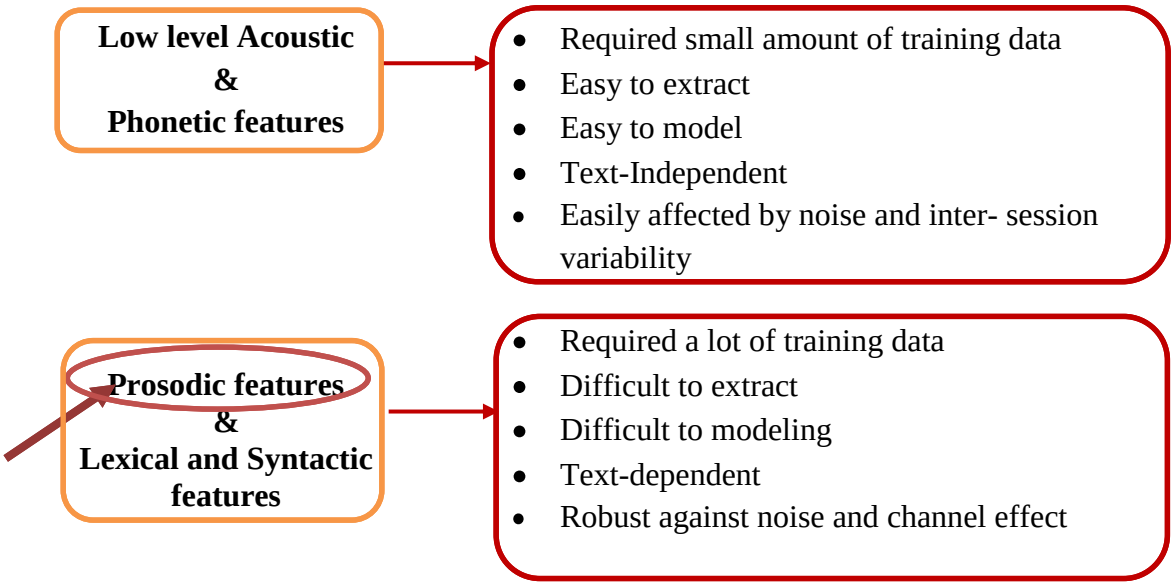


Figure- 1.8: Characteristics of Different Types of Speech Features

Research in speaker recognition is mainly concerned on the development of fast and robust system, which can work in noisy environment as well as channel mismatch. The time required by the traditional classifier system has to be reduced to make the identification system more useful. If classifier GMM is used, then computational complexity can be reduced by two ways; by reducing the number of mixtures or by reducing the dimension of feature vectors. But in case of reducing the dimension of feature vectors speaker identification rate decreases and this results in making the identification system impractical for using. Therefore, another option is to reduce dimension of feature vector, without losing important information contained in the speech features.

System performance depends on feature extraction technique such as MFCC, LPCC, LPC, Prosodic etc. Commonly used feature extraction technique is MFCC but it is not suitable in the case of noisy data while prosodic features are robust against the noisy data [7] [12] [25]. Many researchers have come up with different prosodic features for improving system performance. Hence, the main motivation behind the proposed research is to analyse the different prosodic features available and to find out the way of improving system performance.

1.9 Research Problem

From the foregoing discussion, it is pertinent that security is a big challenge in the current digital era where insecurity is everywhere. The potential of speaker recognition technology is that it relies on a signal (voice) which is natural and available unobtrusively to acquire without any special equipment or training. The primary use of this technology is for remote system accessibility and forensics. Also it is easy to use and portable (portable as handhelds device) and the leading factor is high accuracy. Keeping this in mind, the researcher has formulated a problem as under in order to improve the accuracy of speaker recognition system;

Improving Performance of Speaker Recognition Using Prosodic Features

1.10 Objective of the Research

In order to achieve the most general goal to improve the performance of speaker recognition system using Prosodic features and Gaussian Mixture Model (GMM), the following objectives have been set fourth:

- To review and critically examine the speaker recognition technology and to identify key issues that needs to be addressed in the real-world deployment of this technology.
- To study about the ways for improving speaker recognition performance and noise robustness for real-world operational conditions,
- To study about the alternatives to the present state-of-the-art approaches for speaker recognition, and to identify the ones that offer practical advantages,
- To demonstrate the operation of the speaker recognition technology in real-life or close to real-life scenarios.
- To find out the better modeling technique and pattern matching technique.
- To conduct a detailed study on Prosodic features extraction technique.
- To design an algorithm to extract the features of speaker's voice and to create a database for the same.
- To design and develop a speaker recognition system with better performance.

- To test the proposed system by analyzing the words spoken by a speaker to verify whether the speaker is a true speaker or not (i.e. speaker verification) and to identify a particular speaker among a group of persons by accepting or rejecting.
- To evaluate the performance of the proposed system.
- To compare the performance of the proposed speaker recognition system with the other recognition system in the area.
- To validate the proposed system.

1.11 Methodology Followed

The basic methodology is tantamount to a list of things that we might try in order to reach out to the ultimate goal.

- Conceptualization and Review of the available literature;
- Development of the Conceptual Speaker Recognition Framework;
- Expert-Review and Revision of the proposed Speaker Recognition Framework;
- Implementation of the proposed Framework in order to improve system performance;
- Experimentation;
- Preview and Pre-Tryout;
- Tryout;
- Assessment of Effectiveness;
- Documentation and Finalization

1.12 Expected Deliverables

The following are the expected outcomes of the research:

- The list of speech features which are responsible to enhance the system performance.
- The List of prosodic features which are more robust against noise.
- A framework for developing speaker recognition system.

- Design and development of an improved speaker recognition system with better performance.
- Design and development of an algorithm to identify a registered speaker.
- Design and development of speaker's voice database.
- An experimental study showing the usefulness of the proposed speaker recognition system.
- A comparative study to prove that the proposed speaker recognition system is better one.

1.13 Limitation

In order to keep the research precise and within the time boundary, the thesis has few limitations. These are as follows:

- The proposed work focuses only on speaker Identification and not on speaker verification.
- Voice database is limited and it also contains some background noise.
- The performance of proposed system is tested only within laboratory setup. Implementation of the same has not been done in real scenario.
- Training and testing is performed only on English and Hindi voice database.
- During testing, channel mismatch has not been considered.

1.14 Thesis Outline

It is expected that the proposed research will make the speaker identification task more accurate. The thesis presents detailed study about the same. Apart from annexure, references and other components, this study includes seven chapters. A summary of each chapter is presented below.

Chapter- 2: Literature Review

This chapter provides concise definition and discussion about speaker recognition technology. It presents the literature review, basic terminology of speaker recognition and speaker recognition methodology in details. It also presents the general overview

of human speech production, and consequently introduces the review literature of speaker modeling technology and the estimated model.

Chapter- 3: Framework for Speaker Recognition

In this chapter, a framework for development of speaker recognition system is proposed along with the premises of framework. A guideline is given for the proposed framework. The proposed framework has the six phases including acquiring speech signal, feature extraction, modeling, pattern matching and decision phase and performance evaluation phase. In addition, limitation of the proposed framework has also been discussed.

Chapter- 4: Implementation of the Proposed Framework Using Prosodic Features

In this chapter, a detailed description is presented on how the proposed framework for speaker identification system is implemented. A detailed study of prosodic features with their extraction procedure is also presented. Most of the components of this system are implemented individually by MATLAB as well as Praat software as per the suitability of framework component. Database creation process is also given in detail. Voice database is created for training and testing of speaker recognition. At the time of voice recording some background noise exists. In addition, discussions about prosodic features which are used in this research along with an algorithm for speaker identification are presented.

Chapter- 5: Experiments and Result

In this chapter, the experimental results are presented. During experiment training condition, enrollment condition and test conditions are discussed. Results obtained from the individual feature extraction techniques have also been discussed. This chapter deals with the results obtained in the study and its data analysis. Calculated results are presented graphically. Comparison of accuracy results of speaker recognition systems using different feature extraction methods are calculated and compared. It has also shown that accuracy of automatic speaker recognition system

depends on various factors including which feature extraction technique is used, what speaker modeling technique is used, recording conditions of speech etc.

Chapter- 6: Validation of the Framework

In this chapter, concept of validation has been discussed and methodology for validation has been designed. During validation, hypothesis have been formulated and tested on the basis of statistical analysis. Student t-test has been used for testing the hypothesis. Additionally, some issues have been discussed which occur at the time of development of speaker recognition system. The main issues including noise, headset mismatch, sampling rate of speech signal etc. are discussed here. In addition, the ways to improve speaker recognition system performance have been described.

Chapter- 7: Conclusion and Future Work

This chapter describes the findings of this research. It presents an overview of the research and outlines in terms of its major findings. In addition, it demonstrates the significant contribution of this research in reference to speaker recognition technology. It also discusses probable limitations of the research and proposes directions for future research.

Chapter- 2
Literature Review

CHAPTER-2: LITERATURE REVIEW

The speech of man is like embroidered tapestries, since like them this has to be extended in order to display its patterns, but when it is rolled up it conceals and distorts them.

—Themistocles

2.1 Introduction

Biometric can be considered as a tool used to measure human characteristics or individuality. Biometric includes behavioral or physiological measurements of the individual. Behavior biometric is performed by Voice, Signature, Keystrokes, and Typing etc., whereas physiological biometric includes iris, face, retina, fingerprints, ear, DNA etc. Now a days voice biometrics is emerging research area. From the literature survey, it has been identified that biometric verification or realistic authentication is used to recognize an individual through his/her voice individual characteristic. Recognizing a person's individuality by his/her voice is known as Automatic Speaker Recognition (ASR). Speaker recognition falls in the category of biometric security systems. This chapter provides a review of the relevant literature and terminology related to the research [55] [90].

Technology improvements are required in many fields such as online transactions, communications in the field of banking and networking. Now a day's biometric techniques are used in many fields for security purpose, each voice has their individual characteristic that's made it impossible to clone it. Such type of systems uses some human characteristics which are unique in each person, e.g. fingerprints, voice, retina and DNA, etc. Automatic speaker recognition can be defined as; it is an automatic recognition system that uses a person's individuality for their identification [57] [97].

In this technological era, it is very common that when you receive a phone call and the moment caller saying 'Hello, it is me' and you answer immediately 'Ya- Ya , I recognize your voice'. This is an example of speaker recognition, which occurs in our daily life, this shows that a human easily recognize the caller from his/her voice only. There is a growing demand for person authentication for accessing different services including transaction, access control and forensics, etc. In order to consider

security perspective speaker recognition system is highly recommended. The reason behind its popularity that it is very user friendly, cost effective and robustness etc.

2.2 Background

Voice is the most natural way to express their thoughts for humans; therefore it is the most common medium used for communication. With the developments are made in technologies the way of communication medium also changes. Therefore to act together with machine and computers people needs to interact with computer programmed devices. To make human interaction as easy and appropriate as possible, there has been a great effort made in human-computer research oriented technologies during last few decades and still continues [15]. Automatic speaker recognition systems have many advanced developments in last 6 decades. Today's such type of systems are used in many areas such as access control, authentication, online banking and some forensic applications. The first speaker recognition system comes in to the existence at 1962. In 1962 Lawrence G. Kersta, a Bell Laboratories Physicist published an article entitled, "Voiceprint Identification" published by Nature [30] [65].

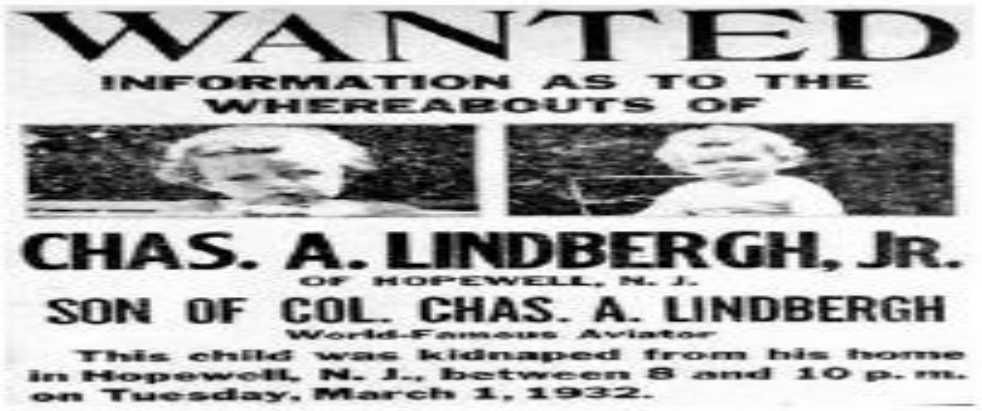


Figure-2.2: Lindbergh Wanted Poster [30]

It was 1932 when baby boy Charles and Anne Lindbergh's kidnapped and killed. During the cycle Charles Lindbergh who sat in nearby car, listen an unknown person voice who says "Hey Doctor, over here, over here" but he not able to see him. After Two and a half years later during the trial of the suspect kidnapper, Charles

Lindbergh had heard the same voice (Bruno Hauptmann voice) without seeing his face and he claimed that this person voice is same as the voice heard during graveyard. This case inspires 'Frances McGehee' to research on the reliability of ear witnesses and it is first academic research. Later she published two articles entitled as "Reliability of the Identification of the Human Voice" and "An Experimental Study Voice Recognition". Due to McGehee research in speaker recognition give another research topic in forensics and psychology [30] [65-66]. In the 1960's to make autonomous speaker recognition system many developments have been done.

A physiological model of human speech developed in 1960 by Gunnar Fant. The Fant model and other parallel research become the base to speech analysis for speaker recognition as well as speech recognition. In 1960's Leonard E. Baum and others established a stochastic model for Markov processes, to determine hidden parameters of a statistical model. Due to observe these features the model is called Hidden Markov Model (HMM). This statistical model broadly used in speech recognition but limited in speaker recognition [30]. The use of voice as a biometric authentication has very long period. The studies have been made in early of twentieth century, on vocal recognition as an academic venture. As the use of computer rises it is led to the development of speaker recognition system. In early days of computing idea of speech recognition was proposed. The first step in the direction of machine based recognition system initiated with the invention of spectrogram during Second World War. It was Kersta who contributed research in 1960's on voiceprint and this contribution opens new way in the field of speaker recognition technology [30] [62] [65-66].

In the early of 1970's the first speaker recognition system was developed. The developed system is text-dependent and multi-model speaker authentication system which is used for access control at Texas Instruments. Later many improvements were made to developing text-independent system. And now more work performs to developing more robust and accurate text-independent speaker recognition system. Many researchers used NIST SRE speech database for evaluation of accuracy/robustness to text-independent speaker recognition system. Now day's speaker recognition is a worldwide activity and a burning area of research. As

compared to other biometrics “speaker recognition” is commercial venture across the world [4] [30] [51] [62].

2.3 Related Study

Even though a majority of speaker recognition system approach are based on cepstral coefficient such as MFCC. Several systems have been proposed by using cepstral features [67- 69]. But in last one decade it is observed that researcher make interest in prosodic features to develop speaker recognition system. Prosodic features produce best result as compared to MFCC’s and more robust against noise [70-73]. The generally used prosodic features are pitch and energy contours [74]. As it is a fact that prosodic features are related to phonemes and syllables (such as pitch and duration) are less sensitive to channel distortion than cepstral features [71] [75]. As result obtained in [102] author said that to obtain best result the prosodic feature (duration, pitch and energy) are more suitable. Prosodic features are based on speakers speaking style and speaker’s intonation. Prosodic systems are required large amount of voice data to train speaker models [71-75].

The work presented in [70] by K. Sonmez et al. propose a system based on distances among pitch histogram values and demonstrate that pitch has a log normal distribution. Also a study made by the same author in [76] proposed a stylization method which is based on segmentation of pitch contour [217]. Here extract a specific set of parameters for each segment such as median, segment feature duration and slope of pitch contour. Model created for each feature by using GMM.

In [77] Licia Sbattella et al. proposed a speaker recognition system based on prosodic features which is able to recognize speaking style and emotions of a speaker. In this study authors used two LDA-based classifiers. These classifiers depend on two sets of prosodic features. The result shows that obtained Ac= 71% for emotions and 86% for speaking style. The PrEmA tool used for extract prosodic characteristics such as speaking style and emotions of a speaker. For experiment and validation taken 18 segments randomly to evaluate each emotion, from total available 90 segments.

As author Luciana Ferrer et al. says in [78] that prosodic features used successfully in speaker recognition from more than last one decade. Best performing prosodic based speaker recognition system to till date has been based on syllable features extracted from speech signal. In this study author used two different modeling techniques to create speaker models. The one is joint factor analysis (JFA) of GMM means and support vector machine (SVM) modeling of GMM weights. The result shows the improvement of 30% in detection cost function (DCF). The experiment conducted for text-independent verification system, where recognized claimed identity of a speaker that is whether claimed identity is true or false.

As authors introduces continuous prosodic features in [17] for speaker recognition and model with JFA. These features also used for language identification. In this study the used prosodic features are syllable like features, pitch and energy based features. Since used prosodic features are continuous hence GMM is used for modeling technique. Author used pseudo syllable as the main unit for extracting prosodic features. The experiment is performed on NIST-2006 database (core condition, English-only trials) and evaluated EER is 16.6% and 14.6% for speaker recognition. As author says about the achieved system performance, that for better performance we need large amount of training data [154].

In [135] authors used prosodic features for speech feature extraction. Prosodic features generally based on syllable level of speech such as contextual, positional and phonological features are extracting from syllable. In this study used prosodic features are duration and pitch (F0) values of the syllable. Neural network used for modeling prosodic features, and also use to extract speaker specific prosodic information and this methodology also useful to enhancing the system performance of speaker, speech and language identification systems. Experiments were done for identification of Hindi dialects.

2.4 First Speaker Recognition System

It was 1962; Lawrence G. Kersta physicist Bell Laboratories published an article entitled “Voiceprint Identification” published by Nature. In 1960 law enforcement agencies approached Bell Laboratories to identifying those callers who had made several bomb threats over telephone. This task was given to Lawrence G. Kersta.

After some time he claimed on the basis of his research that he had a method to recognize individuals with very high success rates. Three other scientists of Bell Laboratories named Potter, Kopp and Green earlier used his method for speaker recognition for military applications during World War II. They have developed a spectrogram for speech signal. Spectrogram contains frequency and intensity of speech signal with respect to time [30] [80-83]. Figure-2.4 shows the sample of spectrogram of a speech signal.

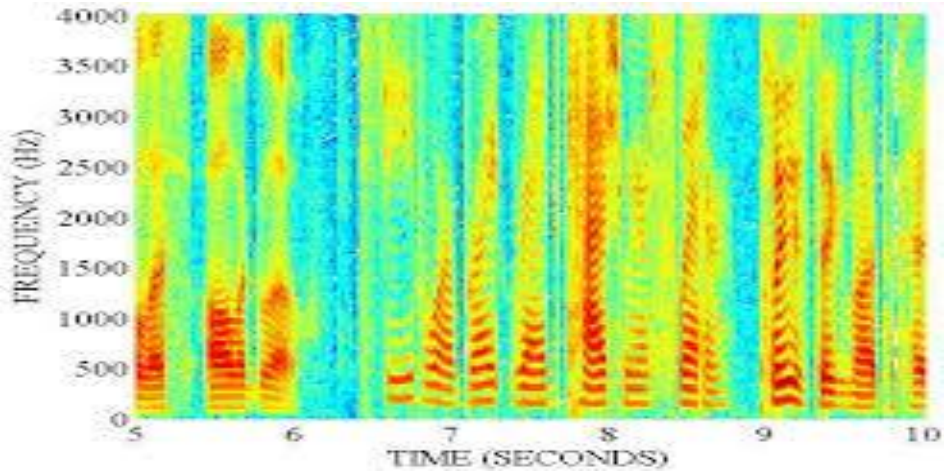


Figure-2.4: Sample of Spectrogram of a Speech Signal [173]

Identification method given by Kersta is based on the aural-visual method. Spectrogram produced by a sample speech signal and this is examined visually for pattern matching. Result achieved by this method can be very high, assumed an expert interpreter and good environmental circumstances. Even with good results achieved, machine required human interaction due to this its use in security applications is limited [30] [65] [84]. Kersta's method is still utilized in some forensic applications such as in FBI but not into real-world speaker recognition system. For automatic authentication of a person, independent recognition system would be needed and also have low error rate [80].

To develop a robust and accurate speaker recognition system feature extraction/selection, modeling and decision making algorithms have important role. This field has made many significant improvements to till date. The hidden Markov Model (stochastic model) developed by Leonard E. Baum and others in 1960's [85]. This method used to determine hidden parameters of a statistical model. Later it was

extensively used in speech as well as speaker recognition system during 1980's. Matsui and Furui said that HMM and Vector quantization (VQ) with enough training data was effective and less computationally demanding as Gaussian mixture model. VQ use for speaker feature vector but it is not consider as a serious approach to speaker recognition. As it is studied that HMM is a standard method for speech recognition and text-dependent speaker recognition systems but it is found not more suitable for text-independent speaker recognition system [86-88] [117][120][237].

2.5 Speech Production Method

Figure- 2.5 shows the human voice/speech production system. Human speech production system is capable to produce several numbers of distinct sounds. The voice production system can be divided into subsystems such as respiratory subsystem, laryngeal subsystem, and articulatory subsystem. These subsystems contains the different kind of information related to human's voice features for example respiratory subsystem (diaphragm, lungs), laryngeal subsystem (larynx) and articulatory subsystem (oral/nasal cavities, soft/hard palate, tongue, jaw, lips and teeth) [39-40]. It is known that the respiratory and laryngeal subsystems be responsible for source signal of speech. Hence F_0 (fundamental frequency i.e. the quasi-periodic cycle frequency of voiced speech) and the glottal waveform differs in a little way in terms of vowels [41- 42].

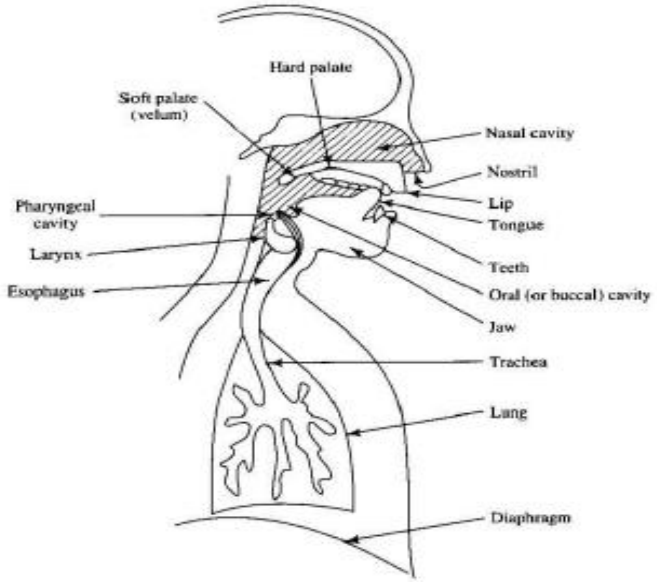


Figure-2.5: Respiratory system involved in speech production [40]

A speaker is recognized by both physiological and behavioral characteristic of their voice. These characteristics are known as vocal tract (spectral envelop) characteristics and voice source (supra-segmental features) characteristics [43]. Speech is unique for individual due to differences in size of vocal tracts of each person. The length of vocal tract is related to resonance frequencies such as formant frequencies [44].

- **Sources of variability in the speech signal:** A speech signal contains many different types of information, and this information intermingled in a way that it is very difficult to decompose and sometimes impossible. The variation in speech can be categorized in general as [45-46]:
- **Speaker Identity:** Information that carry speaker's voice permanent characteristics.
- **Linguistic:** Information that express speaker's purpose to the listener.
- **State:** Information that takes passing states of the speaker that is not relevant to speech signal.
- **Lungs:** During voice production lungs are used for inhalation and exhalation of air. They are also supplies energy to the rest of the blocks in the voice production systems. Exhalation and inhalation are differing to each other in working. By reducing the lung air pressure Inhalation is occur while exhalation initiated by an air pressure increase in the lungs [47].
- **Larynx:** Larynx is a complex system of cartilages, muscles and ligaments. It is also known as 'voice box' [47].

2.6 Biometric and Speaker Recognition

In today's environment, where insecurity is everywhere security has been one of the important issues. For providing security voice biometric is an emerging area especially for the purpose of authentication [89-90]. In voice biometric speaker recognition is performed with the help of the unique characteristics of human voice including physiological and behavioral characteristics [55] [57].

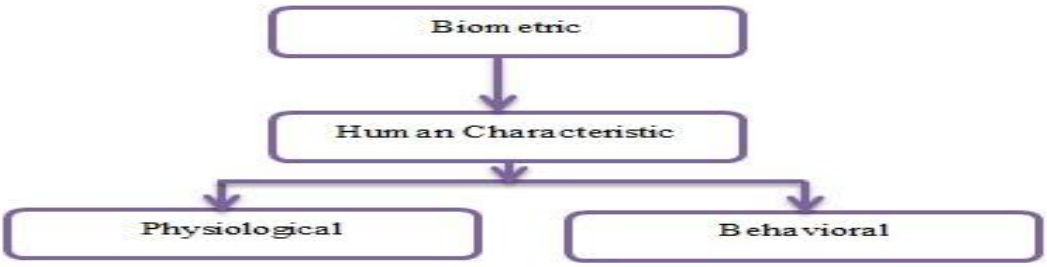


Figure-2.6: Types of Biometric

These characteristics have specific and appropriate features of voice and have potential to recognizing a person [55] [91]. With this approach it is also possible to authenticate a person irrespective of changes of environment or channel. This approach is very useful and cost effective as it is voice based biometric technique which is easily available in this digital era [92-93]. There exist many areas where this technique can be successfully implemented for security and investigation perspective. Few of the application areas of speaker recognition system includes forensics, remote access control security, web services, online calling, personalization of services and customer relationship management, voice based biometric system, voice based banking, surveillance/criminal investigation etc.[92-94] [97]. Figure-2.6 show that how biometric related to human characteristics.

Forensic science is a method of crime investigation by gathering and examining criminal's information [95]. Computer forensics or computer forensic science is an area of digital forensic science. It is used in legal verification through computers (for digital storage media). The aim of digital forensic is to identifying, analyzing, preserving, recovering and presenting specific information and judgment concerning digital information [96]. Digital forensic is mainly associated with investigation related forensic crimes through computers and the results are used in civil legal proceedings. Digital forensic uses various techniques to distinguish identity of individuals. Digital forensic is a very young area for crime investigation [97]. Forensic linguistic is used for voice identification which is also called forensic phonetics. It is performed on the basis of voice acoustic qualities (if the voice is recorded anywhere e.g. on a tape, mobile phone or any other device) [57] [98]. Identifying a speaker with the help of forensic linguistic is called forensic speaker

recognition. It is important and challenging task. Forensic speaker recognition is an application of speaker recognition [99].

The term speaker recognition refers to speaker verification as well as speaker identification. Speaker recognition/Voice recognition is the process of a person's authentication by his/her voice [93]. It is also known as Biometric Identification Technique (BIT) [91]. In this technique human traits are used for identification and verification for the purpose of access control. Speaker recognition technology is mainly used for three purposes. In authentication purpose, forensic scenario, screening and indexing applications. Authentication refers to verify the identity of a user who needs physical or logical access. Voice forensic refers to comparing two voice samples to determine the source of the same. Screening and indexing is refers to search of specific speaker speech from large voice database [94] [96] [100].

2.7 Characteristics and Categorization of Biometric

Biometric is a feature of human being by using which a person can be recognized. The following properties have been considered as guiding light to understand about what biological features are best suited for biometric [101-102]:

- Uniqueness: Something that differentiate individuals.
- Universality: Something that everybody has.
- Stability: Something which is constant over time for each & every person.
- Measurability: Something which is easy to measure.
- Acceptability: Something which is well accepted by people.
- Performance: Something which has speed, accuracy and robustness
- Non Confront: Something which cannot be easily fooled.

The above properties forms the basis to decide what features should be used as biometric. Every biometric has individual purpose for its use such as security system, crime investigation, voting system, Time accounting etc. The selection of biometric depends on the requirement for authentication [99-102]. The available biometric examples include DNA matching, eyes, ear, voice, face, fingerprint, hand geometry, signature/writing etc. The Figure- 2.7 (a) and Figure- 2.7 (b), shows about physiological and behavioral characteristics of any person.

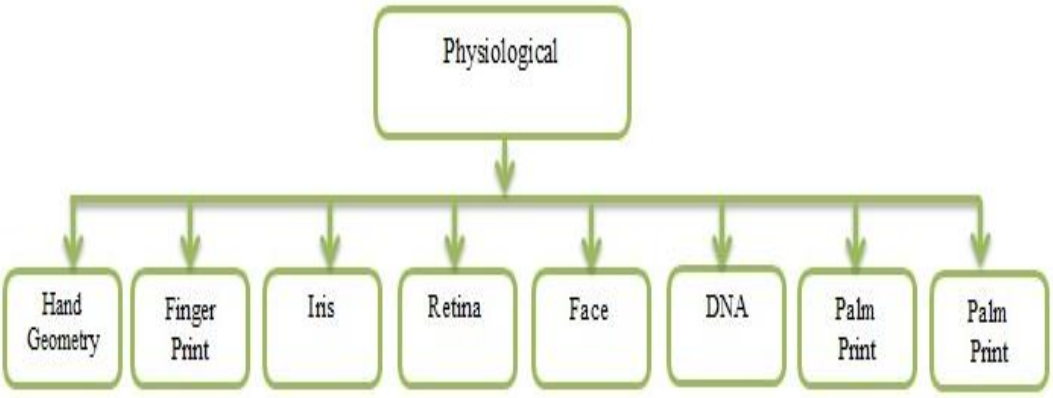


Figure-2.7 (a): Types of Physiological Characteristics of Human

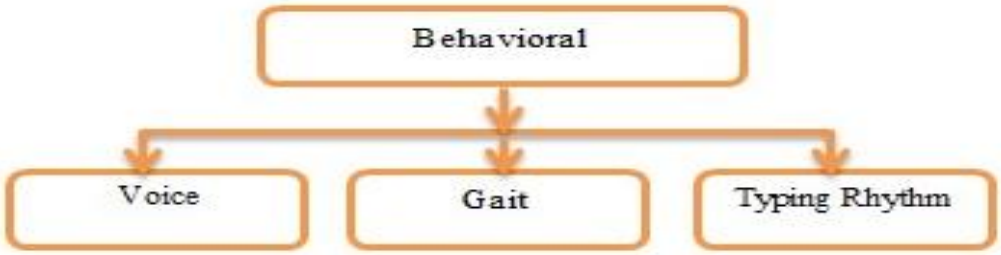


Figure-2.7 (B): Types of Behavioral Characteristics of Human

- DNA: Identification of an individual using the analysis of DNA segments
- Eyes (Eris and Retina): Use of the features found in the eyes to identify an individual.
- Ear: Identification of someone by using the shape of the ear.
- Face: The analysis of facial features for the authentication of an individual.
- Fingerprint: Use of the ridges and valleys found on the surface tips on human fingers.
- Hand geometry: use of the geometric features of the hand
- Signature/writing: The authentication of an individual by the analysis of handwriting style
- Voice (Speaker Recognition): The use of the voice/speech as a method of determining the identity of a speaker

The above mentioned are some biometrics which are used for person authentication. Biometric is considered for Security and accuracy, but the disadvantages are being

offensive of privacy and cost of implementation [99-102]. Following are some advantages of biometrics:

- Convenience
- Increased security
- Accuracy
- Non imitative
- Non sharable
- Cannot be lost
- Reduced paper work
- Easy to access

The above advantage makes biometric systems more secure and accurate. Biometric recognition is the continuously developing branch of science. It is convenient and reliable technology this allows using biometrics in common life by making this technology easier and interesting. Almost every country in the world is using at least one biometric to recognize its nationals. Use of biometric is increasing day to day for security. For example, In India 'Aadhaar' is the largest biometric database in the world, about 480 million 'Aadhaar' numbers have been assigned to the Indian nationals up to 2013 [90] [92].

2.8 Phases of Speaker Recognition System

Speaker recognition is a pattern recognition problem which is a branch of machine learning. Speaker recognition and speech recognition both come under voice recognition, which is also known as voice biometric. Speech and speaker recognition can be distinguished as 'speaker recognition' i.e. 'who is speaking' and 'speech recognition' i.e. 'what is being said' [104-105]. Extraction of speech signal is the main task in the development of speaker recognition system. To extract speech features from speech/voice signal there are many techniques available such as MFCC, LPCC, LPC, Prosodic etc. One of the speech features is spectral features of speech signal which is used to representing speaker's voice characteristics [106-107]. After feature extraction speaker models are created for individual and stored as voice database. To create models for speakers, various modeling techniques including

GMM, HMM, pattern matching, frequency estimation, vector quantization, decision tree and neural networks etc. are used mainly[104].

Automatic speaker recognition system has two phases enrollment phase and verification phase. In enrollment phase speaker's voice is recorded and specific features are extracted from speech signal/voice print. In verification phase a voice sample/utterance is compared to stored template or voice print [91] [106]. Speaker recognition can be classified into speaker verification and speaker identification. Identification and verification can be explained as: In case of speaker identification, the voice sample is compared with multiple templates from stored voice database and the best match is selected e.g. comparing a voice sample of an assaulter from previously predictable voice database of criminals and trying to find best match in this case one speaker's voice is matched against 'n' templates so it is also called 1: n match. While in case of verification speech sample is compared with the claimed voice print e.g. presenting your identity card to security officer, the security agent compare your face to the photo attached in the identity card and verify that either claimed identity is accepted or rejected so it is 1:1 match [91] [105-106] [108]. Figure- 2.8 shows the common steps involved in development of Automatic speaker recognition system.

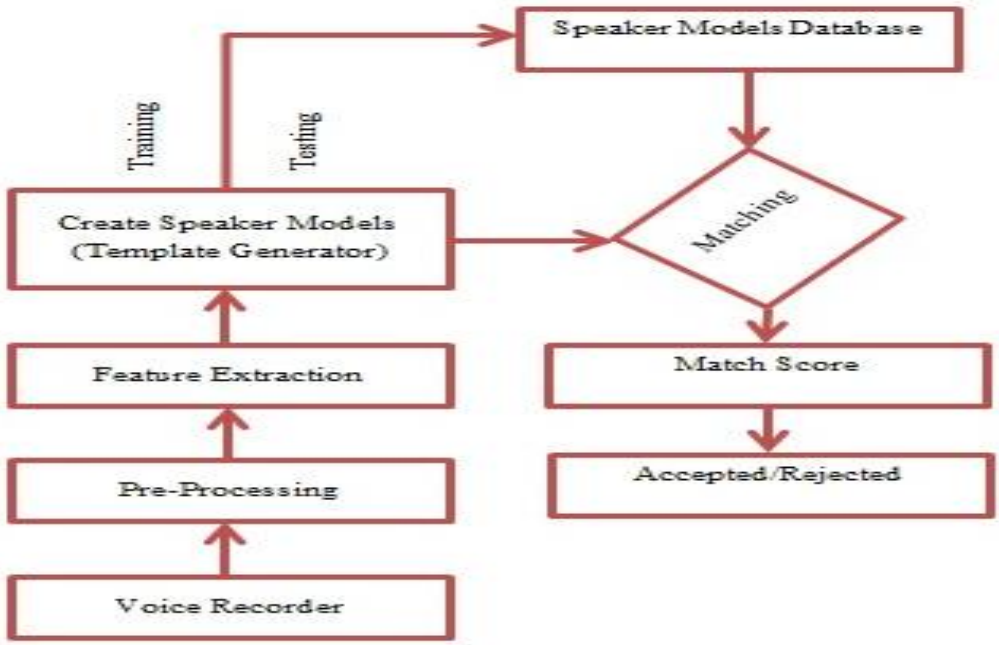


Figure-2.8: Phases of Automatic Speaker Recognition System

On the basis of process involved in verification and identification it can be easily inferred that verification is faster than identification. During the study in this area, it has found that in most of the cases identification is performed first to find the best match then only verification is done to reach out a conclusive result. It can be justified by taking an example suppose that "if a voice sample of a suspected assaulter is captured then this voice sample is matched with the previous formed voice database and tried to find out best match of this sample (that is identification) after that verification is performed then gives conclusive result declaring true or false and the best match voice is belong to that assaulter or not [76] [78] [118].

2.9 Development of Speaker Recognition System

In 1960's various developments has been made to make it possible autonomous speaker recognition system. The development made in this time covered a wide-ranging discipline in the field of speaker recognition system. For example, Gunnar Fant in 1960 developed a physiological model of human voice production system. This model sets a basis for understanding speech analysis for speaker recognition and speech recognition both. Fant idea of physiological model of voice directed future researchers to characterize speech signal as a linear source filter model. Through Fant model it is possible to make many advances to discovering human voice characteristics which is individual recognizable [30] [80] [109]. Figure-2.9 shows the timeline of crucial development in the field of automatic speaker recognition system.

In 1963, Bogert, Healy and Tukey published a research article titled "The Quefrency Alanysis of the Time Series for Echos: Cepstrum, Pseudo-Auto-Covariance, Cross-Cepstrum, and Saphe Cracking" [110]. This article (oddly titled) makes a study about echo detection. In 1965, Cooley and Tukey published their research on digital implementation for the Fourier Transform and later it is known as Cooley-Tukey Fast Fourier Transform (FFT) [111]. Cooley and Tukey method, contributed as efficient method of frequency analysis of digital signal. In 1969 Michael Noll who inspired by the echo detection cepstrum (Bogert, Healy and Tukey article) gave idea for pitch detection of a human voice by using cepstrum [112]. Ronald Schafer who joined Oppenheim research makes contribution to build on Noll's pitch detection which used for cepstral analysis to model speech signal. Later

the developed cepstral speech model was used as an important tool for speaker recognition [113-114].

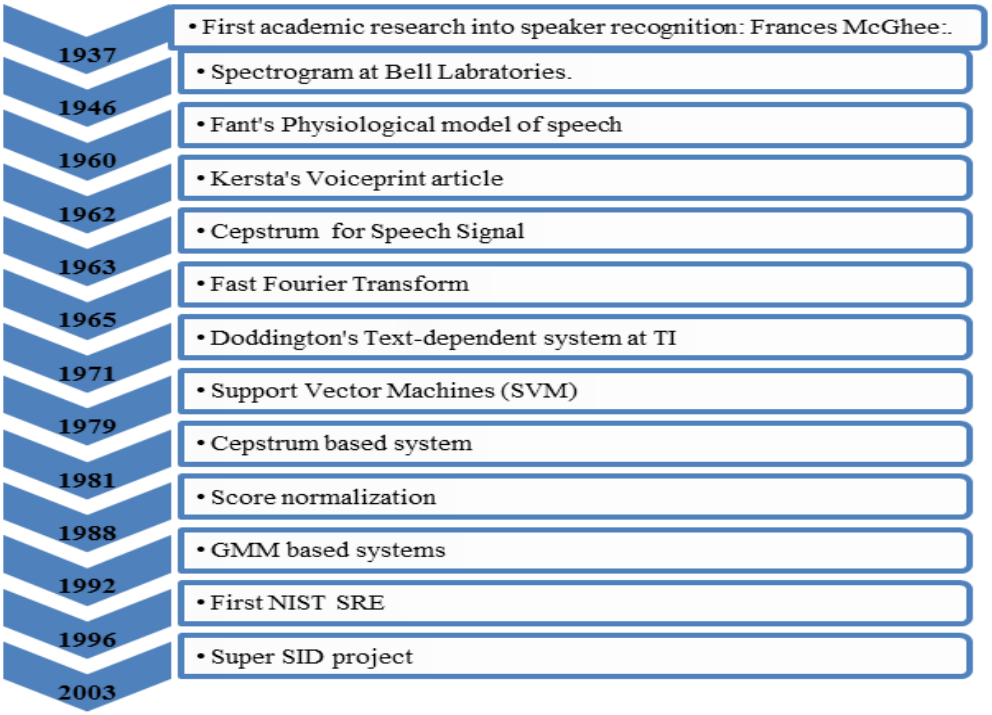


Figure- 2.9: Timeline of Major Speaker Recognition development system

Advancement in the area of ASR 1974 to 2017 has been shown on the basis of different parameters. The summary is presented in table-2.9 the terms defined in columns in table-2.9 are: "Developer/Author/year" refers to who developed and used the particular techniques, "Organization" is the lab or company or institution where the work has been done, "Database (Population)" is the number of speakers in which test has been conducted for speaker verification or identification, "Features extraction techniques" that is refers to the technique used to measure speech signal features, "Modeling" refers to the method is used for the matching of signal, "Voice type" shows how the voice is acquired such as telephone, lab, noisy place etc., "text-type" means the system is text-dependent or text-independent, "accuracy" sows that how much the system is accurate for recognition. The complete information in the table gives a general overview of speaker recognition research from 1974 to 2016. To be focused here have taken some selected & significant studies [63-64].

Speaker recognition system can be used in access control, telephone banking, biometric investigation, crime investigation etc. There is a number of commercial/organizational/personal automatic speaker recognition system including

T-NETIX, ITT, Lernout & Hauspie, Veritel and voice control system. Studies say that 'sprint's voice FONCARD' is the largest scale deployment of any biometric system till date [140-142]. It is very difficult to make a meaningful comparison between text-dependent and text-independent speaker recognition system in absence of standard comparison criteria. As there are different techniques dealing with different recognition problems so it is not easy to decide which one is better. For instance Gish's segmental 'Gaussian model' and Reynolds' 'Gaussian Mixture Model' for text-independent approaches are used to deal with unique problems e.g. sounds or articulations present in the test signal, but not in training voice signal [142] [155].

During the literature survey, it was found that the following areas of speaker recognition have been gaining great attention in terms of research:

- Accuracy of speaker recognition system [141] [143-144].
- Development of more robust system for speaker recognition [145]
- Development of feature extraction techniques for voice feature extraction[146]
- Different speaker modeling techniques for speaker identification and verification etc.[147-148]

Table-2.9 summarizes few relevant research works in the area. Of course, these works have their own worth. Nobody can deny the importance of these works. But still there exist many questions which are still unanswered:

- Is there any standard way available to decide about the number of voice parameters?
- Which voice parameters are essential to include during the development of speaker recognition system?
- What is the maximum time limit for voice recording to achieve maximum accuracy?
- It is possible to make a robust speaker recognition system in real which is not affected by background noise, session variability, recoding environment etc.?
- What and how many speech parameters should be included to develop a robust speaker recognition system?
- What factors are more responsible for enhancing the system performance as well as degradation?

The main objective of the work is to focus on the recent advances and development in the area of speaker recognition and the problems still remains unanswered.

Developer/Author/year	**Organization**	**Database(population)**	**FeaturesExtraction/ Modeling/Matching Method**	**Features**	**Voice type**	**Text type/system type**	**Accuracy (%)**
Moataz El Ayadi et.al./2017 [9]	Dept. of EM & P, Faculty of Engineering, Cairo University, Giza, Egypt/ NIST 20 0 0	NIST 20 0 0(100 male) & COSINE database	GMM-UBM/robust i-vector	MFCC	Lab	Independent	16% relative improvement
Ye Tian, Zhe Chen, Fuliang Yin/ 2017 [161]	SICS, Dalian University of Technology, Dalian, China	100 experiment	DIEKF/IMAGE Method	Acoustic	Microphone array network (room reverberations)	Independent/Speaker Tracking	DIEKF-3 is better than DKF & DIEKF-1
B. S. Atal/ 1974 [28]	Bell laboratories	10 speakers	LPC	Cepstr-um	Lab	Independent	93%
Alfredo Maesa/2012[171]	Voxforge.org	450 speakers	MFCC	spectral subtraction	Audio data-base	Independent/Identification	>96%
Douglas A. Reynolds/1995[149]	Lincoln Laboratory	49	GMM	Short Utterance	Telephone	Independent/Identification	96.8%
Rabah W et.al./2004[172]	King Abdulaziz University	20	SVD-based algorithm	LPC/Cepstral	office	Independent/Identification	94%
Najim Dehak et.al./2007 [17]	NIST-2006	NA	GMM-JFA	prosodic features	Lab	language identification	Improvement 8% (all trials) and 12% (English only
P. Krishnamoorthy/2011[148]	TIMIT	100	GMM-UBM	MFCC	Lab	Independent/Identification	80%
Sriram Ganapathy/2014[178]	SRE database (NIST-2010)	random	AR model	FDLP	Lab	Dependent/Recognition	relative improvements of up to 25%
Hesham Tolba/2011[179]	Arabic speakers	10	HMM/GHMM	MFCC	Lab	Dependent/Independent	80%
Chih-Hung Chou et. Al./2015 [180]	ALTERA DE2-70,	16	VQ/GMM-PQ	OOS	Lab	Dependent	Recognition Rate 88.3%

Emmanual Perrin et. Al./1994 [181]	E-HERRIOT	60	Acoustical Signature	Vocalic Space	Standard Protocol	Dependent	>90%
Ergun Yucesoy et al. /2016[125]	E Gender database INTERSPEECH 2010	299 speakers	GMM-SV	prosodic features	Lab	Dependent	90.4%, 54.1% and 53.5% in gender, age, and age & gender categories
Xuanjing Shen et al. /2014[183]	TIMIT speech database	38(19 Female, 19 Male)	LFA-SVM Gaussian kernel	12-order MFCC,	Lab	NA	81.52%
Anzar S.M et al./ 2016[184]	English language database for adaptive speaker recognition(ELDASR)	50(Male/Female)	GMM/MFCC	MFCC super-template	Lab with intra-class-variations	NA	Improved (% NA)
Isaias Sanchez-Cortina et al./2016 [185]	videoLectures.net, poli Media	NA	logistic regression model	NB model	Online educational lectures	Dependent	Relative improvement between 2% and 7%.

Table- 2.9: Progress in Speaker Recognition in Last Six Decades (Some Selected)

* RMA: Royal Military Academy
* NDSF: Normalized Dynamic Spectral Feature
* VER: Verification error rate
* RER: Rejection error rate
* FDLP: frequency domain linear prediction
* SRE: NIST-2010 speaker recognition evaluation database
* AR Model: Auto Regressive Model
* NB Model: word-dependent naïve Bayes (NB)
* JFA: joint factor analysis
* DIEKF: Distributed iterated extended Kalman filter

Speaker recognition is one of the emerging research domain for persons authentication and enhances the security in the areas including access control, voice authentication, banking by telephone and many more [122][145]. It is very difficult to find the fix voice parameters by which a good speaker recognition system with maximum accuracy can be developed. Therefore to design and develop robust speaker recognition system, continuous effort is needed. Speaker recognition technology has many advancement and development till date but technology development and evaluation are two sides of the same coin. So keeping this point in mind it can be concluded that without having a good measure of progress nobody can make valuable progress [122]. Till date various investigations have been proposed for evaluation of speaker recognition but in real a complete tool has not yet been developed [12] [62].

2.10 Principles of Speaker Recognition

On the basis of speaker recognition application, it can be divided into three specific tasks named as speaker identification, verification, segmentation and clustering [116-119]. Speaker identification is defined as which speaker's belong the voice sample, from a group of known speakers. Speaker verification is defined as, whether a speaker is who s/he claims to be according s/he voice sample. This task also known as voice authentication and speaker detection [4].

Speaker segmentation and clustering is defined as, it is a way to index audio records to make retrieval easy. Generally it is consider that the speech of a particular speaker is available for processing. But suppose this is not the case and the speech is mixed with other speakers then it is required to segregate the speech sample into segments to perform recognition process. So the aim of this task is to divide the given speech sample into identical segments and then find them via speaker identity [20] [62] [217].

A speaker recognition system has two phases named as training phase and testing phase. In training phase a speaker register by providing a speech sample to the system. The system extracts speech features from the speech sample to create a speaker voice model of the enrolled speaker. While in testing phase a speaker provide a speech sample (test sample) that is use to make a decision on the basis of speaker's

voice models which already register in the system. In speaker identification test voice sample compare with all the voice sample models. In speaker verification test sample compare with only to the speaker model for which identity claimed [4] [12] [62].

In the modern digital era where insecurity is prevailing everywhere maintaining security is a big challenge. Lots of cases are being reported in daily life related to edit audio clips and wrong claim for identity. Speaker recognition is a technique to automatically recognizing a speaker on the basis of information extracted by his/her speech. It can be divided into two categories; speaker identification and speaker verification. This method provides security in confidential areas. For example, to prove the claimed identity of a person, his/her voice is treated through forensic test [120]. This technique is very useful to authenticate a person's identity. The aim of automatic speaker recognition is to acquire the voice of speakers and to create voice model for each speaker and finally to compares these models with an utterance of the speaker to prove his/her identity. Different individuals have different voice. Even voice of a person may differ time to time. The variation in different people's voices is termed as inter-speaker variability and the variation in the same person's voice is termed as intra-speaker variability [121-123]. The speaker recognition relies on the ability of human being to identify other person's voice with the following observations:

- Human being is able to recognize the voice of any person (whom he/she knows and communicated to each other frequently).
- A person is able to identify other person's voice to whom he/she communicate frequently irrespective of the communication medium or background noise.
- A person is able to recognize other person even if communication happens after a long gap (even years).
- A human is able to identify the 'state of mind' person is also able to recognize the 'state of mind' (emotions level that is speaker is happy, sad, neutral, cold, some health issue etc.) of the speaker by listening his/her voice.

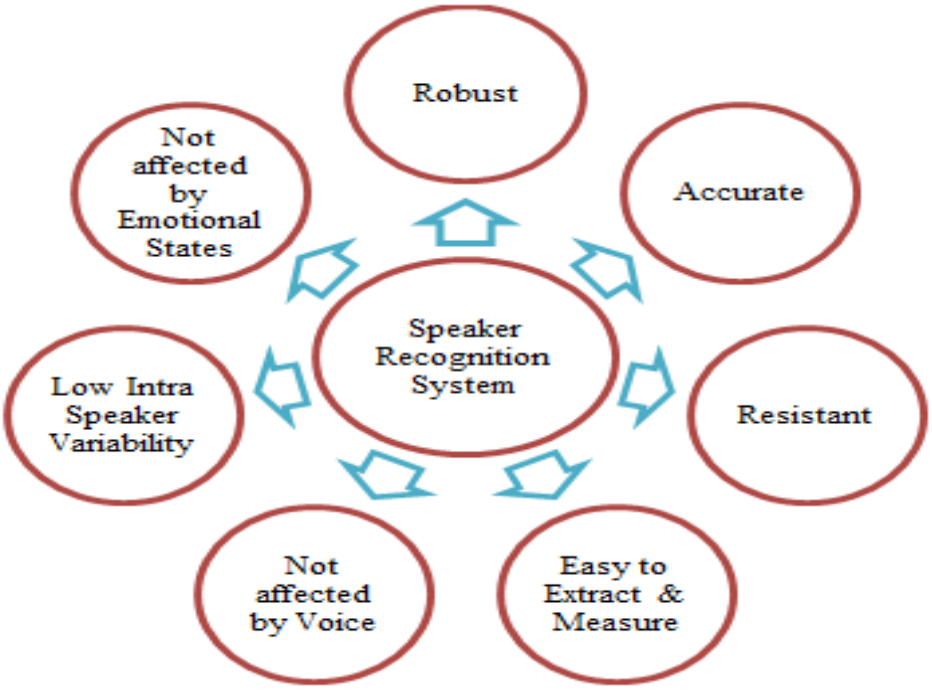


Figure-2.10: Characteristics of a Robust Speaker Recognition System

A strong speaker recognition mechanism must focuses on the factors on the basis of which a person recognizes the voice of other persons. If it can be known that how human recognizes the voice of any person (whom he/she communicate frequently) then this will help to make a robust speaker recognition system which is more and more accurate. Figure-2.10: shows the characteristics of speaker recognition system. A speaker recognition system should be robust, accurate, resistant, easy to extract and measure, not affected by voice, low intra-speaker variability and not affected by emotional states etc. [210].

2.11 Challenges of Speaker Recognition

In speaker recognition several sources of errors occur. These errors may be related to speaker or it may be technical errors [7] [238]. There are some challenges/errors such as: Intra-variation and Inter-variation, Voice disguise and mimicry, Mismatch condition (Channel, Recording, and Environment etc.), Background noise etc. are discussed here.

2.11.1 Intra-Variation and Inter-Variation

Variation in voice is the main factor which affecting the speaker recognition system performance, it is also known as session variability. It can be session variability or it is occurred due to health issue e.g. cold, depression, aging or some other physiological changes etc. [162-163]. Intra-variation problem occur individual such as pitch mismatch.

2.11.2 Channel Mismatch Condition

In the recognition mismatch condition is the more severe problem. Such type of errors comes into the category of technical errors for example recording device mismatch, recording quality mismatch background noise etc. it is happens naturally that a speaker speaks training utterances in clean environment while for testing it may be speaks in noisy condition e.g. street, market, office etc.[163-166]

2.11.3 Voice Disguise and Mimicry

Mimicry (imitation) is a kind of voice disguise, where people produce voice like another one. Whereas voice disguise is the case where one's voice changing intentionally. Generally voice disguise is used by fraudster or criminals. For example when a fraudster/criminal make a blackmail call may keep a cloth in his/her mouth or during investigation he/she changes their voice [167-169].

2.12 Suitability of Speaker Recognition System

How much speaker recognition useful and suitable as a biometric authentication purpose, is judge on the following criteria to evaluate the system suitability?

- **Uniqueness:** As it is discussed that voice is the combination of behavioural and physiological factors. Voice is unique for individual due to their glottal structure, so the voice has more unique feature than other biometric such as finger print [124] [174].
- **Universality**: It is not a universal solution for authentication, since for that people who are having problem with their voice or have severe illness this biometric solution is not useful [124].

- **Acceptability:** speaking is a natural process for communication of human so speaker recognition is well accepted in an unobtrusive manner, there is no need for any unusual task. So it is well accepted as compared to other biometric such as retina scans required direct contact and it is demonstrate pregnancy and other medical conditions such as high blood pressure [124] [186] [200].
- **Stability:** since speech is time varying, it changes with aging and is also determine by some factors such as stress, sickness, emotion, etc. [124] [186].
- **Performance:** Generally performance is measured by speed, accuracy and robustness. Robustness should be maintained by the authentication system. The speaker recognition system robust against channel mismatch and noise also since it is behavioural biometric, robust against misreading and mispronunciation [174] [186] [200].
- **Availability:** In this digital era where lots of electronic devices are available for communication, acquiring voice is a very easy task. The advantage of speaker recognition that it is cost effective that means it is not required any expensive hardware. It can be perform on the telephone or using other communication medium by recording someone's voice [124].
- **Circumvention:** spoofing is a major issue with speaker recognition. Since voice is recorded for training and testing phases, also voice recording for identification process. So it is ensure that risk of spoofing with voice recording can be mitigated. So an imitator cannot anticipate the random phrase that will be required for identification. System must be not attempt a playback spoofing attack [40] [57] [124].

2.13 Speech Feature Extraction

Speaker recognition is speech dependent system. Speech signal analyzed to obtain less variability and more discriminative features by converting speech signal to parametric values. To extract these parametric values different methods are used called features of speech signal. Features obtained from speech signal are used to create speaker models which contained different information of speech. There are different techniques are available for feature extraction such as LPC, LPCC, MFCC, Prosodic features of speech etc. Every method used for specific feature extraction

from speech signal such as MFCC and PLP methods used to extract spectral features of speech signal, and formant, intensity, pitch and harmonicity are represent to prosodic features of speech [126]. Feature extraction is the process of finding suitable voice features for speaker identification. A feature extraction technique has the following properties [127]:

- Easy to measure
- Robust against noise and channel distortion
- Robust against mimicry and disguise
- Low intra-speaker variation
- High inter-speaker variable

To convert raw speech into a categorization of acoustic features, an audio processing technique is used. Acoustic features are contains characteristics information about the speech signal. This process is called preprocessing (feature extraction) also called front-end in the literature. Commonly used acoustic features are Mel Frequency Cepstral Coefficients (MFCC), Linear

Prediction Cepstral Coefficients (LPCC), and Perceptual Linear Prediction Cepstral Coefficients (PLPC) etc. these acoustic features are founded by the spectral information of a speech signal which is a short windowed segment of a speech. MFCC features are more commonly used to develop many speaker recognition systems such as most of the NIST (in 1998) used MFCC speech features while some use LPCC speech features [5] [159]. However from last one decade researcher gave attention to prosodic information for developing speaker recognition systems. There are many reasons to using prosodic features such as it carrying speaker's speaking style and intonation related information. In addition prosodic features such as pitch and duration (phoneme and syllable) are more robust against channel effects as compared to cepstral features [89] [145]. Following are some feature extraction techniques:

- **Linear Predictive Coding (LPC)**

LPC is an encoding method for speech processing, based on the linear predictive model. Linear predictive model are estimated as a linear function. LPC is used to

evaluate specific components of a frequency spectrum of speech signal, called formants [28].

- **Mel-frequency Cepstrum Coefficient (MFCC)**

MFCC features are most commonly used speech features in speech and speaker recognition both. Drawback of such type of features that it is highly affected by noise and not able to capture some information of speech e.g. tones. So that such type of speech features is not suitable for recognition when data is noisy [7] [126].

- **Prosodic**

ASR has three common steps. These steps include data acquisition, feature extraction and modeling techniques. To extract features from speech signal, many feature extraction techniques are available such as MFCC, LPC, LPCC, Prosodic etc. and for modeling HMM, GMM, UBM etc. techniques are available [17-18] [35].

Automatic Speaker Recognition (ASR) is a procedure to recognize a person using a machine by his/her spoken words/sentence. This technology is useful to maintain security in various fields such as crime investigation, access control, and voice based banking and authenticity etc. To develop ASR system, extraction of characteristic of speech signal is one of the core tasks. The primary work of feature extraction technique is to extract voice characteristics from the speech signal. These voice characteristics are unique to each and every person to be used to distinguish them [129,130]. Prosodic features for speech signal are expressed in terms of stress (the relative prominence of a syllable or musical note (especially with regard to stress or pitch), rhythm (recurring at regular intervals) and intonation (rise and fall of the voice pitch). These convey required information to identify the spoken language. Figure- 2.13 shows the proposed system for speaker recognition using prosodic features.

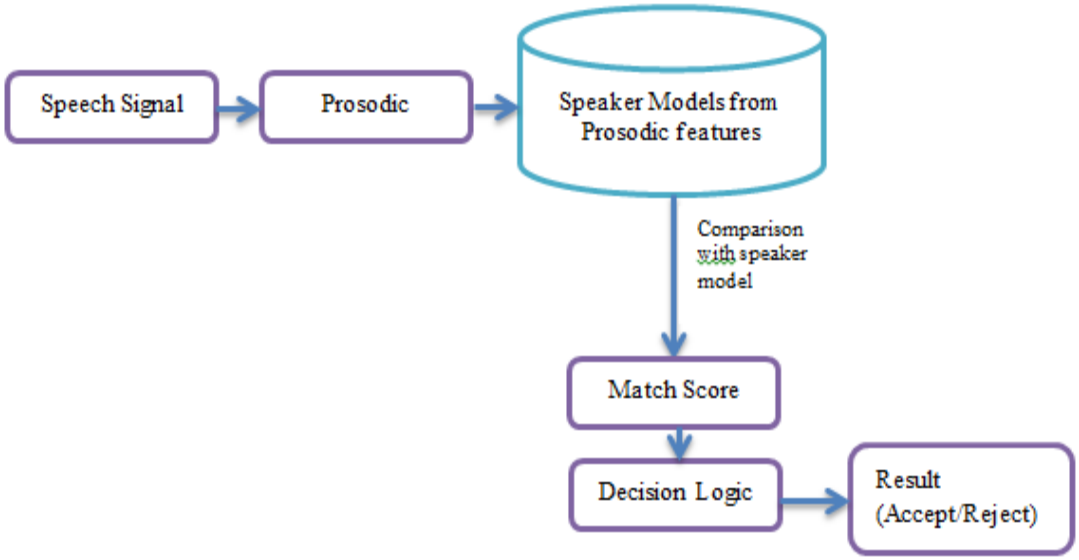


Figure- 2.13: Proposed System for Speaker Recognition

Speech features used for speaker recognition may be spectral (cepstral) features, phonetic features and prosodic features [131-133].Traditional speaker recognition systems depend on 'spectral features' which are extracted from very short segment of speech signal. This technique is adequate for clean data but the system performance degrades if the data is noisy or there is handset variability [134]. Unfortunately this technique is not suitable to extract long-range speech features related to person's speaking behaviour such as prosodic, lexical and discourse related habits. The purpose for using such long-range features in speaker recognition system is to increase system performance as compared to system using cepstral/spectral features [135].

As authors said in [136-137] that it has been found that by using long-range speech features, the system performance has been improved. Another advantage of such type of technique is to use long-range speech features which replicate person's behavioral characteristics of speech features. Such type of features could have potential for recognizing speakers as well as recognizing characteristics of the speech, for example speaking rate, speaking style, pitch etc. The main purpose of research on long-range speech features is to understand speaking behaviour [138]. It is assumed that prosody is associated to linguistic component of voice such as syllables and it is noticeable that changes occur in measurable parameters for example fundamental frequency F_0, energy and duration of speech [139].

2.14 Available Approaches for Modeling and Classification

Speaker modeling is one of the important components of speaker recognition. For every registered speaker, speaker models are created during training and testing phases after the computation of speech feature vectors. Training phase involves creating voice database for registered speakers whereas in testing phase, claimed voice input is matched with the training database [202]. During matching (classification), received speech (either known or unknown) is compared with the speaker model to evaluate a score value (match score). By using this score value, it is decided whether the speaker is accepted or rejected [124] [165]. Table- 2.14 shows the comparative study of different modeling techniques on the basis of different parameters.

Speaker Modeling Methods	**Comparison Based on Characteristics**				
	Text-independent/ Text-dependent	**Robustness**	**Purpose/Used For**	**Model Type**	**Approach used**
GMM	Text-independent and text-dependent	• Not affected with time variability. • Robust against Noise [25] [62] [67].	• Useful for classification • Reduced computing (posterior probability) complexity. • Speaker Recognition. • Gives high recognition accuracy [25] [62] [67].	Stochastic models [25]	Generative classifiers [25].
HMM	Text-independent and text-dependent both	Robustness against utterance variations [25] [62] [67].	• Acoustic feature space. • Multiple-state ergodic HMM [25] [62] [67].	Stochastic models [25].	Generative classifiers [25].
SVM	Text-independent and text-dependent	Most robust classifiers in Speaker Verification. [25] [62] [67].	• Not useful for classification • Try to minimize the classification error on a set of training data. • Speaker recognition and pattern classification [25] [62] [67].	Classifier Method [25] [62] [67].	Discriminative approach [25].
VQ	Text-independent and text-dependent	Affected by Time variability [25] [62] [67].	• Reduces storage requirements. • Speaker Verification [25] [62] [67].	Template models [25].	Clustering methods [25].

ANN	Text-independent and text-dependent.	Less robust classifiers [25] [62] [67].	• Used for classification methods. • Speaker recognition [25] [62] [67].	Classical pattern [25] [62] [67].	Discriminative approach [25].
DTW	Text-independent and text-dependent	Use for non-uniformity problem [25] [62] [67].	Resolve the matching problem [25] [62] [67].	Template Model [62].	Dynamic programming [25] [62] [67].

Table- 2.14: Comparative Study of Different Modeling Techniques on the Basis of Different Parameters.

Speaker models can be categorized as stochastic and template models also known as generative models and discriminative models respectively. In stochastic modeling, speaker models are created by using probability density function. During training phase, probability density function parameters are estimated from the given speech. And for matching a likelihood of the utterance is evaluated. Whereas in template modeling, training and testing models are directly compared with each other and the quantity of falsification between these two is the degree of similarity [25] [62] [67] [170]. A short description of selected modeling techniques is given in the following subsections.

- ***Support Vector Machine (SVM):*** It is a binary discriminative classifier. SVM uses boundary between two classes for creation of speaker models. One class contains training data vectors for target speaker's which are labeled +1 while the other class contains imposters training data vectors from a large data set and labeled as -1 [25] [69].
- ***Hidden Markov Model (HMM):*** It is the most popular modeling methodology for text-independent and text-dependent speaker recognition. HMM, a doubly stochastic process was developed in 1980s. The term hidden is use because it has an underlying stochastic process that is not observable. To observe the hidden process another stochastic process are used [88] [62] [171].
- ***Gaussian Mixture Model (GMM):*** It is comparatively more suitable method for speaker modeling. For creation of speaker model it uses speech features as a linear combination of finite mixture of multivariate Gaussian components. It uses Expectation-Maximization (EM) algorithm and maximum likelihood (ML) estimation for estimation of GMM parameters [25] [156-157].

- ***Vector quantization (VQ):*** It is a classification method for speaker verification. It uses clustering methods e.g. K-means use to reduce training vector. Each cluster is represented by code vector and this code vector is the centroid of that cluster. Collection of centroid vectors is called codebook. During verification process the training data of a registered speaker is used to create a codebook which is the model for specific speaker. If the provided speech is of an unknown speaker then the matching is determined by evaluating distance between the testing data feature vector and the target speaker nearest vector codebook. The evaluated distance is called score value of a verified speaker [237] [257-258].
- ***Artificial Neural Networks (ANNs)*****:** It is a discriminative methodology used for speaker classification. There are several types of neural networks, for speaker recognition generally Multi-Layers Perception (MLP) is used. MLP is a feed-forward network. In this network, multiple layers of nodes are achieved and collectively these are used for complex machine learning task. For each node, weighted sum are calculated for the inputs. Here, weights are adjustable parameters. After that transfer function is applied for calculating output of that node. Back propagation algorithm is use for determining the weight parameters [79] [151].

2.14.1 Characteristics of GMM

The Gaussian mixture speaker model was introduced in 1990 by Rose and Reynolds [132]. Then Gaussian Mixture Modeling (GMM) technique for text-independent speaker recognition introduced by Reynolds in 1992 came into picture [26]. The specialty of GMM is that any distribution can be modeled by using Gaussian mixture modeling technique. The reason behind is that it is able to provide large number of mixture components of voice. This modeling method is useful in text-independent speaker identification as well as speaker verification [67]. The GMM provides high recognition accuracy and effective speaker representations which is also computationally inexpensive [149].

GMM has widely used for speaker modeling in text-independent speaker recognition system. It has the following characteristics which make it more useful for modeling [4] [26] [149], such as:

- GMM has the ability to form smooth approximations to arbitrarily shaped density. It is based on a linear combination of Gaussian basis functions which are capable of representing a large class of arbitrary densities.
- In GMM, for each Gaussian component, an implicit realization of probabilistic modeling of speaker dependent acoustic classes to a broad acoustic class such as vowels, nasals and fricatives etc. is used.
- GMM is not susceptible to natural changes such as aging or cold.

2.14.2 Gaussian Components

The task of speaker recognition is done by using individual mixture components i.e. Gaussian components [26] [149]. For speaker recognition, features are obtained from the speech signal. The fundamental information of speaker discrimination can be characterized by Gaussians. For the expected GMM parameters i.e. covariance and Gaussian component, weight is associated to the location of formant, magnitude of speech signal and bandwidth of speech signal [132]. As discussed in [132], for good quality system performance at least 8 to 16 Gaussian components are mandatory where voice/speech is considered as noiseless. The GMM is created by using these components through diagonal covariance matrix. To build multi conditional robust systems, the minimum number of essential Gaussian components involving 64 and 128 are required. The main advantage of GMM involves its likelihood function being computationally inexpensive. GMM is collected from a finite mixture of Gaussian components. Since Gaussian components have potential to characterize discriminative information of speaker, it is widely used for speaker recognition [26] [149].

2.14.3 Gaussian Mixture Model and Speaker Recognition

GMM takes sequence of vectors provided by feature extraction technique and use it to create speaker model. These models are called Gaussian Mixture Models. The Gaussian mixture model is ‘mixture density’, categorized as a sum of M Gaussian component densities. Component density is a product of ‘mixture weight’ with a ‘Gaussian component’. Individual ‘Gaussian component’ represent acoustic classes and these classes reflect speaker specific vocal tract information therefore is useful for modeling speaker identity [4] [26] [149].

The GMM is used to represent speaker's model in the speaker recognition systems. The distribution of feature vectors extracted from a speaker's speech signals is modeled by Gaussian mixture density function [149] [132]. Equation 2.13.3 (a) for a D-dimensional feature vector denoted as x, the mixture density function P for speaker s is:

$$P\ (x|\lambda s) = \sum_{i=1}^{M} p_i^s b_i^s\ (x) \qquad 2.13.3\ (a)$$

In equation 2.13.3 (b) the density is a weighted linear combination of M component Gaussian densities, $b_i^s(x)$ each parameterized by a mean vector, μ_i^s and covariance matrix, $\sum_i^s$

$$b_i^s(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i^s|^{1/2}} \times \exp\{-\tfrac{1}{2}(x-\mu_i^s)'(\Sigma_i^s)^{-1}(x-\mu_i^s)\} \qquad 2.13.3\ (b)$$

The mixture weights are p_i^s and is represented as

$$\sum_{i=1}^{M} p_i^s = 1$$

And $\lambda_s = \{p_i^s, \mu_i^s, \Sigma_i^s\},\ i = 1, \ldots, M.$

For an input data X and a number of mixtures M (assume a priori), data can be fit using M Gaussian distributions. The figure- 2.14.3 (a) shows the Gaussian component of a speech signal and figure- 2.14.3 (b) represents the process of computing the probability of a feature vector given a GMM model [132].

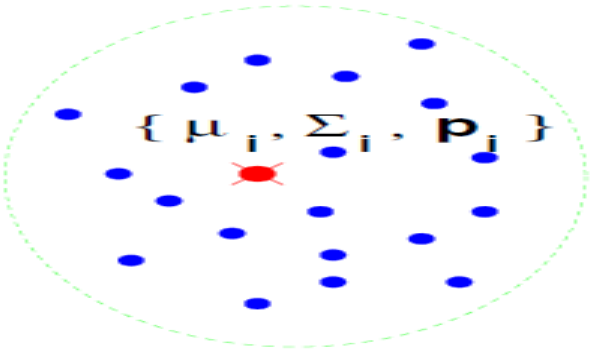


Figure- 2.14.3(a): One Component of a GMM Speaker Model [7]

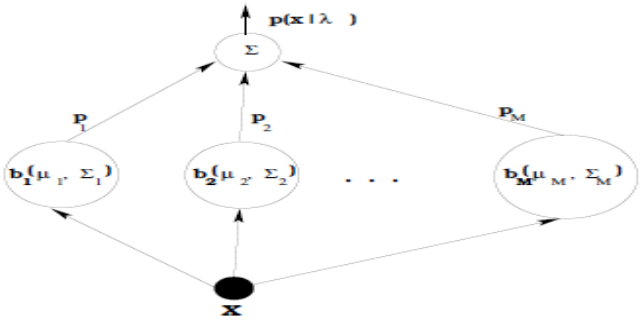


Figure- 2.14.3 (b): Process of Computing the Probability of a Feature Vector given a GMM Model

2.15 Performance Measurements of Biometric Systems

The performance of speaker recognition system or biometric systems depends on the following factors [158] [186]:

- False Acceptance Rate (FAR)
- False Rejection Rate (FRR)
- Relative operating characteristic (ROC)
- Equal Error Rate
- Template Capacity

➢ ***False Acceptance Rate (FAR)***

It is the probability that the recognition system matches incorrectly (input pattern) with the non-matching template in the speaker database, i.e. it measures the percent of imposter recognition acceptance [172-174].

- ***False Rejection Rate (FRR)***

It is the probability that the recognition system misses the correct input pattern template which exists in the speaker database, i.e. it measures the percent of true speaker which are falsely rejected [172-174].

- ***Relative Operating Characteristic (ROC)***

It is a graphical classification between the FAR and FRR. Speaker recognition system performs on the basis of the result (accepted/rejected) of matching algorithm. Matching algorithm gives results on a threshold; threshold determines the recognition value close to a template stored in the database. Higher threshold reduces the FAR but increases the FRR, while when threshold is reduced then lower false non-match are found and more imposters are accepted [172-174].

- ***Equal Error Rate***

It is the rate where FAR and FRR both are equal. EER is a prompt method to compare the accuracy of any system with the help of different ROC curve. The lowest EER shows more accurate system [4] [131].

- ***Template capacity***

Template capacity of any system is defined as the maximum number of sets of data which can be stored in the system [184]. Speaker recognition performance is usually calculated by detection error trade-off (DET) curves, equal error rate, and a weighted cost value. To be more accurate, a speaker recognition system should have lowest EER. It also has high storage capacity [176-178] [184].

2.16 Performance Dependent Task

Performance of speaker recognition system is highly affected by the following factor [58]:

- Background and environmental noise
- Session gap between training and testing data

- Channel mismatch between training and test set
- Speaker variability (speaking rate, cold, speaking level etc.)
- Short time changes (speaking style)

2.17 Conclusion

In this chapter, basic concepts of automatic speaker recognition systems, modeling technique etc. has been discussed. Speaker recognition is method of designing a system for identity of an individual based on his/her voice. Speaker recognition has a significant potential as it is convenient biometric method for security. The Speaker Recognition task is normally achieved by acquiring speech signal, feature extraction, modeling speech features for speaker, pattern matching and obtaining match score.

Till date a lot of work has been done in automatic speaker recognition field, but still many realistic problems need to be solved. Researcher has tried to examine areas of possible improvements in the field of speaker recognition. As it is clear from the literature, that the prosodic features are more robust against noise. Therefore main goal of the research is to improve the performance of speaker recognition system by modeling prosodic features. In addition, the researcher has proposed a framework for speaker recognition system.

Chapter- 3
Framework for Speaker Recognition

CHAPTER-3: FRAMEWORK FOR SPEAKER RECOGNITION

The choice of a point of view is the initial act of culture.
—José Ortega y Gasset

3.1 Introduction

Human voice or speech signal encloses with rich information about the individual such as speaker emotion, speaker identity, language, message content, speaker temperament etc. speech processing has the following task for example speech analysis, synthesis, coding and recognition etc. further it is classified as speech recognition, language recognition and speaker recognition etc. [19] [160]. Speaker recognition is the process of extracting voice features for personal identity by the analysis of speech utterances. It is a biometric technology which is used in many security areas for secure access control and forensic investigation. In today's digital era where insecurity is everywhere, speaker recognition technologies provide a secure solution in our daily life. ASR systems provides many services e.g. voice based banking, voice database access, voicemail, remote access to personal computers, voice based access control devices, and many other authentication areas [19][186].

In past few years, speaker recognition technology has many significant developments which is now used in several authentication applications such as physical and logical access control systems [188]. Also Current scenario shows that speaker recognition is a growing research area in speech signal processing [19]. Speaker recognition/voice recognition is a process of identifying of an individual on the basis of his/her voice. Voice has the characteristics of both physiological and a behavioral biometric features. There is a difference between speaker recognition and speech recognition. Speaker recognition, recognizing who is speaking while speech recognition recognizing what is being said [177].The current speaker recognition (text-independent system) is language independent but their performance affected in multilingual trial condition [189]. Prosodic features are robust against technical mismatch hence system performance improved by using prosodic features [190].

Speaker recognition has the main task are speech feature extraction or front-end processing and modeling. Feature extraction is the process of selecting required speech features that is further used for speaker modeling. Several speech features has been proposed to till date, each feature have their advantages and disadvantages. These speech features are used for different type of speech processing e.g. speech recognition, language identification and speaker recognition. Process of developing speaker recognition system has two phase, the training i.e. enrollment phase and the testing phase [182]. During training phase, speech samples are collected and system is trained by the collected speech samples. Whereas in the testing phase, provided speech sample is matched by system for identification or verification of speaker. Speaker recognition categorize in two main task speaker identification and speaker verification. Further it is divided into text-dependent and text-independent speaker recognition system. The recognition is said to be text-dependent when the speaker use same linguistic for both i.e. training and testing else recognition is text-independent. Furthermore system is either open-set or closed set, if the system has the chances to reject the speaker who is not enrolled in the speaker database else it is closed-set database system [176] [191].

3.2 Background

Human speech is a natural way of communication with each other. It is a medium by which human express their emotions, thoughts and share messages. Speech is a complex signal and it contains several information of the speaker's and language. For example excitation source information, vocal tract system (while producing voice), linguistic information, emotional states of speaker, supra-segmental information (prosodic features e.g. pitch and energy). Human speech is unique for individual due to differences in shape of vocal cord, size of larynx and other voice production organs [171]. Now a days voice based authentication technology has grown up quickly and it is used for authenticating of individuals. This technology can be used but not limited to crime investigation, forensic, personal authentication, voice based system etc. [9] [74] [44] [192].

In recent digital era where everything is going to be digitalized, human authentication is also done by machines. It is fact that human being is able to easily

distinguish among voices of different persons. To become a recognizer like human the machine should be robust and reliable. Speaker recognition, speech recognition and language identification are the most commonly used authentication processes. Speaker recognition is emergent area in speech signal processing. It is concerned to the identity of a person, based on his/her voice characteristics. Speaker recognition has many applications in distinguished areas such as personal authentication, forensic, security check in military etc. [77] [194]. For example, in digital forensic through voice, a suspected person can be recognized by tapped telephone conversation of criminals/terrorists.

One of the popular areas of speaker recognition is authentication.Process of automatic speaker recognition is based on acquiring speech signal; creating speaker models which are used to compare with available models [195]. In general, speaker recognition is sub divided into speaker verification and speaker identification. Speaker verification is used to confirm (accept/reject) to an identity claimed by a speaker. For example it is useful in case of access control where voice is used as biometric feature. In speaker identification, a speaker is selected from known speaker's set for which speech sample (speaker model) is previously available. This system may also be able to take decision whether the acquired new speech signal matches with the existing stored speech models or this is an unknown speaker [196].

Though, there are numerous researches going on in the area. To development of a robust and accurate speaker identification system is still a big challenge. Lots of efforts have been made to improve the recognition system performance [7] [197-200] but the progress still needs improvement. The framework propose by the researcher is a generic framework for speaker recognition system. It includes complete procedure to design a speaker recognition system. It provides with several choices for its implementation it is implementer's choice to select medium through which speakers signal acquired, to decide on the size of segment of speech signal; to choose the suitable feature extraction technique; to select modeling technique and to select technique for matching score. The framework is applicable for both recognition system i.e. text-dependent/text-independent automatic speaker recognition system.

The framework provides a methodology for speaker recognition. During enrolment speech signal is acquired for each speaker to extract speech features. From

speech features equal number of speaker models is created for every registered candidate to create voice training database. Recognition process is done by matching the utterance with each registered speaker's models/template. Speaker recognition system selects the template whose match score matches most closely to the model available in training database. The framework integrates the whole process involved in the speaker recognition system. The purpose of the proposed framework is to identify individual's long utterance of the speech with the help of prosodic statistics. The framework provides both static and dynamic characteristics for creating speaker model.

3.3 The Framework

A framework can be defined as, the structure (real or theoretical) supposed to help as a guide for developing software or hardware or anything that uses it to produce something valuable [194].The proposed framework for speaker recognition system is universal in nature i.e. it can be used by anyone to design a recognition system.

3.3.1 Premises

A framework is a structured or a logical way to organize a process to achieve anything [194]. It is reusable set of component which is used to manage system. As every proposal may possess its own premises the framework for improving the performance of speaker recognition system has the following assumptions:

- The framework is designed for text-dependent/text-independent speaker recognition system.
- One can choose a subset of the available techniques and methodology for the development of recognition system.
- To make a robust and accurate system one should keep into account certain things like channel mismatch, background noise, recording conditions etc.
- The framework is not explicitly discussing about noise removal/addition. As it is assumed that it is the part of feature extraction process.

In order to make system more robust one should try to select the speech features which are more robust against noise and have useful information related to

speaker. Also a modelling technique which is suitable according to a particular application may be chosen to create speakers model for training and testing voice database.

3.3.2 Guidelines

The aim of the framework is to identify as well as verify speakers. To fulfil the aim, the proposed framework for speaker recognition system has the following phases. These are namely:

- Sample collection and preparation
- Feature extraction
- Model creation
- Feature Matching
- Decision
- Performance evaluation

In the first phase i.e. during sample collection, speaker's voice is collected through the available communication mediums. In the next phase i.e. during sample preparation collected voice samples are broken into small pieces for further process. Further, selection of speech features, feature selection method, modeling technique and matching method is performed. During model creation, speaker voice models for training and testing purpose are created using modeling technique selected in previous phase. Matching is performed by comparing a voice sample with the sample in the database. On the basis of match score it is decided that identity is found or not.

3.3.3 Framework Development

The goal of developing the speaker recognition framework is to recognizing a speaker either in closed-set or open-set. It is supposed that the speech segment is taken by either known or unknown speaker. Proposed framework is shown in the figure-3.3.3(a). All the phases in the speaker recognition have been discussed in detail in the following subsections:

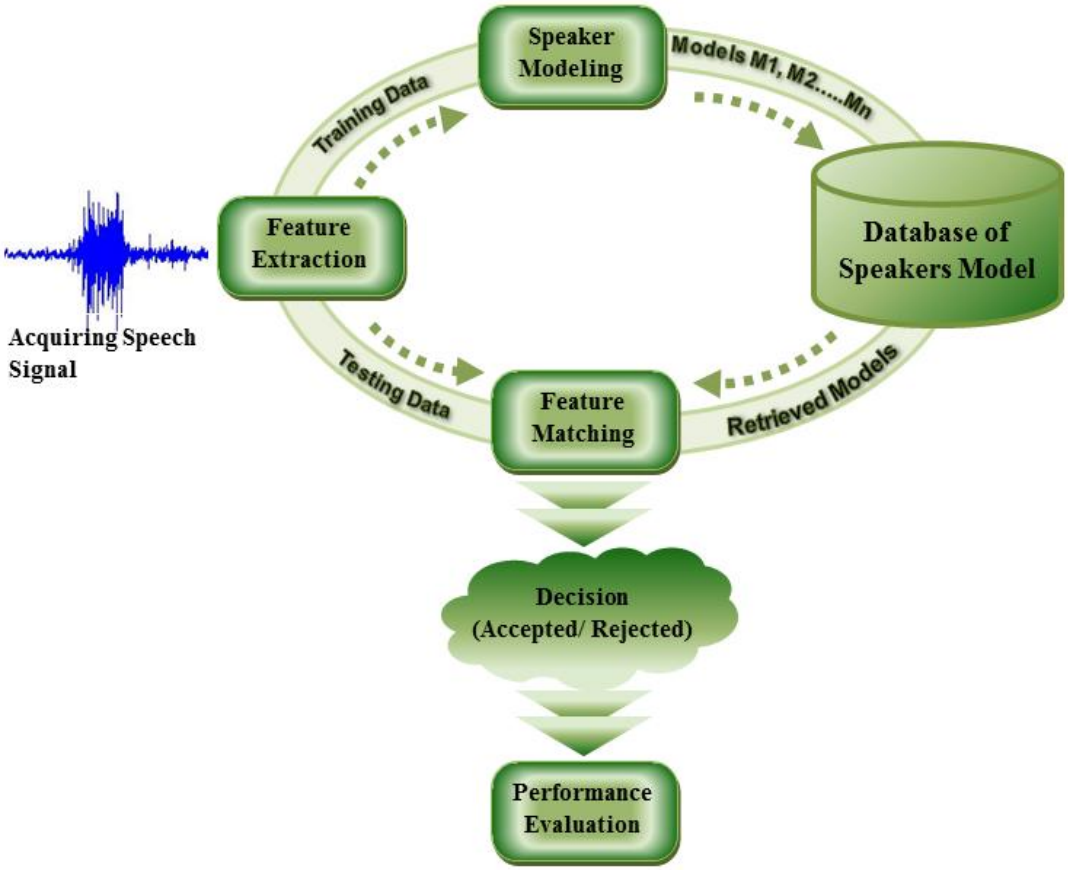


Figure-3.3.3 (a): Framework for Development of Speaker Recognition System

(i) **Acquiring Speech Signal**

The first phase of the framework is acquiring speech signal for sample preparation. The first step of design and development of speaker recognition system is acquiring speaker's voice. There is various ways for speech signal acquisition. For example telephonic conversation, voice recording by using headphones in lab, acquiring voice in noisy environment, old recorded voice (CD, magnetic tape etc.). During acquisition of speech signal, sound wave is transformed into a digital signal for processing. A microphone, headphone/telephone or any other voice recording device can be used for converting acoustic wave into an analog signal. The recorded voice is used for authentication of any speaker [203]. After acquiring speech signal, framing and windowing is performed on the voice samples.

- ***Framing/Windowing***

After the acquisition of speech signal framing is performed. During frame blocking the signal splits into equal frames of length N. After framing windowing is performed. There are many types of windowing such as triangular windowing, rectangular windowing Bartlett, Blackman, Hamming, Hanning, Kaiser, Lanczos and Tukey window functions etc. The simplest is rectangular window (no windowing). This window has no discontinuity at beginning and end of frame [7] [52]. Figure- 3.3.3(b) shows the frames and window in a speech signal.

Selection of speech frame is an important task and deciding frame length is an essential parameter for spectral analysis of a speech signal. Generally standard frame length 10-30 milliseconds are used for MFCC [52] [204]. Size of the window is related such that it should be large enough for adequate frequency resolution and short enough to capture the spectral properties.

Windowing of a speech signal is done to find out effect of spectral artifacts in the framing process [7] [205-206] [125]. For windowing several smoothing windows are used such as Rectangular (none), Hanning, Hamming, Blackman-Harris, Exact Blackman, Blackman, Flat Top etc. Hanning window is use for evaluating transients and its shape like a shape of a half cycle of a cosine wave. Modified version of hanning window is known as Hamming window. Shape of hamming window is like to a cosine wave [207]. In general Hamming window (it gives better spectral performance [208] is used for calculate window function of speech signal [201]. The Hamming window is defined as [7] [52].

$$w[n] = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{N}, & 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{cases}$$

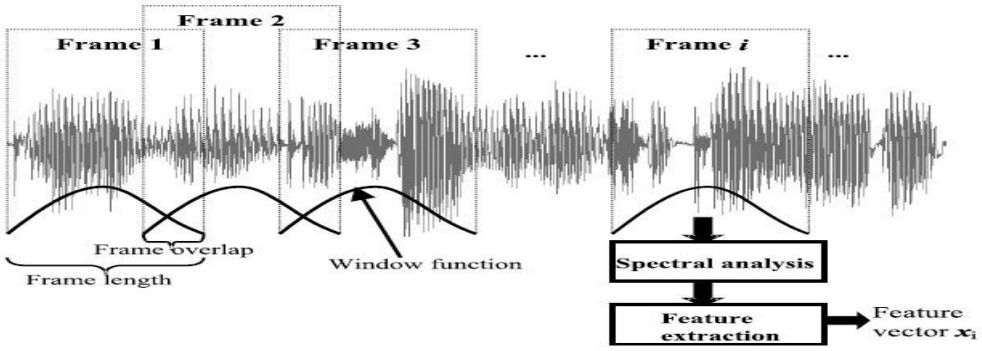


Figure- 3.3.3 (b): Frame & Window of a Speech Signal [7]

(ii) Speech Feature Extraction

Feature extraction is the process of converting a raw speech signal into a sequence of acoustic feature vectors which contain the characteristic information about speaker. The following suggestions must be taken into account while selecting speech features [30] [247]:

- Speech features should be resistant against voice and channel distortion.
- Speech features should not be affected by variation in voice (e.g. by speaker's health or aging)
- Feature extracted from speech signal should be easy to estimate.
- Speech features should be able to maintain high inter-speaker discrimination and less of intra-speaker variability.
- Feature extraction method should be difficult to mimic against speech of imposters.

The above mentioned characteristics of a speech feature extraction methodology are difficult to achieve in individual feature extraction technique. Since some features such as fundamental frequency (F0) is robust against noise but required long speech segments hence prosodic features are individually capable to build speaker recognition system [225] [231].

➢ **Selection of Speech Features**

To select appropriate speech features and methods to extract selected speech features is known as feature selection and feature extraction [203]. Feature extraction is the

main task of speaker recognition or speech recognition. It is well known that speech signal is a complex signal which contains several features of voice. To recognize a speaker it is necessary to extract speech features of the speaker. These features are categorized as physiological and behavioral speech features of the speaker [211]. Physiological features are such as hand geometry, finger print, iris, retina, and face, DNA etc. and behavioral features such as voice, Gait and typing rhythm etc. The next section discuss about the criteria of speech features selection [212].

- **Criteria for Speech Feature Selection:**

To develop a robust speaker recognition system there must be some specific criteria to select properties of speech signal after framing/windowing. To create a good system, the speech features selected should possess the following properties [213]:

- It should be robust against noise and distortion
- It should be occur naturally and frequently
- It should not be affected by speakers health or age
- It should be difficult to mimic
- It should be Not affected by speaker variability
- It should not affected by channel mismatch

A single speech feature has not fulfilled the entire above mentioned prerequisite. Therefore selection of speech features depends on the application of authentication such as security level, environmental noise, size of database, type of speakers (co-operative/non co-operative) etc. For example spectral features are extremely discriminative, they calculated from very short segments of speech signal (1-5 sec) but easily affected by noise. F_0 statistics require large amount of speech data but robust against channel and noise (technical) mismatches [213].

- **Analysis and Categorization of Speech Features**

A speaker recognition system can be designed by using one or more (combination) of the speech features. The selection of features will depend on the requirement of the system. For example short-term spectral features are highly discriminative and they can be reliably measured from short segments (1-5 seconds) but these features are

easily affected by noise (when transmitted over a noisy channel) [211] [214]. Fundamental frequency (F_0) measurements are robust against channel mismatch but require long speech segments and are not discriminative. In addition, selection of speech features basically depends on the environment where the system is to be deployed such as co-operative/non co-operative speakers, security/convenience balance, database size, amount of environmental noise etc. There are many speech features available for speaker recognition. These features can be categorized as follows:

- Spectral features
- High-level features
- Supra-segmental / Prosodic features
- Source features
- Dynamic features etc.

Feature Type	**Examples**
Spectral features	MFCC, LPCC, LSF Long-term average spectrum (LTAS) Formant frequencies and bandwidths
High-level features	Idiosyncratic word usage Pronunciation
Supra-segmental/Prosodic features	F0 contours Intensity contours Micro-prosody
Source features	F0 mean Glottal pulse shape
Dynamic features	Delta features Modulation frequencies Vector autoregressive coefficients

Table -3.3.3: List out the Categories of the Speech Features along with their Examples

The type of authentication system will decide ‘which’ and ‘how many’ features are to be selected. Table- 3.3.3 shows the example of each feature type. Spectral

feature are in the form of short-term speech spectrum and describe physical characteristic of vocal tract. High-level features represent the symbolic information e.g. characteristic of word usage. Supra-segmental or prosodic features represent speaking rate, rhythm, intonation pattern and stress etc. Source features represent glottal voice source features. Dynamic features are related with time evolution of spectral features.

(iii) Speaker Modelling and Database Creation

Speaker modeling involves two phases training phase and testing phase. Speaker models are created by using specific speech feature. There are two types of model creation methods; stochastic models and template models. These modeling methods are used for constructing speaker models using the features extracted from the speech signal. In this phase, a speech model based on the extracted features of speech signal is created and stored. During authentication of a speaker, matching algorithm compares the models of the claimed user. In stochastic models, pattern matching is probabilistic and the result is measurement of likelihood, or conditional probability of the given observational model. The template method can be dependent or independent of time. VQ modeling is an example of time-independent template model. Time-dependent template model are more complicated because it must accommodate human speaking rate variability [52] [203] [237]. Stochastic models are more flexible and result is more reliable due to probabilistic likelihood score as compared to template models.

- **Characteristics of a Good Speaker Model**

A good speaker model is one which can rapidly able to adapt voice differences. During construction of speaker models, a number of design goals need to follow. It is very difficult to achieve these goals. However, by choosing a good speaker model these goals can be achieved. Following are the characteristics of a good speaker model [191] [215]. Figure- 3.3.3 (c): shows the characteristic of good speaker model

- **Consistent within speaker:** speaker model for a particular speaker should avoid speaker's voice inrta-variability [215]. It should be able to neglect differences occurred in the voice of same speaker over the time.

Figure- 3.3.3 (c): Characteristic of Good Speaker Model

(i) **Distinguish individual speakers:** Individual speaker should be represented distinctly.

(ii) **Perceptual significance:** those speakers voice that are nearby the suspected should be widely separated either similar or different while judged.

(iii) **Compactness:** Compactness should be achieved only if the models have low dimension. This allows the model and the application which is using it. bring together new speaker models covered by training speaker models.

(iv) **Text-independent:** It should be text-independent. There is no need to utter the same phrase or sentence during training and testing phase.

(v) **Rapidity of Formation:** models should be generated as rapidly as possible by using information from speech signal.

(vi) **Robust against noise:** Modeling techniques should be robust against noise. For a given speech signal the model should be free from noise.

(vii) **Thoroughness:** the model should contain all the required information about speaker to make conscientious decision.

The above are the characteristics of a good speaker model which can enable to achieve the goal of developing robust speaker recognition system. The aim is to develop general models of speakers that can be used successfully to a wide variety of applications in the area of speaker recognition [162] [215] [219].

After modelling, the speaker's models are stored in a database which can be referred while matching.

(iv) Feature Matching

Matching is the process of comparing the extracted speech features of a person with stored speaker models/templates. The comparison quantifies the similarity between the voice (record for identification) and a speaker model from voice database. Selection of a matching technique is an important task. There are many prevalent classifications / matching technique such as Hidden Markov Models (HMM), Vector Quantization (VQ) and Dynamic Time Warping (DTW) [4] [7] [216]. Speaker model is used for comparing with a particular input signal. Speaker models are stored in a database. Two kinds of database are there: training database and testing database. Model comparison method involves the following:

- ✓ Matches training database of the target speaker with his/her testing database.
- ✓ Match score is calculated.
- ✓ If the match score is greater than or equal to the threshold then the target speaker is accepted by the system otherwise rejected.

During matching created speaker models, may be speaker-dependent in case of speaker recognition system and speaker-independent in case of forensic speaker recognition system. There is predefined specific criterion for creating speaker models [152] [195].

(v) **Decision Phase**

The decision process depends on the kind of the system i.e. closed-set system or open-set system. In case of closed-set identification system, the decision can be made by selecting that model which is most similar to the test sample speech signal. In case of open-set, system requires a threshold to verify that similarity is valid. As there may be chances that a system rejects a registered speaker, hence cost of making an error is considered in the decision process. For example, in case of a bank to allow an imposter will prove to be more costly than to reject a true customer. The Decision is determined by particular matching and modelling algorithms. For example, in case of template matching decision is given by computed distance between speakers models whereas in stochastic matching calculated result is based on the computed probabilities [4] [19] [26] [253]. Figure- 3.3.3 (d) shows the decision process of a speaker identification system.

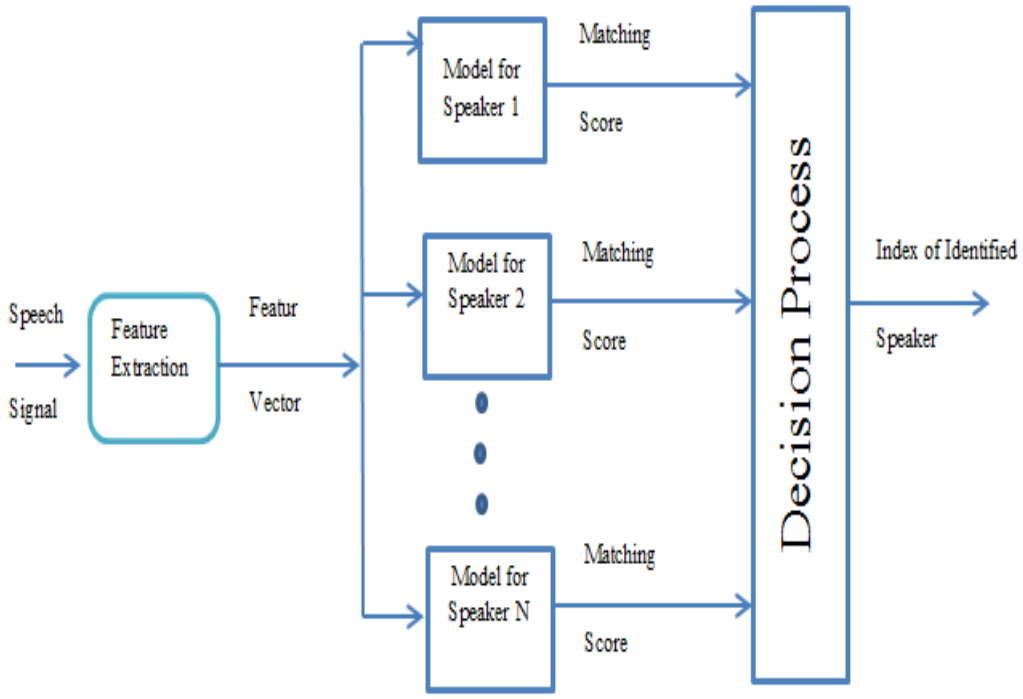


Figure-3.3.3 (d): Decision Process of Speaker Identification

3.3.4 Performance Evaluation Phase

The framework also provides the way to measure performance of the speaker recognition system. For the purpose, various available metrics can be used. The commonly used metrics are the False Acceptance Rate (FAR) and False Rejection

Rate (FRR). For making speaker recognition system more accurate, FAR is should be minimum [7] [220]. In addition, performance of speaker identification system is also decided by Equal Error Rate (EER). It is the most common method used to evaluate system performance. EER is a point where probability of False Acceptance (FA) and probability of False Rejection (FR) both are equal [4].

3.4 Framework Significance

The proposed framework has the following significance:

1. The proposed framework will help to develop a speaker recognition system using speech signal.
2. The proposed framework is applicable for both; text-dependent or text-independent speaker recognition system.
3. It gives the flexibility to choose any type of speech features for developing speaker recognition system.
4. The proposed framework is independent from the specific modelling technique.
5. During pattern matching there is no compulsion for selecting a specific pattern matching algorithm or technique.
6. The proposed framework is universal framework which includes all the necessary steps which is required to develop a speaker identification/ recognition system.
7. In every step, the proposed framework facilitates the user with the option to go back to the previous phase, if required.

3.5 Limitation of Developed Framework

The proposed framework, however, has been designed with the following limitations:

- Though it provides step by step solution to develop a speaker recognition system but it may not include exhaustive steps to be performed at any phase.
- In order to improve performance, the proposed framework requires large voice data is required and a broad training has to be performed in advance.

3.6 Conclusion

In this chapter authors proposed a framework for development of a robust and more accurate speaker recognition system. The proposed framework provides a method for speaker recognition. Speaker enrollment is the first step towards creating speaker recognition system, in this each registered candidate provides a set of utterances. For training, equal number of speech templates is created for each speaker (for individual speaker templates need not be of same duration) but for testing it may be vary. Speaker's template set is used as a model for the individual. During matching process, system selects the template whose match score is more closed to the test template. The proposed framework explains the overall process involve in the development of speaker recognition system. The proposed framework has a credential for long-term statistics. In the framework either static or dynamic speaker characteristics of speech features are used for speaker recognition. The proposed framework has the potential to additional embodiment and modification would be possible as per the requirement.

Chapter- 4
Implementation of the Proposed Framework Using Prosodic Features

CHAPTER-4: IMPLEMENTATION OF THE PROPOSED FRAMEWORK USING PROSODIC FEATURES

"Emotion and feelings may not be intruders in the bastion of reason at all: they may be enmeshed in its networks, for worse and for better."

- Antonio R. Damasio

4.1 Introduction

Implementation is the process of defining a method, model, design, standard steps for doing something (that is how a specific system is build). It also states the basic tasks those should be performed to happen something actually. The proposed framework for speaker recognition has some prescriptive steps in each of the individual stage. These steps are speech signal acquisition, feature extraction, speaker modeling, matching, decision phase and performance evaluation phase. Implementation of the framework is important to justify the usefulness and effectiveness of the proposed framework. Implementation of the framework involves developing a systematic and well planned method for recognizing a speaker through his voice by following the guidelines given by the proposed framework.

Speaker's voice features must be stored in a database for authentication process of a particular speaker. Hence, a database of the speech of enrolled speakers is created. Now, implementation of the proposed framework has been performed step by step. Implementation of the first phase i.e. acquisition phase is done by acquiring speech signal of a particular speaker. After acquiring speech signal, speech sample is prepared. During sample preparation, speech signal are broken down in small speech frames and windows for extracting speech features. In the second phase, prosodic features are extracted from the acquired speech signal. The reason behind implementation of this phase by extracting prosodic features is that these features enhance system performance [7] [40].

After feature extraction of a speech signal, speaker's models are created and stored. Modeling phase of the framework is implemented by creating Gaussian Mixture Model of the extracted speech features. During matching phase, the stored

Models are compared with the voice models of the speaker in question. Match score is calculated by using K-means algorithm for individual speaker. According to the match score, the decision is made whether the speaker is accepted or rejected. Lastly, system performance has been calculated by using Equal Error rate.

4.2 Implementation of the Proposed Framework

The framework for development of speaker recognition system proposed in chapter 3 is implemented in this chapter. It is divided into five phases namely speech signals acquisition, feature extraction, speaker modeling, matching and decision process. Each of the phases is reviewed and revised accordingly. In speech acquisition, speech signal is acquired for training and testing purpose. After acquiring speech signal features are extracted to create speaker models for training and testing. After creating models, matching is performed by testing speech signal and calculating match score. On the basis of match score result, decision is taken that whether the speaker is accepted or rejected. The last phase includes the evaluation of system performance.

Various automatic speaker recognition systems are developed in different input speech condition and by using different features extraction and modeling techniques. But the major limitations of their work are that many of them are tested for only specific language (e. g. native language). In almost all the approaches, voice is recorded in lab condition to find out clean speech which makes them less useful in real conditions as it is almost sure that the testing data cannot be noise less in real life situations. In addition, various approaches have used many number of prosodic speech features to devise the recognition system which makes these system complex.

The researcher has tried to overcome all the limitations of the approaches discussed above. The framework for devising speaker recognition system is implemented using comparative less number of prosodic features; testing data is collected in normal conditions with background noise; the system can be easily used to identify Hindi (mother tongue) as well as English. Following is the step by step implementation of the proposed framework. Figure- 4.2 shows the phases involved in the proposed framework for development of speaker recognition system.

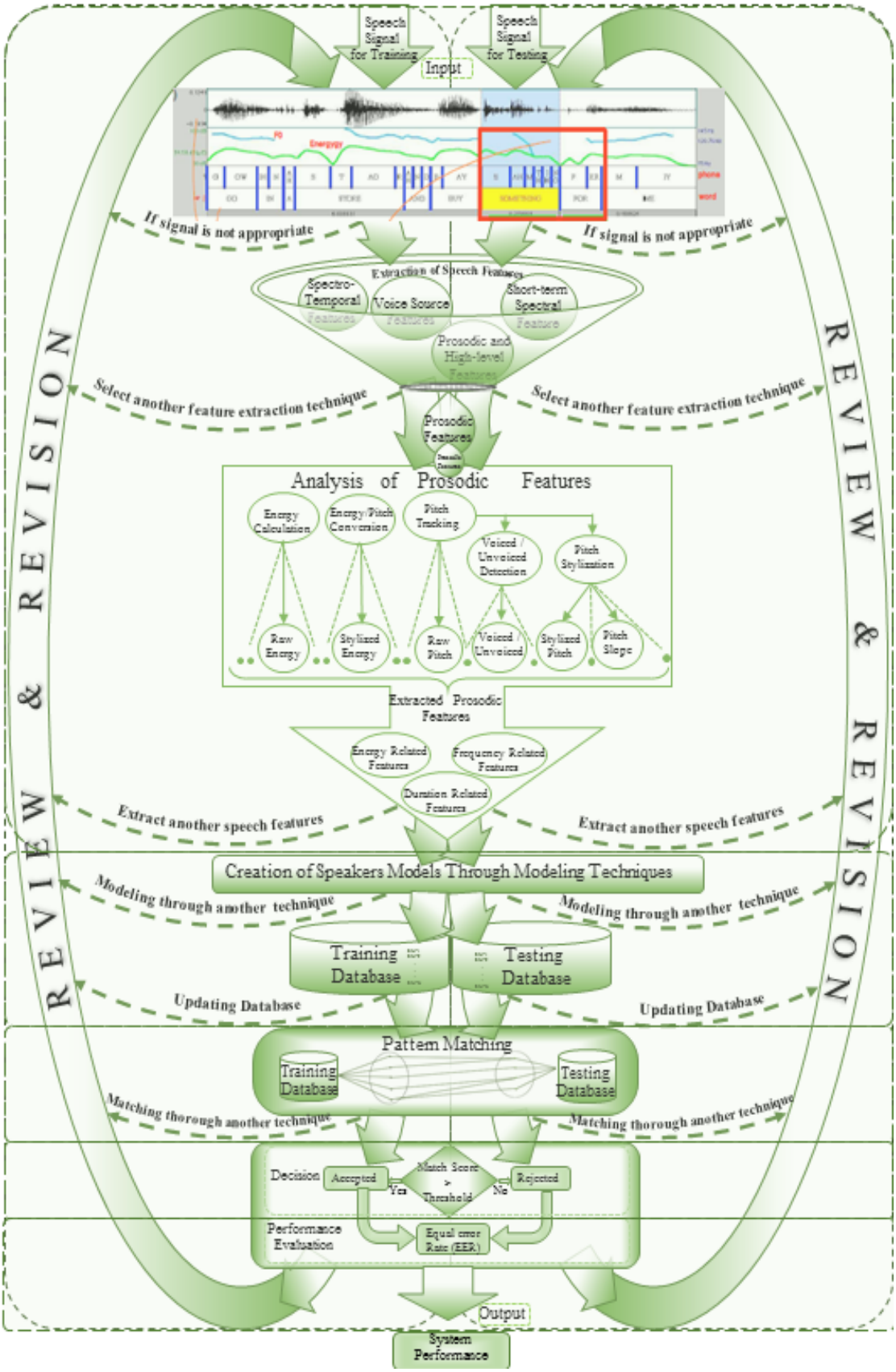


Figure-4.2: Implementation of the Proposed Framework for Development of Speaker Recognition System

4.2.1 Acquisition of Speech Signal

During implementation, experiments have been carried out for speaker recognition. For experiment 57 speakers were involved (both male and female). The range of speaker's age lies between 22 to 45 years. Each speaker was given to read different content for 2 to 3 minutes. Speakers were allowed to speak in their own reading style. Reading material included articles from newspaper and books. Every speaker read the randomly selected articles in Hindi & English both. The voice has been recorded in a normal lab environment with a sampling frequency of 8000 using a mono channel. At the time of voice recording in the lab, electric equipment was switched on. After recording of each speaker's voice, speech signal have been cut into smaller clips and then randomly four pieces have been selected out of those. These four pieces of speech signals are used for training and testing of speakers. Then preprocessing is applied on the speech samples and speech features are extracted by using prosodic feature extraction techniques. After feature extraction, these features are used to create speaker models. Finally these models are stored in the database. The voice database recording conditions have been given in table- 4.2.1(a) and description of the database has been given in table- 4.2.1 (b).

S. No.	**Device No.**	**Device Category**	**Sampling Rate**	**File Format**
1	Device 1	Head phone/microphone	16 kHz	.wav
2	Device 2	Portable voice recorder	44.1 kHz	Mp3
3	Device 3	Laptop Microphone	16 kHz	.wav
4	Device 4	Using MATLAB code	16 kHz	.wav

Table- 4.2.1(a): Recording of Voice and Device Specification

Language	**Hindi, English**
Speakers	57 (19 Males and 38 Females)
Speech Type	Read Articles Paragraphs
Recording Condition	Lab, Microphone Relatively clean
Handset Mismatch	No
Sampling Frequency	16 kHz
Quantization	16 bit
Training Speech Duration	Approximately 2 to 3 minutes
Evaluation Speech Duration	In between 5-12 seconds

Table- 4.2.1(b): Description of Developed Database

4.2.2 Extraction of Speech Features

There are many Feature extraction methods available to extract the speech specific features. For the purpose of analyzing a speech signal, it is converted into parametric values in the speech based recognition systems. Parametric values mean those values which have more discriminative and less variable speech features [126] [193]. During the implementation of this phase, prosodic features have been extracted. The following subsections are describing the extraction of prosodic features.

a) Analysis of Prosodic Features

Many researchers have investigated prosodic features of speech signal which are used to characterizing human behaviors over speaking expressions [221-223]. Prosodic features are rich in containing speech information about the speaker's gender, emotion, age, physical condition, attitude, intention etc. In this work researcher used the prosodic features such as duration, intonation, pauses, loudness, rhythm and timbre [77]. Prosodic features are also known as supra-segmental features. Prosodic features such as intonation patterns, rhythm, syllable stress and speaking rate are the non-segmental characteristics of speech [224]. Speech features that define prosody are stated as prosodic features of speech [39]. Table-4.2.2(a) shows different types of speech features and their example. On the basis of duration of the analyzed speech segment, prosodic features can be categorized into the following [52]:

- High-level features – Long-time features, time duration of a word or utterance.
- Source features – Prosodic features within a single glottal period;
- Supra-segmental features – Prosodic features spanning a few glottal periods;

Type of speech features	**Examples**
Spectral features	MFCC, LPCC, LFCC
Prosodic features	Pitch and energy contours
Dynamic features	Velocity/acceleration features Feature fusion multivariate auto-regression (MAR)
Supra-segmental features	F_0 contours Intensity contours

Source features	Glottal pulse shape
High level features	High level features

Table- 4.2.2 (a): Types of Speech Features and Examples

Prosodic features are commonly represented by pitch, energy and duration of syllables [189]. These are generally used to extract the information about speakers speaking style. The most common prosodic features are formants, fundamental frequency and energy of a speech signal. To improve performance of speaker recognition system a set of statistical constraints should be estimated based on the temporal parameters [52]. The authors proposed a feature extraction methodology in [225] and have made a number of improvements to the calculation of fundamental frequency and accent.

Some of the prosodic features include intensity/loudness, duration/rhythm, pitch/fundamental frequency etc. The table- 4.2.2 (b) shows some of the standard prosodic features which are globally defined:

Feature Name	**Description**
Mean	Mean F0 frequency (Hz)
Median	Median F0 frequency (Hz)
Max	99% value of F0 frequency (Hz)
Min	1% value of F0 frequency (Hz)
fracMax	95% value of F0 frequency (Hz)
fracMin	5% value of F0 frequency (Hz)
fracRange	5–95% F0 frequency range (Hz)
Range	1–99% F0 frequency range (Hz)
F0Var	F0 variance
Shimmer	Mean proportional random intensity perturbation
Jitter	Trend corrected mean proportional random F0 perturbation
LFE1000	Proportion of Low Frequency Energy under 1000Hz
LFE500	Proportion of Low Frequency Energy under 500Hz
GDposav	Average F0 rise during cont. voiced segment (Hz)
GDnegav	Average F0 fall during cont. voiced segment (Hz)
GDriseav	Average F0 rise steepness (Hz/cycle)

GDfallav	Average F0 fall steepness (Hz/cycle)
GDrisemax	Max rise of F0 during continuous voiced segment (Hz)
GDfallmin	Max fall of F0 during continuous voiced segment (Hz)
GDmax	Max steepness of F0 rise (Hz/cycle)
GDmin	Max steepness of F0 fall (Hz/cycle)
MeanInt	Mean RMS intensity (abs., dB)
MedianInt	Median RMS intensity (abs., dB)
MaxInt	Max RMS intensity (abs., dB)
MinInt	Min RMS intensity (abs., dB)
IntRange	Intensity range (abs., dB)
fracMaxInt	95% value of intensity (abs., dB)
fracMinInt	5% value of intensity (abs., dB)
fracIntRange	5–95% intensity range (abs., dB)
IntVar	Intensity variance (abs., dB)
Mavlngth	Average length of voiced runs
Manlngth	Average length of unvoiced segments shorter than 300 ms
Maslngth	Average length of silence segments shorter than 250 ms
Mlnlngth	Average length of unvoiced segments longer than 300 ms
Mlslngth	Average length of silence segments longer than 250 ms
max_vlngth	Max length of voiced segments
max_nlngth	Max length of unvoiced segments
max_slngth	Max length of silence segments
perc_short	Percentage of pauses shorter than 50ms
perc_mid	Percentage of 50–250 ms pauses
perc_long	Percentage of 250–700 ms pauses
Spratio	Ratio of speech against unvoiced segments >300ms
Vratio	Ratio of voicing against unvoiced segments
Sratio	Ratio of silence against speech
norm_intvar	Normalized segment intensity distribution width
norm_freqvar	Normalized segment frequency distribution width

Table- 4.2.2 (b): Prosodic features and some related acoustic features [40]

b) Pitch and Fundamental Frequency

Pitch contour is directly related to fundamental frequency (F_0). It is the most important prosodic property of speech. In the proposed work, F_0 is calculated from human utterance by using some statistics such as autocorrelation method or by Praat Software [218]. It was extracted randomly from 10ms to 30ms intervals or 30ms to 35ms intervals. The frequency range was set differently for male and female; for male speakers range is 75Hz to 400 Hz and for female speaker's frequency range is 100Hz to 600 Hz [40][225-227] [254].

c) Speech and Emotion

To represent emotional states voice is a strong and reliable indicator. The relationship between voice and emotion such as changes occur in articulation of sounds, breathing and phonation represent noticeable changes in the acoustic parameters related to the production of speech. Various theories have been established to find relation between emotion and speech characteristics. For example, sadness is expressed as low tone with long pauses and slow speech rate; fear is expressed as tight tone tense; joy is expressed as sharp tone with acceleration in speech rate. The communication process has been studied from many points of view. Speech is not only the medium to convey information and emotions but it also characterized by a particular communication style. Individuals have their communication style which changes over time and depends on mental status (happy, angry, rejecting/welcoming, friendly etc.) [77][228-229]. Table-4.2.2(c) represents the emotion on the basis of speaking style.

Emotion	Characterization of Prosodic Feature
Happiness	Moderate duration variations
	Brilliant tone
	Articulation predominantly detached
	High average sound level
	Tendency to tighten up the contrasts between long and short words
	Slightly rising intonation
	Quick attacks
	Slight or missing vibrato
Anger/ Fear	High noise level
	Articulation mostly not linked
	Distorted notes
	Relatively sharp contrasts between long and short words
	Very dry attacks
	Sharply stamp
Sadness	Relatively large variations in duration
	Low noise level
	Final delaying
	Soft attacks
	Tendency to attenuate the contrasts between long and short words
	Articulation linked
	Intonation (at times) slightly declining
	Soft tone
	Slow and wide vibrato

Table- 4.2.2(c): Prosodic Characterization of Emotions [77]

d) Feature Description

It is noticed that researchers are interested into make use of wide range of knowledge sources to extract features of speech signal. These features are related to dialog and sentence or word boundaries. The advantage of direct modeling approach is that no intermediate labeling of the prosody is required. Using prosodic information several

language processing task can be performed such as speaker recognition, sentence segmentation and tagging, dialog act segmentation and tagging and disfluency detection etc. A prosody model is created on the basis of these features of speech and then is used to build an outcome detection system [230-232].

Prosodic features extraction differs from implementation methods. For example pitch tracking is performed by SRI's model. It is the piecewise linear stylization algorithm. Implementation of the same is done by Praat's software which is autocorrelation based. In this work, researcher has taken various types of slope detection tasks so all the prosodic features are extracted around each word preceding a boundary and each word following a boundary [70] [231]. Figure-4.2.2 shows the procedure of extraction of prosodic features of speech signal. Some of the prosodic features are as:

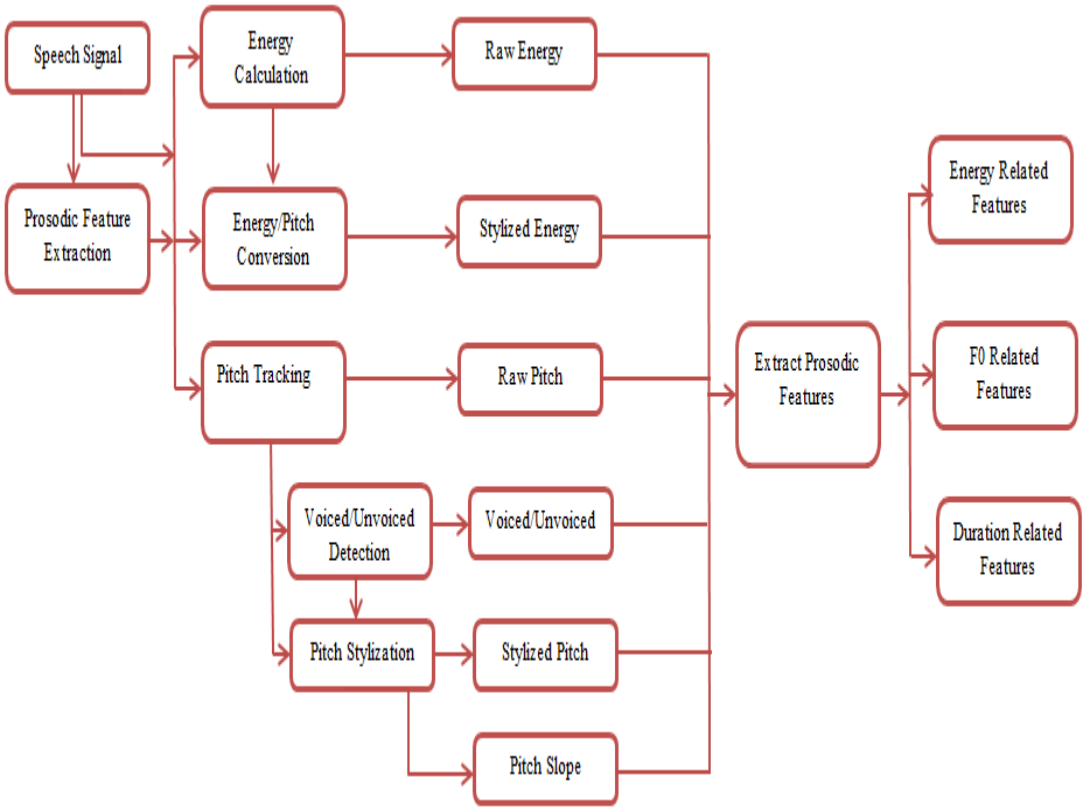


Figure- 4.2.2: Procedures for Extraction of Prosodic Feature of Speech

- **Duration Features:** These features are affected by device alignment and word of human transcriptions. Duration is the most important features of prosodic [192]. It represents the length of phonetic segments. Duration features contains a wide range characteristic of speech such as emotions (fear, anger, and happiness) [70] [77].
- **F_0 Features:** Extensive F_0 features are extracted from the preprocessed F_0 data. These features contain range features, boundary tone or movement features and slope features etc. [192] [77]
    - **Range Features:** These features basically are used to distinguish between offset and baseline at the offset position and mean and the baseline at the offset position. These features also contain the maximum, minimum, mean and last F_0 values of a word or window related to each word boundary. These include pitch range of a single word or a window with word boundary. Range features are normalized by the topline F_0, the baseline F_0 and pitch range by log difference and linear difference [40] [77] [222-223].
    - **Movement Features:** These features are measured from the variation of the F_0, for the voiced sections of speech or window following a boundary. F_0 values are calculated against the maximum, minimum, mean, first and last stylized F_0. It is compared with particular word or window through log ratio and log difference [77] [222-223].
    - **Slope Features**: Slope features are generally related to pitch slope. These features are generated from the stylized pitch value. In this case, features are computed as, the last slope value (word preceding a boundary) and the first slope value (word following a boundary). These features also include slope difference such as falling, rising and unvoiced as slope features [222- 223].
- **Energy Features**: These features are computed based on the intensity contour. As F0 features, several energy features also estimated. Energy features include movement features and slope features using numerous normalization methods [222].
- **Other Features**: In this study researcher have also included other features such as gender type [58].

Prosodic also use some non-verbal information such as stress, intonation and rhythm. This information is also known as prosody. Generally, for recognition purpose acoustic features are used such as formant, pitch and intensity [126] [233-234]. In this study prosodic properties of speaker's voice are represented in 49 dimensional feature vectors. The derived features are listed in table- 4.2.2(e).

Feature Sets	**Statistical Term**	**Number of Features (49)**
Pitch	Mean, median, standard deviation, maximum, minimum	5
Energy	Low, high, mean	3
Jitter, Shimmer	Jitter(local), Jitter(absolute), Jitter(rap), Jitter(db) Shimmer(local), Shimmer(local, dB)	6
Formants	Mean value of first, second and third formant (e.g. F1, F2, F3) Standard deviations of F1, F2 and F3 Band width of F1, F2 and F3	9
Intensity	Mean, median, maximum, minimum, standard deviation	5
Harmonicity	Mean autocorrelation Mean harmonic to noise ratio Mean noise to harmonic ratio	3
Long term average	Mean, maximum, minimum, slope, kurtosis, skewness Amplitude of the first and second harmonic	18

Table- 4.2.2(e): Statistical Term used for Prosodic Features of Speech

To extract these speech features MATLAB as well as Praat software have been used. In this analysis autocorrelation method is used for pitch estimation, with 12.5ms time step, 60-350 Hz frequency interval. Burg algorithm is used for formant prediction with 10ms; 25ms time step window width and 5500 Hz maximum formant frequency. Table- 4.2.2 (f) shows some Prosodic and Acoustic features of speech.

Feature Name	Short Description of Feature
Mean	Mean F0 frequency (Hz)
Median	Median F0 frequency (Hz)
Max	99% value of F0 frequency (Hz)
Min	1% value of F0 frequency (Hz)
fracMax	95% value of F0 frequency (Hz)
fracMin	5% value of F0 frequency (Hz)
fracRange	5–95% F0 frequency range (Hz)
Range	1–99% F0 frequency range (Hz)
F0Var	F0 variance
Shimmer	Mean proportional random intensity perturbation
Jitter	Trend corrected mean proportional random F0 perturbation
MeanInt	Mean RMS intensity (abs., dB)
MedianInt	Median RMS intensity (abs., dB)
MaxInt	Max RMS intensity (abs., dB)
MinInt	Min RMS intensity (abs., dB)
IntRange	Intensity range (abs., dB)

Table- 4.2.2 (f): Prosodic Features and other Related Acoustic Features [40]

e) Implementation Details

We have implemented prosodic features extraction by MATLAB and Praat software. The reason behind using Praat is that it provides an existing suite of high quality speech analysis routines, such as pitch tracking [226] [231]. There are some additional reasons for using Praat such as:

- It is widely used for speech analysis toolkit that is supported by many platforms e.g. Windows, Macintosh, Linux, and Solaris etc.
- It provides different type of valuable data to represent various types of information from speech signal e.g. TextGrid, PitchTier etc.
- It is especially useful for extracting prosodic features of speech.
- It has built-in programming scripting language to extending its capability.

It is clear that Praat is a best platform for extracting speech features of speech signal.

4.2.3 Creation of Speaker Models

To create speaker model Gaussian Mixture Model (GMM) is used. GMM modeling technique is used for the development of speaker models. It was first introduced by Reynolds in 1992. Initially it was not popular but shortly it became more popular to develop a speaker recognition system. It has been revealed that Gaussian Mixture Model is very much appropriate for voice modeling in speaker recognition system. For Speaker recognition, Gaussian mixture model is an essential appliance of statistical clustering [23] [149]. For every utterance we calculate likelihood of the Gaussian mixture model for each parameter. Then the likelihoods are converted into posterior probabilities using bayes rule and a set of prior probabilities [253].

To create speaker models by using Prosodic speech features different types of segmental features of speech are used. For example pitch, energy, Jitter, Shimmer, spectral tilt, log normalized duration etc. The features were then modeled with Gaussian mixture model. This modeling was done for both i.e. training and testing. After that the created models were adapted from each speech parameter. A 'maximum a posteriori' (MAP) is commonly used for adaption in speaker recognition. This approach is good because it permits regularization parameter and use significance factors that restraint the quantity of covariance, global means and weight etc. It should be adapted to data from each speech features parameter [192].

4.2.4 Pattern Matching

Matching (pattern matching) phase is responsible for comparing the estimated speech features with the stored speaker models. In this phase, previously stored speaker model are matched with the claimed or new acquired speech signal. In case of open-set (SV & SI) estimated speech features are also compared with unknown speaker models. In case of verification, resultant is the similarity score between claimed identity and test sample. While in case of identification, resultant is the similarity score for all the stored speaker models. Now, on the basis of similarity score (statistical or deterministic) decision has been made. Suppose that M is the number of stored speakers models and V is the test vector such as $V = \{v_1 \ldots . v_t\}$, extracted from the unknown speakers speech sample. To define match score $s(V, M) \in M_i$ specifying the similarity between V and Mi. Matching with speaker model be determined by the

type of model, dissimilarity value, likelihood ratio etc. Now the enrolled speaker having best match is selected [4] [75] [144] [235].

4.2.5 Decision

The aim of this phase is to declare whether a speaker who is claiming to be present is actually present or not. For decision, a threshold is set. If the resultant is greater than or equal to threshold then the speaker is accepted by the system otherwise rejected. In this research, likelihood is used for decision making. Generally likelihood-ratio test is used to take a decision that claimed speaker is accepted or rejected [4] [149-150]. Suppose for an utterance V, claimed speaker recognition $\overline{Vm}$ model is Vm and remaining available non-claimed model are then the likelihood ratio Lr is:

$$\frac{\text{V is the Claimed Speaker}}{\text{V is not from the Claimed Speaker}} = \frac{Vm/V}{V\overline{m}/V} \qquad (4.2.5)$$

After evaluation of likelihood ratio Lr, it is compared with a defined threshold (let threshold is λ). A speaker is accepted if Lr ≥ λ and rejected if Lr < λ.

4.2.6 Performance Evaluation

It is already discussed that by implementing the proposed framework using prosodic features performance of the speaker recognition system can be improved. The researcher has implemented proposed framework by using prosodic features. Now, in this phase performance of the speaker recognition system which is developed by implementing the proposed framework is evaluated. For evaluation of the proposed speaker recognition system, the equation 4.2.6(a) has been used:

$$SRR = \frac{\text{Number of Correctly Recognized Speaker}}{\text{Number of Speaker}} X\ 100 \qquad (4.2.6(a))$$

Where SRR → Speaker Recognition Rate

For a closed-set identification system, accuracy of identification is used to measure the system performance. As speaker verification systems have two types of errors namely; false acceptance and false rejection, the performance measurement

depends on these two factors. The metrics used for the purpose are the False Acceptance Rate (FAR) and False Rejection Rate (FRR). FAR and FRR are key metric used to make matches or rejections quickly. To make speaker recognition system more accurate FAR is should be minimum [173].

Performance of speaker identification system is decided by Equal Error Rate (EER). It is the most common method used to evaluate system performance. EER is a point where probability of False Acceptance (FA) and probability of False Rejection (FR) both are equal [4] [187]. FAR and FRR can be measure by using equation 4.2.6 (b) and 4.2.6 (c). The metrics are defined as:

(i) FAR is the probability that the system incorrectly match an imposter. It is calculated in percentage. So it also defined as **"the percentage of invalid inputs which are incorrectly accepted"** [173]**.**

$$\text{FAR} = \frac{\text{Number of False Acceptance}}{\text{Number of Identification Attempt}} \qquad (4.2.6(b))$$

(ii) FRR is the probability that system incorrectly rejects to a registered speaker or "**the percentage of valid inputs which are incorrectly rejected**" [43].

$$\text{FRR} = \frac{\text{Number of False Rejection}}{\text{Number of Identification Attempt}} \qquad (4.2.6(c))$$

The speaker identification rate can be calculated by using the formula Equal Error Rate [150]. Then the performance of speaker recognition system is measured using the speaker identification rate metric given in 4.2.6(d). EER is defined as:

(iii) Equal Error Rate (ERR): Equal error rate is an algorithm for biometric security system used to predetermine the threshold value. The reliability of recognition system decision is main concern in speaker recognition applications. The common method used in speaker recognition is adjustment between false acceptances and false rejections. To evaluate a biometric system performance EER unit is calculated. It estimates the error rate of a system. This error rate is measured on the basis of threshold. The threshold is adjusted to generate equality between false acceptances and false rejections [236]. Table- 4.2.6 describes about the possibility of identity authentication.

$$EER = \begin{cases} \frac{FAR(t1)+FRR(t1)}{2} & ; \text{If } FAR(t1)-FRR(t1) <= FRR(t2)-FAR(t2) \\ \frac{FAR(t2)+FRR(t2)}{2} & ; \text{otherwise} \end{cases}$$

	Matches Value	
Type of Access	Measured data matches expected value	Measured data does not matches expected value
Authorized individual access	True Accept (TA) Access correctly granted	False Reject (FR) Access incorrectly denied (type I error)
Unauthorized individual access	False Accept (FA) Access incorrectly granted (type II error)	True Reject (TR) Access correctly denied

Table-4.2.6: Possibilities of Identity Authentication (Matrix)

Generally, system performance is measured as a decision cost function for example NIST SRE evaluates system performance on the basis of Decision Cost Function (DCF). This function measures as the weighted probability of falsely rejected user and weighted probability of falsely accepted imposter. The final result is evaluated in Detection Error Tradeoff (DET) curves [80-81]. Today, majority of recognition systems are text-independent speaker recognition systems. These systems are used in many applications performing with low error rates. For example, forensic applications require much improved error rates for identification purposes [80] [82].

Many researchers concentrate on reducing error rate. Decreasing error rates and hence improving system performance is still a research area in the field. The identification rate can be measure by using equation 4.2.6(d):

$$\% \text{ Speaker Identification Rate} = \frac{\textbf{No. of Correctly Identified Speaker}}{\textbf{Total Number of Segments}} * \textbf{100\%} \qquad (4.2.6(d))$$

The identification system performance is measured using identification error. The identification error can be measure by using equation 4.2.6(e):

$$\% \text{ Identification Error} = \frac{\textbf{Incorrectly Identified Speaker}}{\textbf{Total no.of Speaker}} * \mathbf{100} \qquad (4.2.6(e))$$

System performance also depends on the following parameters [7] [253]:

- The number of Gaussian mixture components
- Type of speech feature used

The system is tested using 32, 64, 128 and 512 Gaussian mixture component and different types of feature extracted from speech signal (Prosodic and MFCC)

4.3 Summary of Steps for Speaker Identification

The overall process of recognizing a speaker has been summarized in the form of an algorithm. Figure-4.3 represents the structure of GMM based speaker identification system. In speaker identification system there are two phases training and recognition (identification). In training phase, a mathematical model (GMM in our case) has been constructed for each speaker through speaker's voice and it is stored in the database. On the other hand in recognition phase, speech data is examined for best match with the speaker models available in the database [237]. The identification procedure is stated as follows:

1. Extracts the features of speech signal
2. Compute the set of speech feature $P = \{p_i\}$
3. For each speaker model M_i, DO (Compute the distortion $D_i = d(P, M_i)$ between P and M_i)
4. Identify the value of the unknown speaker Id as the one with the smallest distortion, i.e. $Id = \underset{i=1\ldots.N}{\text{argmin}} \{D_i\}$

The distortion measured in the second step approximates the dissimilarity between the feature vectors $M_i = \{m_{i1}, m_{i2}, \ldots, m_{iK}\}$ and the vector set $P = \{p_1, p_2, \ldots, p_n\}$. In this, the most intuitive distortion measure has been used for mapping each vector in P to the nearest code vector in Mi and for computing the average of these distances by using the following formula:

$$\frac{1}{N}\sum_{j=1}^{N} \min_{k=1}^{K} d_E (P_j, M_{iK}) \qquad (4.3(a))$$

Where d_E is the Euclidean metric:

$$d_E(x, y) = \sqrt{\sum_{i=1}^{dim} (x_i - y_i)^2} \qquad (4.3(b))$$

The distortion measure, known as the Mean Square Error (MSE) also gives a measure for the quality of the codebook constructed from the training set P.

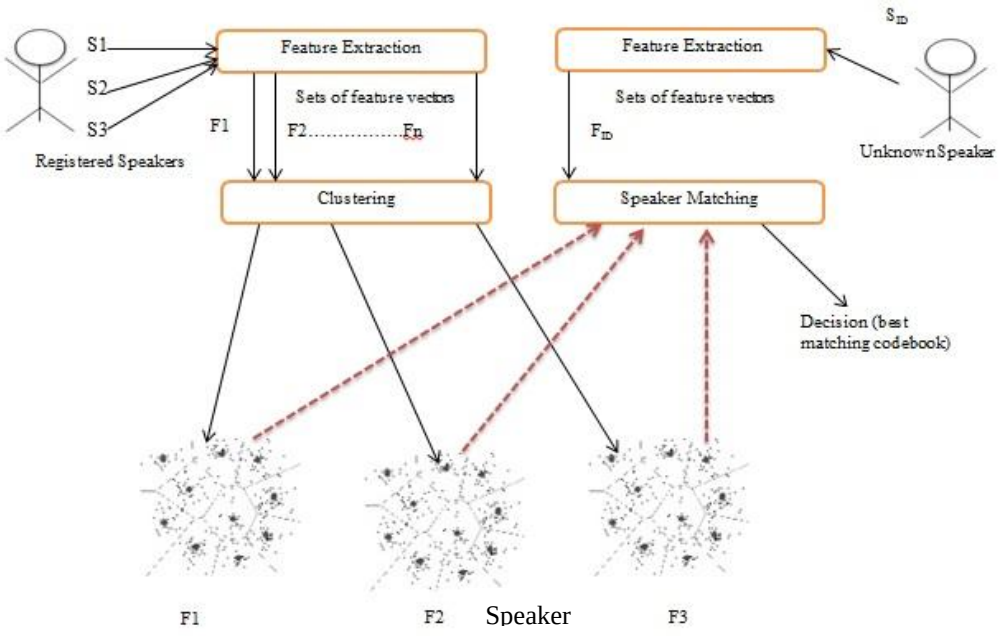


Figure-4.3: Structure of Speaker Identification System

4.4 Conclusion

The term distinctiveness is the key of authentication techniques. Voice biometric is the branch of science and engineering. Authentication based on human biometric is the procedure to analyze the legal problem of digital evidence. It is the combination of science and law. Improvement in the speaker recognition system performance is

needed as these systems are used in many important and secure fields including voice based banking, access control, crime investigation purpose etc.

Automatic Speaker Recognition (ASR) is a technique for recognizing an individual by his/her voice. Controlling access privileges and forensic is the major application areas of speaker recognition systems. To develop a robust speaker recognition system it is required that the system is able to provide acceptable performance with several operating conditions. Though various developments have been done in the area but there are still many improvements required. In this chapter researcher has implemented the framework proposed in chapter-3 for developing a speaker recognition system. The researcher has also claimed that by choosing prosodic features for implementing the framework, system performance has been improved.

Chapter- 5
Experiments and Result

CHAPTER-5: EXPERIMENTS AND RESULT

Once a decision was made, I did not worry about it afterward.

—Harry Truman

5.1 Introduction

In developing speaker recognition system performance is the main task. In addition, system consistency is also a key factor. To make more consistent system longer voice database is required. The purpose of this research is to study English vs. Hindi and Hindi vs. English voiced samples using different recording mediums. During recording different equipment are used. The environmental noise and the time delay occurred when taking voice samples. In this chapter different testing results are shown for proposed text-independent system. The purpose of this research is to investigate the prosodic features which are more useful to improve speaker recognition system performance.

Researcher has tried to found the main obstacle which degrades the system performance and found solution to overcome from that. In speaker recognition noise is the main reason to degrade the system performance. Hence in presence of noise, identification is more difficult. In chapter 4, researcher investigated that prosodic features are robust against the noise. Therefore this research focused on the prosodic features of speech. A number of experiments are conducted for different prosodic features and Gaussian mixture model as well as MFCC with Gaussian mixture model. Modeling is also an important factor for speaker recognition. As studies shows that Gaussian mixture model is most commonly used method for modeling [132] [142] [153]. The results obtained by the experiments are used for the analysis purpose of system performance. The main objective of this research is to find the speech features which are more robust against noise and time variability. In addition system performance is calculated in terms of identification rate.

5.2 Training Conditions

The training condition is defined as amount of data (voice) and resource of acquiring speech signal to develop a speaker recognition system. It is evaluated as fixed or open training condition [239].

- **Fixed-**In this approach acquiring speech signal limits the system to specific data set for training condition. For example; new voice database creation or use of previous available voice database.
- **Open-** It removes the conditions exist in fixed training condition approach. For training purpose other available data (publicly available data) can be used.

5.3 Enrollment Conditions

During enrolment of speaker's speech, it is required to acquire at least that amount of data which is sufficient to create speaker models.

5.4 Test Conditions

The test condition is the process of recognizing a speaker from enrolled data to a given speech segment and automatically decide whether target speaker identified or not. For testing, there must be created some rules which vary system to system for recognition.

5.5 Performance Measurement

To calculate the performance of identification system, the following equation (5.5) is applied:

$$I = \frac{CI}{N} * 100 \qquad (5.5)$$

Where I is the identification rate (total correctly identified speakers), CI is the number of correctly identified speakers, and N is the total number of speakers [239] [259].

5.6 Experiments with Prosodic Features

In the research the speaker identification system is developed by using prosodic and Gaussian mixture model. Initially a speech signal is pre-emphasis before speech segments are framed. The frame size of speech signal is 30-45ms with different frequency for male and female. The research experiments are based on Hindi and English language only for training and testing both. Testing sample length has been

extracted from the speech sample recorded form enrolled speakers. The voice data recorded by headphone is only considered for training and testing purpose.

In the research experiment is carried out in a training database developed in lab condition. This experiment contains 57 speakers and 1561 utterances. For training every speaker's 2-3 minutes recording is available. During testing speech segment length in between 30 second to 45 second. Speaker recognition system is GMM based speaker recognition system. In our system we use prosodic (49 dimensional) features extracted from the created database and 32, 64, 128, 512 Gaussians as speaker models under the 30-45 second training condition. During experiment extracted pitch, duration and energy related features and break the speech signal into segments. By using Legendre polynomial expansions we estimated pitch and energy contours for each segment. And calculate one feature vector for each segment after that model these features by using GMM. Table- 5.6 experiment conditions of speaker identification systems.

Task	**Text-independent Speaker Identification System**
Language	English, Hindi
Front-ends	Prosodic
Back-end	Gaussian Mixture Models (GMM) with 32, 64, 128, 512 Gaussian components
Number of coefficients in a feature vector	34(10static + 12delta + 10delta-delta+ 2 energy) for MFCC
Window size	35-45ms
Step size	10ms
Sampling rate	16kHz
Pitch floor	75 Hz
Number of pitch (per Speaker)	5
Pitch ceiling	350 Hz
Training set	57 Speaker
Testing	51 Speaker
Platform	Laptop, PC, portable Recording Device
Programming Language	MATLAB

Table- 5.6: Experiment Conditions of Speaker Identification Systems

5.7 Result and Discussion

Figure- 5.7 (b) shows the speaker recognition (Identification) performance for different training data. The EER for matched case is varying approximately 5-15% as reported in [67] [150] [115] [216] [206-208]. The result obtained in this thesis the EER is 4.26 – 5.39 % is clearly better. The EER is represented for the testing conversation, and it is the statistically significant for the work. Table- 5.7 (a) shows the training and test condition for proposed speaker recognition system and table- 5.7(b): EER values of speaker recognition system using prosodic features.

Training Channel	**Test Channel**	**No. of Utterances**
Head phone/microphone	Head phone/microphone	584
	Portable voice recorder	317
Portable voice recorder	Head phone/microphone	437
	Portable voice recorder	223
	Total	**1561**

Table- 5.7 (a): Training and Test Condition for Proposed Speaker Recognition System

Training Language	**Testing Language**	**Recognition Rate (%)**	**EER (%)**
English	English	95.74	4.26
	Hindi	94.61	5.39
Hindi	English	94.61	5.39
	Hindi	95.74	4.26

Table- 5.7(b): EER Values of Speaker Recognition System using Prosodic Features

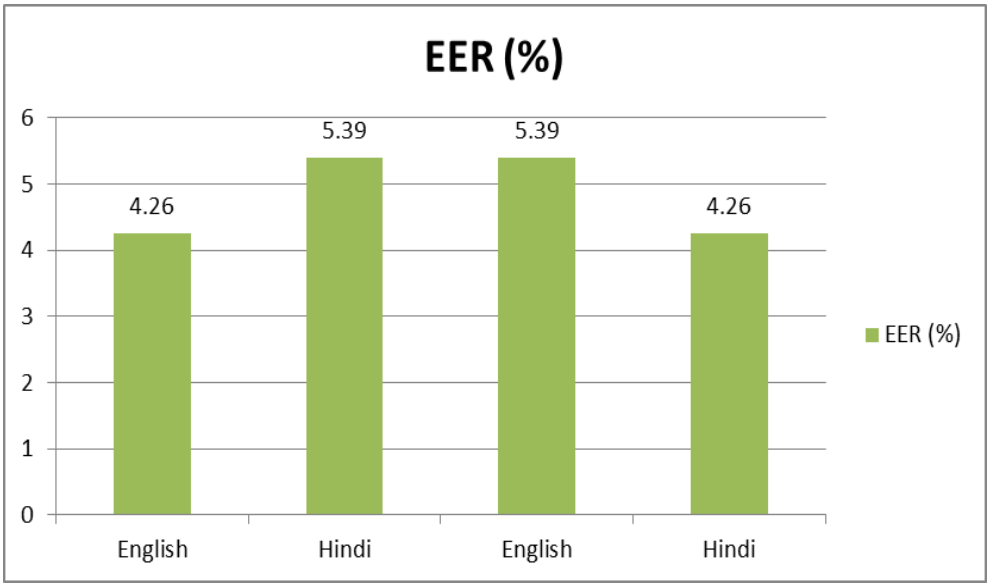


Figure- 5.7(b): EER Values of Speaker Recognition System Using Prosodic Features

5.8 Comparison of Speaker Recognition Performance

The researcher has proposed a speaker recognition system using Prosodic features of speech. It has been concluded from the literature that prosodic features could be used to enhance the performance of speaker recognition system [223] [240] [259]. In the research work, features have been extracted using MFCC and Prosodic; later speaker modeling have been performed. These speaker models are used for pattern matching and then match score is calculated of the individual speaker.

Almost all the implementation and experiment have been performed using MATLAB and few with the help of Praat software. In the session, all data collection, implementation results have been retrieved and comparative analysis has been performed and discussed in subsequent section. The results shown in this chapter prove that prosodic features produce better result than MFCC. In addition, it is also noticed that prosodic is individually helpful to produce a robust speaker recognition system.

5.8.1 Mel-Frequency Cepstrum Coefficients

Mel Frequency Cepstrum Coefficient (MFCC) is one of the most commonly used methods to feature extraction for automatic speaker recognition system. MFCC are coefficients that represent sound based on human perception. This is also one of the

most popular methods used for speaker recognition in base speaker recognition system [241]. But the drawback of MFCC is that it highly affected by noisy data. MFCC are derived by taking the Fourier Transform of the signal, warping it to by using a Mel-filter bank that closely mimic the Mel-scale, the final step is to perform Discrete Cosine Transform on the logarithm power of the speech frame from the Mel-scale output. Table- 5.8.1 shows the system performance using MFCC features extraction method and figure- 5.8.1 is graphical representation of this. Performance of a speaker recognition system can be calculated using the following formula:

$$\text{Performance (\%)} = \frac{no.of\ speech\ samples\ identified}{total\ no.\ of\ speech\ samples\ tested} \times 100 \qquad (5.8.1)$$

Training Language	Testing Language	Recognition Rate (%) MFCC
English	English	89.70
	Hindi	86.50
Hindi	English	86.50
	Hindi	89.70

Table- 5.8.1: Performance of Automatic Speaker Recognition System using MFCC

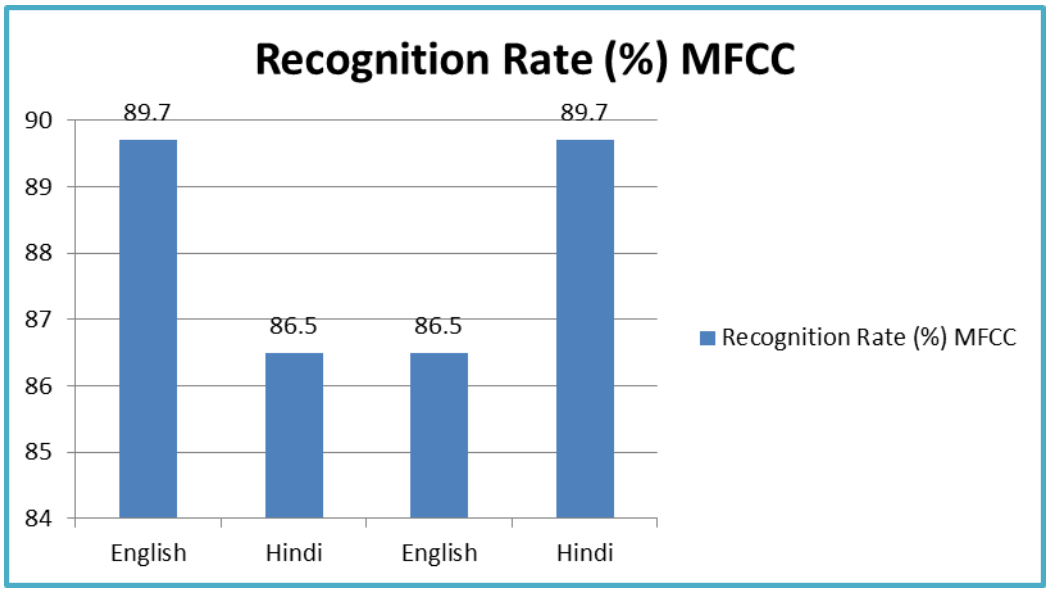


Figure- 5.8.1: Performance of Automatic Speaker Recognition System using MFCC

The result shows the recognition rate of MFCC for Hindi and English is respectively ranging from 86.5 % and 89.7%. MFCC using different coefficients (MFCC20, MFCC23, MFCC32 etc.) have the different recognition rate. Overall, the MFCC recognition rate is less when compared to the Prosodic. It did not achieve a 100% for speaker recognition but the recognition rates are consistent for clear noise but its performance degrades when the speech signal is noisy.

5.8.2 Prosodic

The common Prosodic features of a speech signal are duration, energy, pitch, intensity, speaking rate. Pitch is helpful in extracting noise-robust features of speech [18] [40]. Automatically recognizing the expressive state of speakers from their voice is known as emotion recognition. Attributes of voice are variable due to different moods of speaker and the affected attributes are pitch, speaking rate, intonation [240] [242]. Prosodic features of a speech signal changes continuously throughout the utterance. Hence, during analysis short time speech signal is frequently used and generally the length of speech signal is 10-30ms [4] [7]. The performance measurements revealed that the extracted basic prosodic parameters are quite robust to noise environment [243]. The errors were found to be unbiased as result shown in table- 5.8.2. Graphical representation of recognition accuracy with feature dimension 12, 16, 34 and 49 is shown in figure- 5.8.2.

$$\text{Performance (\%)} = \frac{no.of\ speech\ samples\ identified}{total\ no.\ of\ speech\ samples\ tested} \times 100 \qquad (5.8.2)$$

Training Language	**Testing Language**	**Recognition Rate (%) Prosodic**
English	English	95.74
	Hindi	94.61
Hindi	English	94.61
	Hindi	95.74

Table- 5.8.2: Performance of Automatic Speaker Recognition System using Prosodic

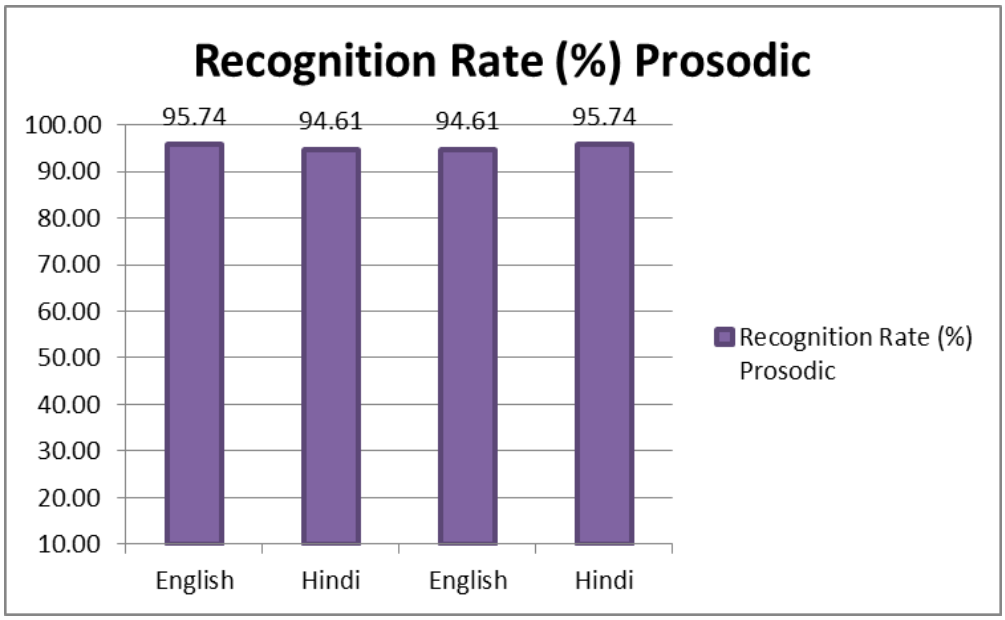


Figure- 5.8.2: Performance of Automatic Speaker Recognition System using Prosodic

The results show that the performance of the recognition rate of features extracted using Prosodic ranging from 95.74% to 94.61%. Results show that by using prosodic using the characteristics has the better recognition rate can be achieved. The recognition rate peaks at the 49th order of the prosodic for all Gaussian components. The findings are expected as prosodic alone is efficient when higher order coefficients are used as superfluous information or noise will be included in the modeling of the speaker model resulting in lower recognition rates. The results also show Prosodic to be more robust and accurate than MFCC.

5.9 Speaker Recognition Lab

Figure- 5.9 (a), 5.9 (b), 5.9 (c) and 5.9 (d) shows the lab setup of speaker recognition lab and experimentation with enrolled speakers.

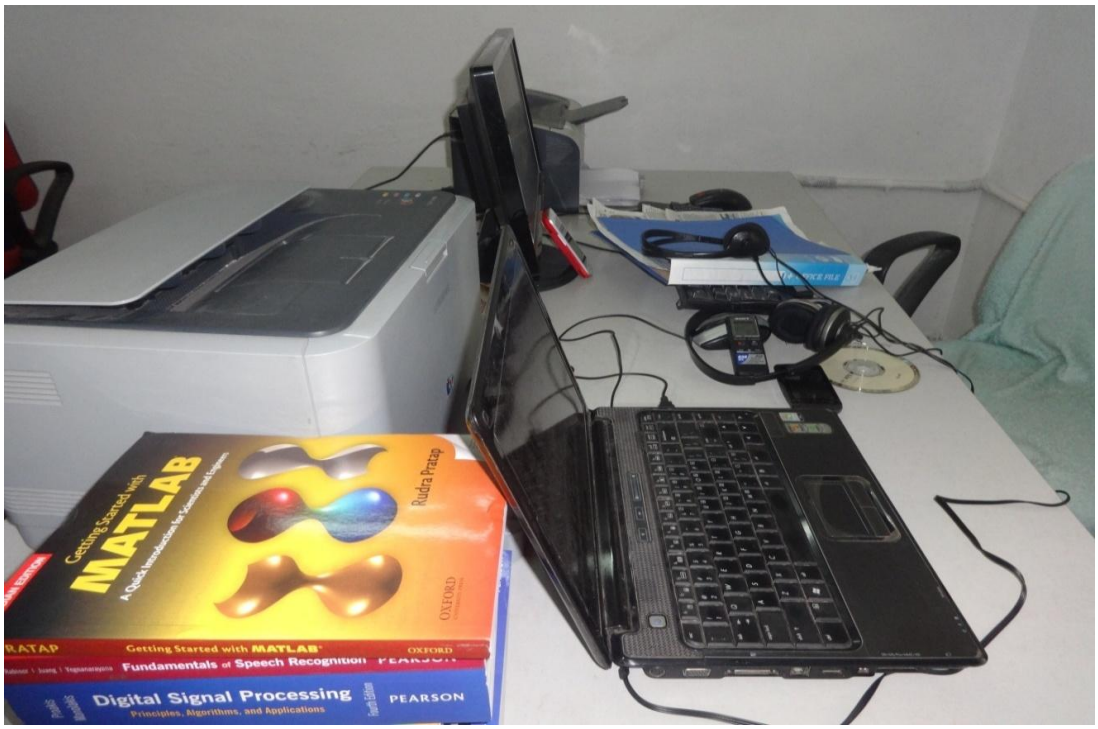


Figure-5.9 (a): Test Bed Setup of Speaker Recognition

Figure-5.9 (b): Voice Data Acquisition Screenshot

Figure-5.9 (c): Test Bed Setup of Speaker Recognition

Figure-5.9 (d): Voice Data Acquisition Screenshot

5.10 Screen Shots of Calculated Results

The screenshots during experimentation have been shown in the figure- 5.10 (a) to figure- 5.10 (r).

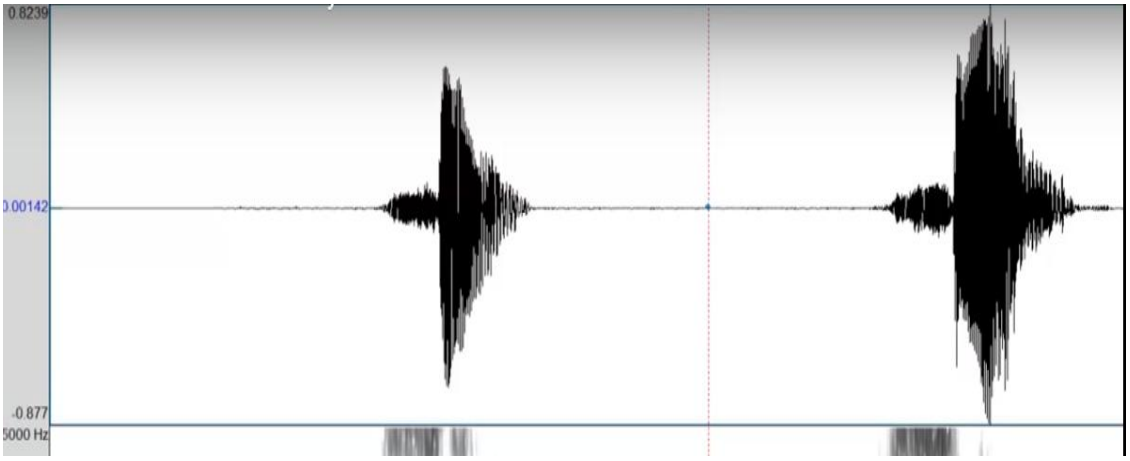


Figure- 5.10 (a): Spectrogram of a Speech Signal

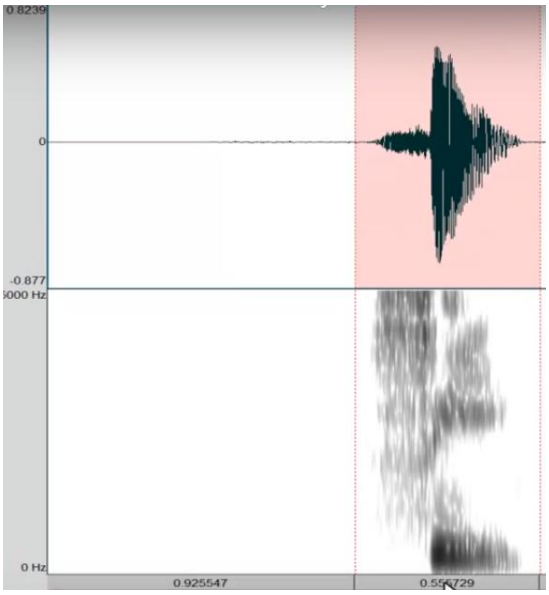


Figure- 5.10 (b): Formant of a Speech Signal

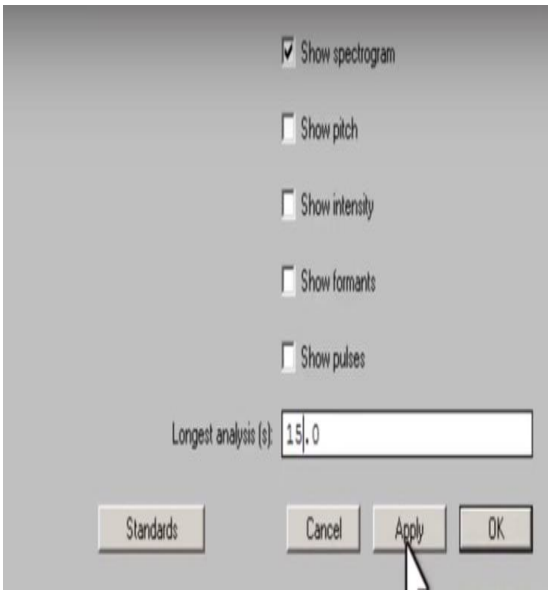


Figure- 5.10 (c): Extracting Different Features from Speech Signal

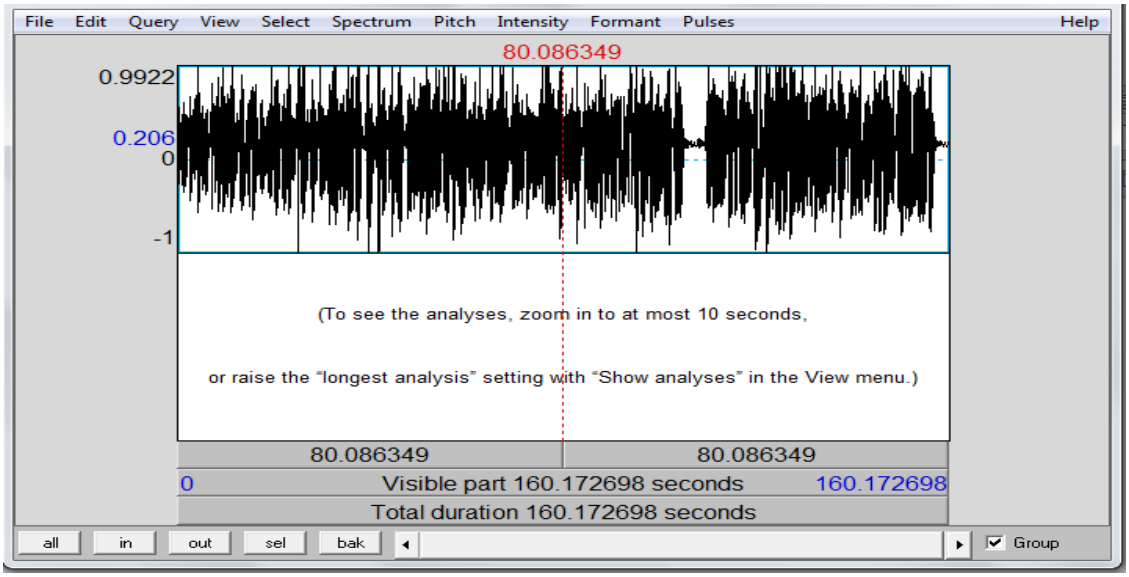


Figure-5.10 (d): Read a Sound File by Praat Software

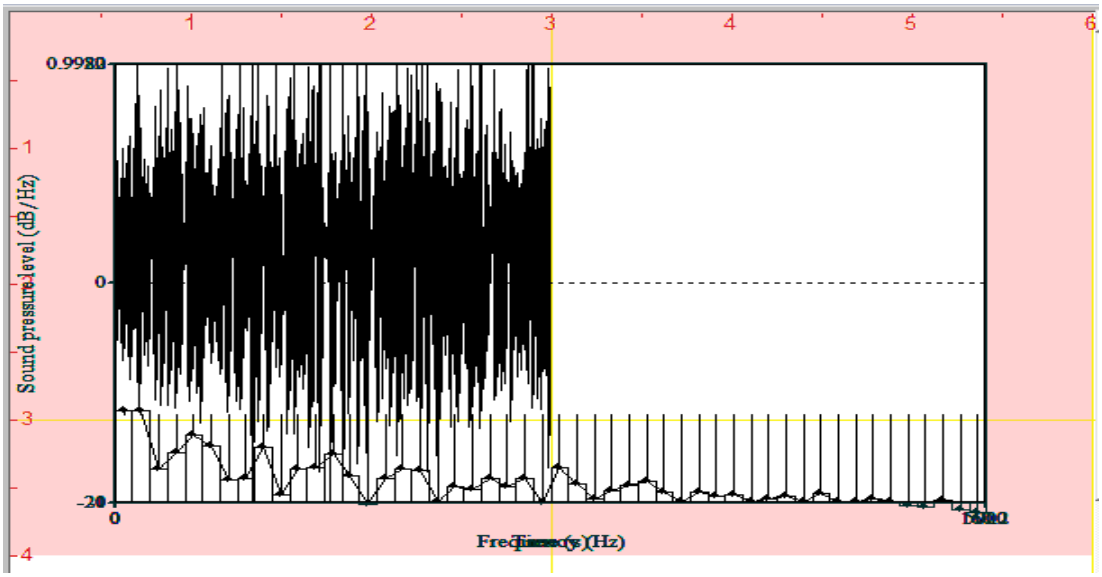


Figure- 5.10 (e): Sound Pressure of a Speech Signal

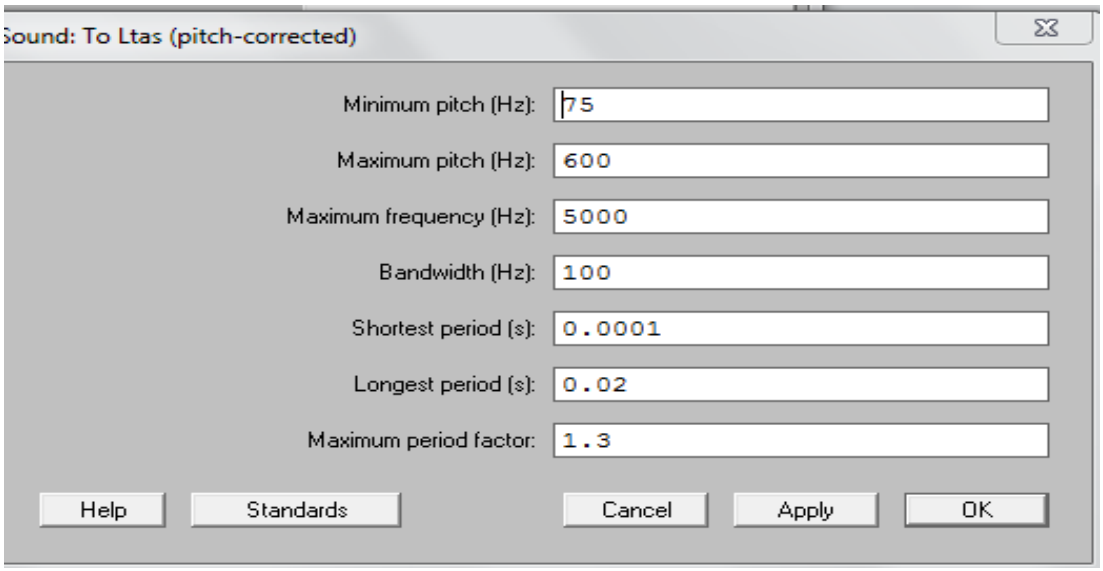


Figure- 5.10 (f): Pitch of the Voice

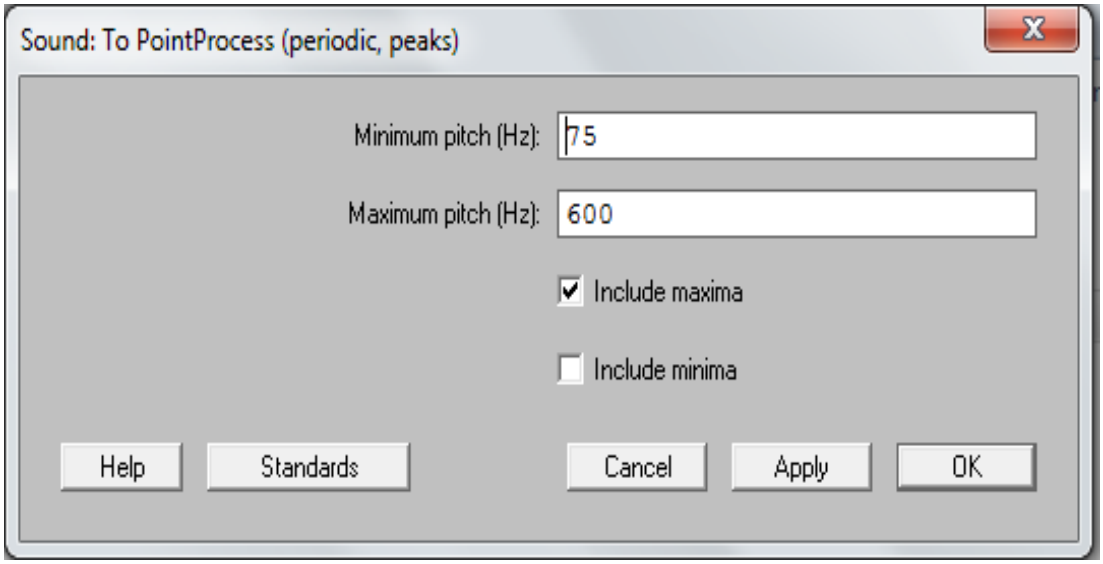


Figure- 5.10 (g): Maximum and Minimum Pitch Voice

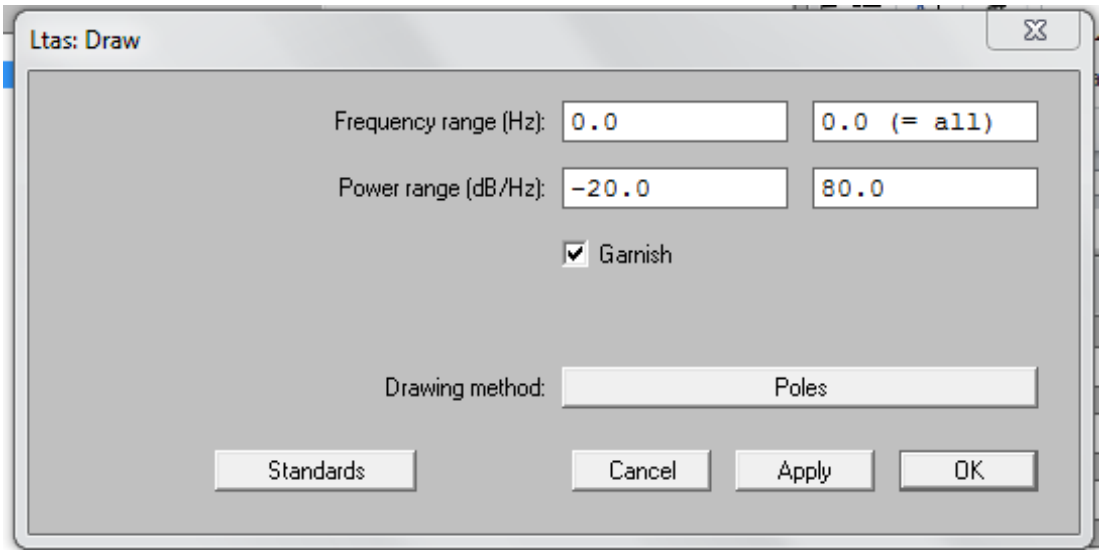


Figure- 5.10 (h): Frequency of the Voice

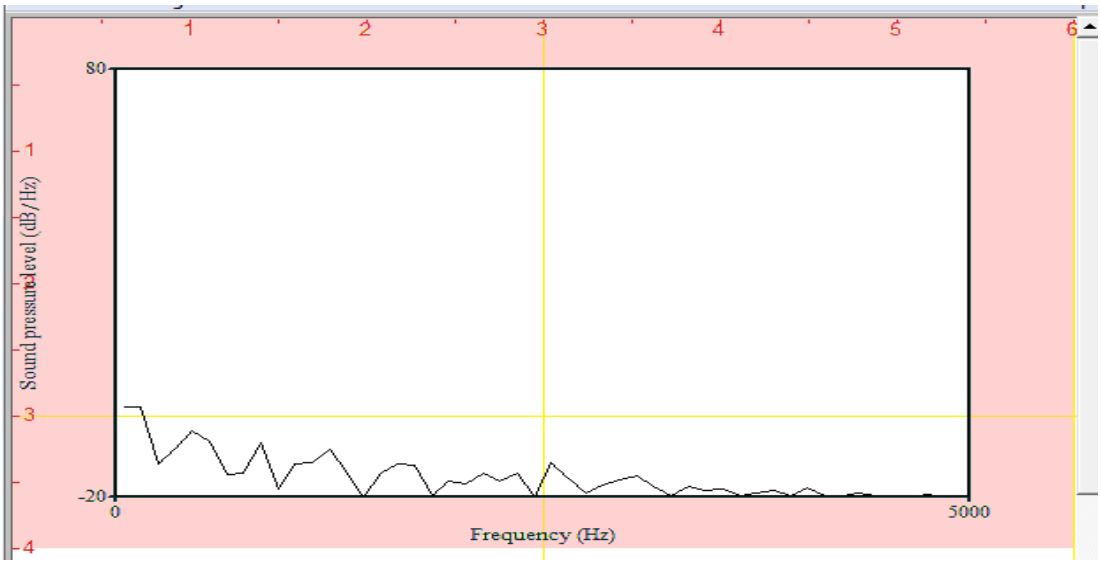


Figure- 5.10(i): Frequency in Curve form

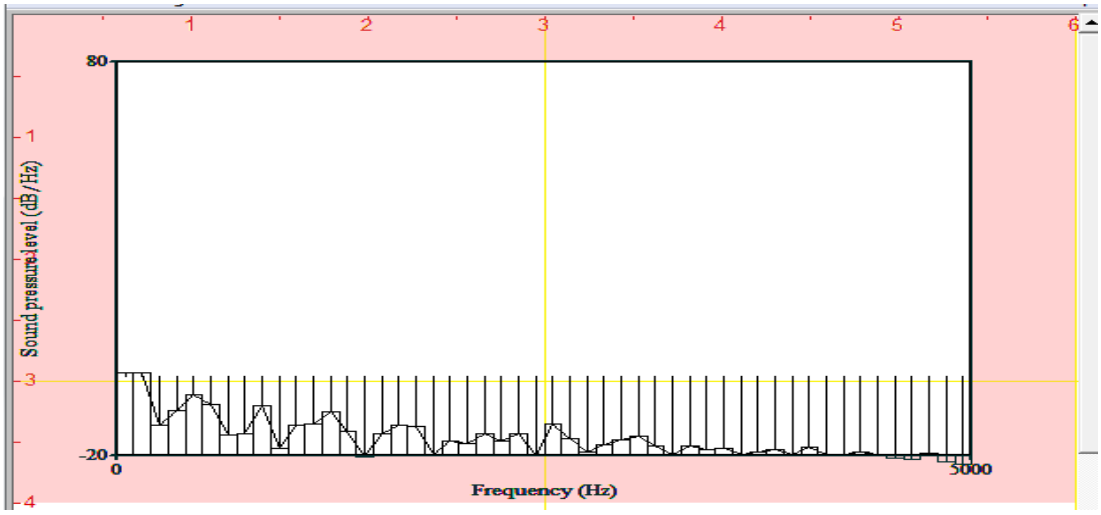


Figure- 5.10(j): Frequency in Poles Form

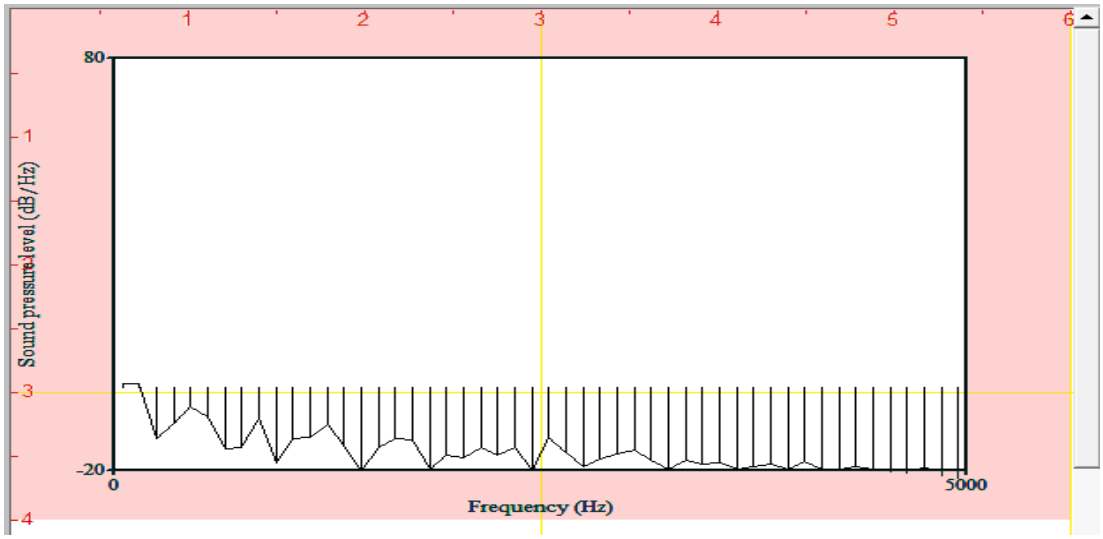


Figure- 5.10(k): Frequency in Bars Form

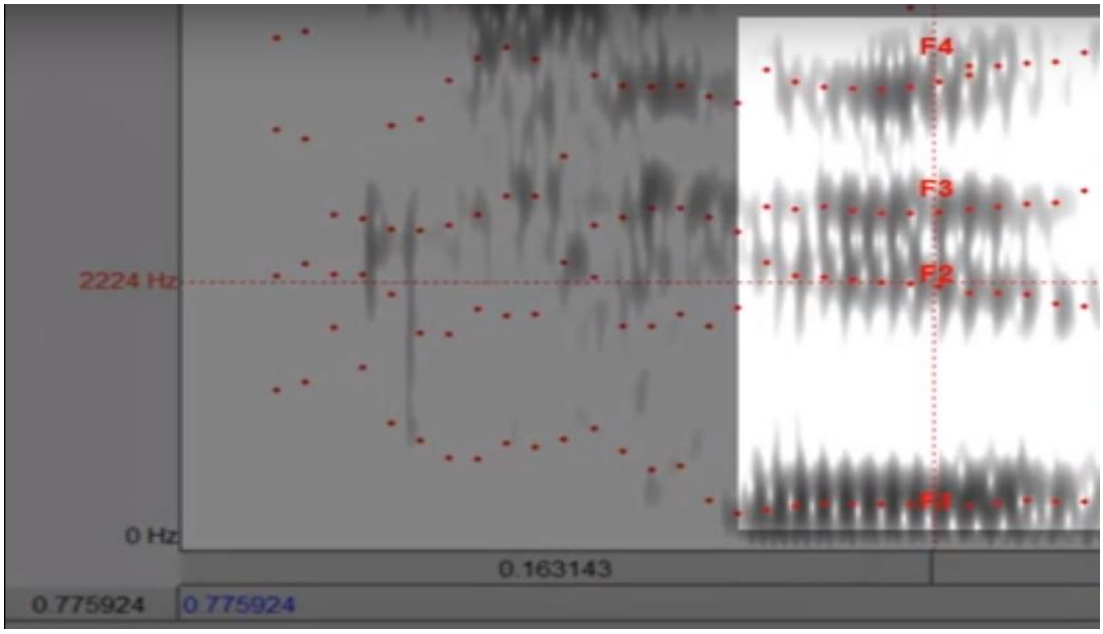


Figure- 5.10 (l): Formant of a speech signal

Time_s	F1_Hz	F2_Hz	F3_Hz	F4_Hz
0.896944	288.578652	1999.752403	2627.437169	3703.792735
0.903194	328.776232	2377.236961	2839.993394	3976.668789
0.909444	363.270055	2267.544926	2821.701333	3879.671972
0.915694	368.982948	2254.635314	2836.599128	3834.168962
0.921944	364.857185	2240.242444	2807.477726	3819.161963
0.928194	369.826204	2211.173995	2783.291246	3814.488684
0.934444	366.346189	2202.098228	2791.478781	3836.518507
0.940694	359.627986	2176.786136	2799.871748	3874.750275
0.946944	362.110690	2127.150653	2814.592841	3930.690113
0.953194	376.940620	2120.890676	2846.203697	4007.274589
0.959444	408.530031	2109.730391	2867.248497	4032.514961
0.965694	389.911624	2031.520963	2872.899571	4049.015374
0.971944	387.054771	2016.114800	2972.912232	4117.345651
0.978194	406.166715	1991.358602	3262.513585	4077.019800
0.984444	346.741412	1973.376933	2634.142926	3975.127987

Figure- 5.10(m): Formant History

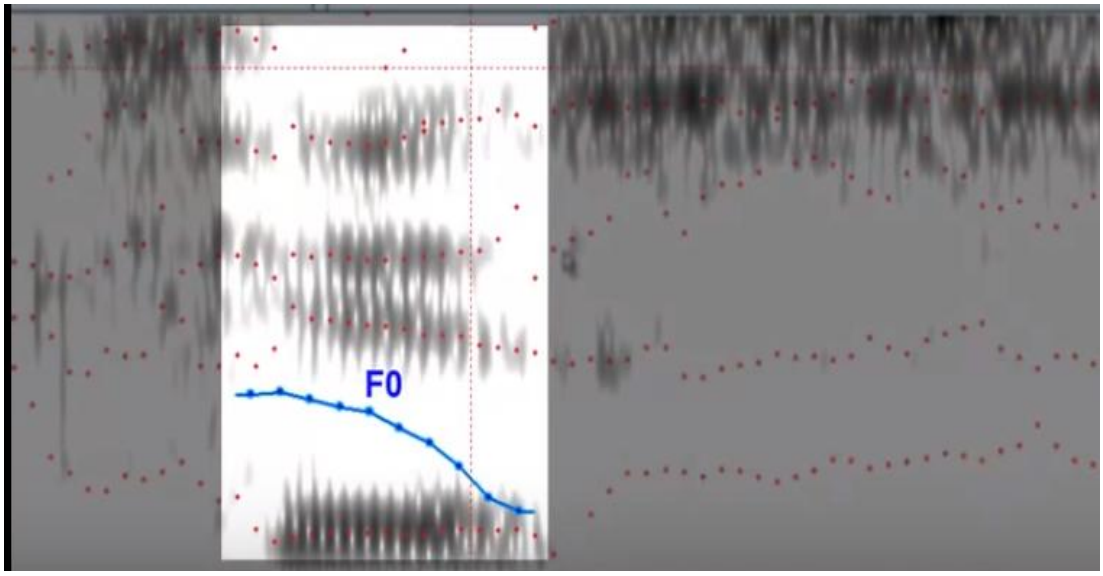


Figure- 5.10 (n): Pitch Analysis (Represented by Blue Lines)

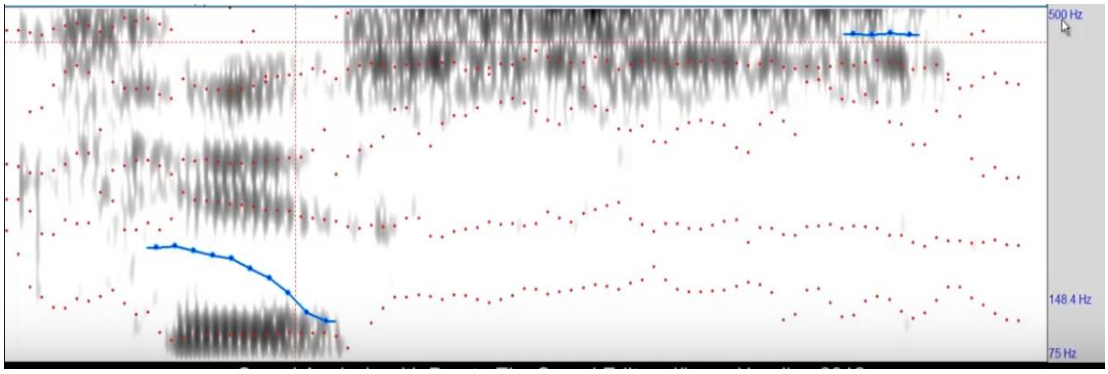


Figure- 5.10 (o): Range of Pitch for a Specific Word

Time_s	F0_Hz
0.898821	213.404682
0.908821	207.581655
0.918821	201.672129
0.928821	198.078240
0.938821	185.864576
0.948821	174.974048
0.958821	157.185632
0.968821	133.281639
0.978821	122.968320
0.988821	--undefined--

Figure- 5.10 (p): Example of Pitch Value

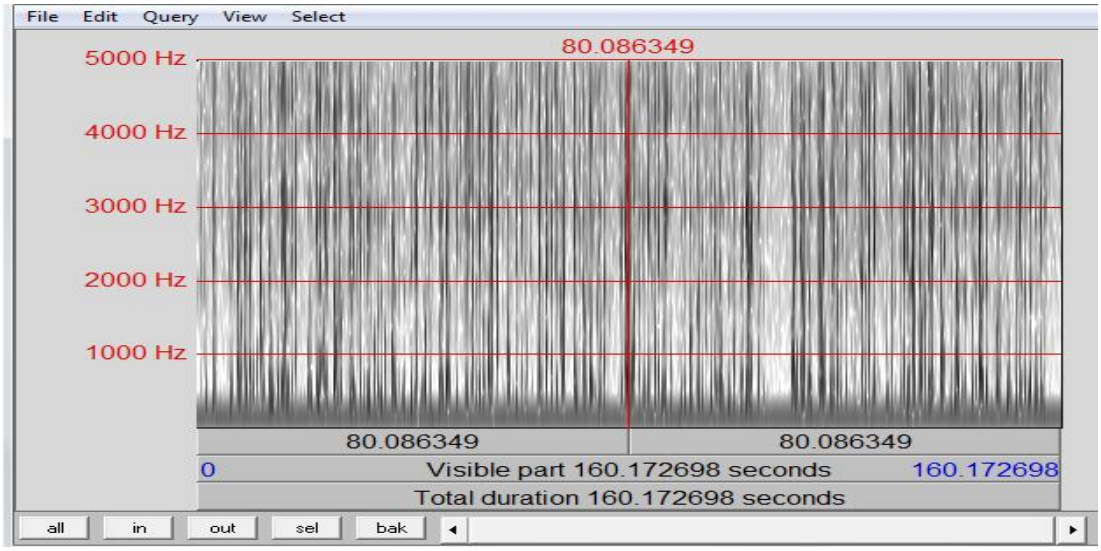


Figure- 5.10 (q): Spectrogram of a Speech Signal

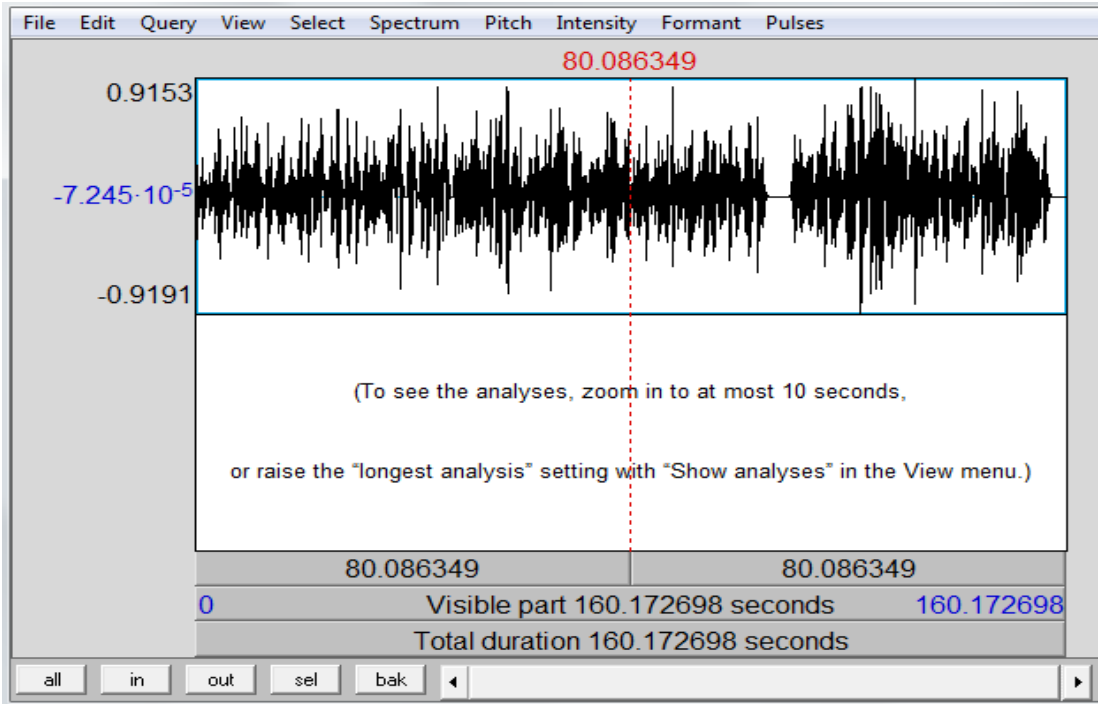


Figure- 5.10 (r): Noise Removal Process

5.11 Conclusion

In this chapter researcher explained and verified the results. In addition, a proposed system is tested for recognizing a speaker by using both Prosodic and Spectral (MFCC) features. Speech features are extracted at the frame level for creating speaker models. Results show that the proposed methodology gives better performance with Gaussian mixture models. In addition, the research provides more detailed information than many other proposed systems in the literature. Calculated results have been shown here in tabular as well as in graphical representation.

Results confirmed that speaker recognition system performance is improved by the proposed methodology. A significance improvement is achieved in system performance when prosodic features are used. Experiments are performed on the voice database created by the researcher in lab condition having background noise. The database contains the voice of male and female members both.

Chapter- 6
Validation of the Framework

CHAPTER-6: VALIDATION OF THE FRAMEWORK

To be or not to be, that is the question

-Shakespeare

6.1 Introduction

Validation of the research work is required to provide certain definite assurance for consistency, accuracy and robustness of an automated system. There are numerous available methodologies by using them data validation can be defined, designed and deploy in different context. Validation is the procedure to test the trustworthiness of any model/methodology/approach. Validation process assures that a product, system and service is fulfilling the required condition [244-246].There are some attributes which mostly includes in the process of validation such as:

- Accuracy
- System suitability
- Specificity
- Repeatability
- Reproducibility
- Quantification limit
- Detection limit (in case of trace element)

System suitability testing is an essential part of several analytical procedures. Validation tests are depending on the concept of analytical operations, samples to be analyzed, equipment and electronics. System suitability procedure test parameters to be established for a particular method depend on the type of method being validated. Model validation is not a decision however it is a necessity in a dynamic modeling development [247] [248]. The definitions of validation as stated by different authors are:

- Validation is the process of establishing confidence in the usefulness of a model (Coyle 1977).

- The process of determining the degree to which a model is an accurate representation of the real-world from the perspective of the intended uses of the model (DoD 2002).
- Substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model (Sargent 2003).

6.2 Methodology for Validation

Validation is a procedure to estimate the comparison between computational results from the simulation and the actual (hypothesis) data from the system. Primary goal of validation is identification and quantification of the error, uncertainty in the conceptual models, and calculation of the numerical error in the computational solution, evaluation of the simulation uncertainty, and at last, comparison between the computational results and the actual data. Therefore, accuracy is calculated in terms to real/hypothetical data. But this approach does not believe that the real/hypothetical data are more accurate than the computational results [249-250]. Some characteristics of validation are discussed below:

- A model should be evaluated for its effectiveness rather than its complete validity.
- No specific test by which the model validity can be checked
- A model cannot have complete validity however it should be fulfill the requirement for which it is constructed
- For validation of model qualitative and quantitative criterion provide more credence.
- For Validation to pass test does not assure that model is valid and failing a test help to reject a negative hypothesis.

Validation is the most incomprehensible part of developing a model, no model can be accepted without it has passed validation test. Validation is a process to determine the trustworthiness of the model and generally it is framework based and dynamic [251]. As author's says in [252-253] that validation is applied on all simulation models irrespective either corresponding system exists presently or would be developed in future. Validation of models using statistical techniques and reasoned that the technique applied for validation would depend on the availability of data in the real system. While some author says [254] [251] that there is complete valid

model, reliability of a model can be justified only for the proposed use of the model and the recommended conditions under which the model has been tested [254].

Tryout

For validation of the proposed approach an experimental tryout has been carried out. For data collection first we created a voice database for speaker's (male & female) after that feature extracted from speech signal using different feature extraction methods such as MFCC and Prosodic. Using feature extraction method accuracy of a speaker recognition system was calculated. Procedure of calculation using MFCC has been discussed in section 5.8.1 and using Prosodic has been discussed in section 5.8.2. Calculated values of recognition rate using MFCC, Prosodic have been given in table-6.2.

Training Language	**Testing Language**	**Recognition Rate (%) MFCC**	**Recognition Rate (%) Prosodic**
English	English	89.70	95.74
	Hindi	86.50	94.61
Hindi	English	86.50	94.61
	Hindi	89.70	95.74

Table- 6.2 Calculated values for MFCC and Prosodic

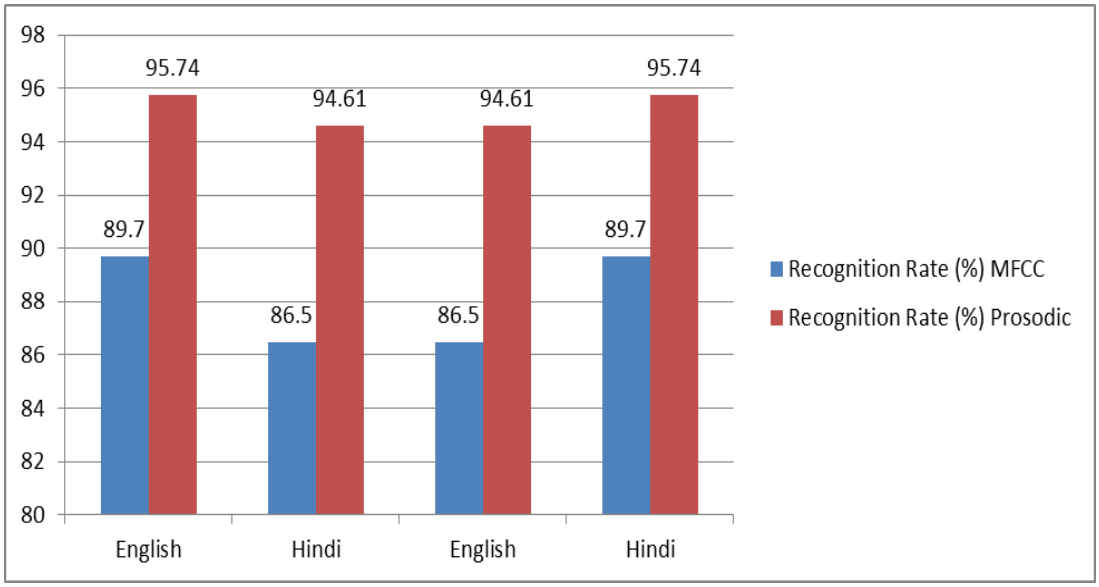


Figure- 6.2: Comparison Performance of Automatic Speaker Recognition System using Prosodic and MFCC

6.3 Statistical Analysis

Student's t- Test is one of the most commonly used techniques for testing a hypothesis on the basis of a difference between sample means [255]. As the number of observations is small, the student t-test is applied for finding out the level of significance and rejection of the null hypothesis. Since the rejection or acceptance of a null hypothesis is based upon either (0.05) alpha (α) or (0.01) alpha (α) level of significance for one tailed or two tailed test, (0.01) alpha (α) level of significance for a two tailed test is taken for rejection of the null hypothesis [256-257]. Statistical analysis process includes some sequential steps. Initially, null hypothesis and alternate hypothesis has been formulated and later statistical analysis has been performed. During statistical analysis, performance of developed approach, Prosodic is compared with MFCC. As a result of statistical analysis it has been concluded that whether there is significance difference between the previous methodologies (MFCC) and the developed Prosodic approach. The obtained t- value will determine whether to reject the null hypothesis and accept the alternative hypothesis.

- **Hypothesis Testing**

A null hypothesis reflects that there is no significant relationship between two or more parameters whereas alternate hypothesis affirms the relationship. Rejection of a null hypothesis provides a stronger base to accept the relationship or to accept the alternate hypothesis [258]. This study relates improvement of speaker recognition system performance by using prosodic features. During the research work a framework has been developed by using Prosodic features of speech and modeling by Gaussian mixture model.

- ✓ **Null Hypothesis (H_{01}):** Performance of speaker recognition system cannot be improved by using Prosodic.
- ✓ **Alternative Hypothesis (H_{11}):** Performance of speaker recognition system can be improved by using Prosodic.

- **Interpretation**

By observing calculated values in table 6.3 (a), it can be observed very easily that the prosodic features equal error rate shows that system performance is improved. The evaluated values for the recognition rate of Prosodic and MFCC are shown in table- 5.8.1 and table- 5.8.2. The result shows that table- 5.8.1 is comparatively improved than in the table- 5.8.2. Experimental values show that methodology followed for improving system performance is appropriate. Hence it has been conclude that performance of recognition system could be improved. Hence the initial claim that prosodic feature is able to improve speaker recognition system performance proved true. Figure- 6.3 shows the EER of MFCC and Prosodic.

Training Language	Testing Language	EER (MFCC)	EER (Prosodic)
English	English	10.3	4.26
	Hindi	13.5	5.39
Hindi	English	13.5	5.39
	Hindi	10.3	4.26

Table- 6.3 (a): Calculated values of EER for Speaker Recognition

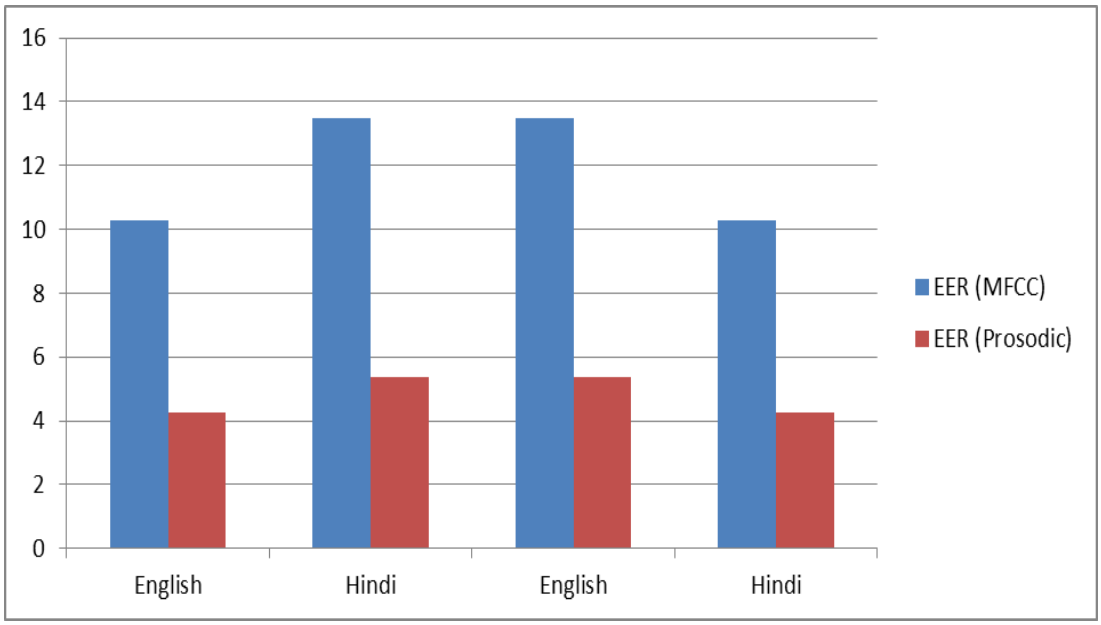


Figure- 6.3: EER of MFCC and Prosodic

- **Level of Significance of the proposed Framework**

To find out the significance difference between MFCC and Prosodic; the means of MFCC and Prosodic are calculated as shown in table- 6.3 (b). Pearson coefficient of correlation is 1. The degree of freedom is 4. For application of the t-test in the scenario, homogeneity of variances i.e. F value must be tested. The homogeneity can be obtained by dividing the larger variance by the lower. The large variance is 3.4134 for MFCC and the smaller one is 0.4256 for Prosodic.

The t value comes out to be -7.2218. As the value exceeds the t critical value of 0.0019 for a two tail test at the 0.01 level for 4 degree of freedom, the null hypothesis H_{01} is strongly rejected and the alternate hypothesis H_{11} is accepted. Hence it is validated that performance of speaker recognition system can be improved by using Prosodic features.

- **T-Test: Two-Sample Assuming Unequal Variances**

Statistical Observation	Prosodic	MFCC
Mean	4.825	11.9
Standard deviation	0.565	1.6
Variance	0.425633333	3.413333333
Observations	4	4
Pearson Coef. of Correlation(r)		1
Hypothesized Mean Difference	0	
Degree of Freedom (df)		4
t Stat		-7.221863395
P(T<=t) one-tail		0.000974887
t Critical one-tail		2.131846786
P(T<=t) two-tail		0.001949775
t Critical two-tail		2.776445105
Test for homogeneity of variances(F)		0.124697265624998

Table-6.3 (b): T-Test: Two-Sample Assuming Unequal Variances

Acceptance of any new approach by society or industry depends upon validation of that approach. It is the validation which proves the usefulness of the approach in society or in industry. For testing the usefulness of the integrated approach for recognition improvement rate of an Automatic Speaker Recognition System, a systematic validation is carried out.

6.4 Conclusion

The research aimed to investigate the advantages and drawbacks of the existing methodologies of the text-independent speaker verification system, and to propose methods that could lead to an improved system performance. Validation of newly developed method is the most essential phase of the research work. It is required to validate methods to make it acceptable to the society. In this chapter framework has been validated. Student t-test has been used to test hypothesis. Hypothesis testing is based on the concept of system performance improvement and it has been analyzed that performance of a recognition system can be improved by the proposed Prosodic features of speech.

Chapter- 7
Conclusion and Future Work

CHAPTER-7: CONCLUSION AND FUTURE WORK

"A conclusion is the place where you get tired of thinking."

- Arthur Bloch

7.1 Introduction

This chapter summarizes the key outcomes of the research and highlights the potential future work for feature extraction in the area of speaker recognition. The key challenges that have been noticed, and those remaining for the field has been discussed in addition to limitations of the work. In this research, speaker recognition from speech is studied using prosodic related speech features. It is assumed that speaker recognition process have certain speaker-specific characteristics, have an effect on the acoustic parameters of the source-filter model. In addition, it is well known fact that the uses of prosodic features enhance speaker's recognition system performance. The research proposed in the thesis has utilized the same fact. In the proposed research by using prosodic speech features statistical models have been created to be used to derive speaker models for speaker recognition process.

It has been shown that systems using prosodic and higher level features out performed standard systems, especially when the amount of available training data was may be noisy. This confirms the assumption that short-term cepstrum systems generally perform better is not always true since MFCC easily affected by noise [212]. However, Prosodic features of speech easily not affected by noise [209]. In addition, higher-level features also have the potential of increased robustness to channel variation, since lexical usage or temporal patterns do not change with the change of acoustic conditions.

7.2 Research Contribution

In the research, some basic but important factors which are a crucial part of the Automatic Speaker recognition (ASR) System such as Voice fundamental, sampling rate, pitch, equal error rate etc. have been discussed. These factors are used to design and construct an Automatic Speaker Recognition System. ASR is a technology by which a person can be recognized by his/ her voice using a recognition system where different comparisons are made with training and testing data. This technology is based on the voice features of a person which is most suitable to recognize a person or

authenticating the person. ASR system can be designed for Speaker Verification & Speaker Identification. Speaker Recognition/ voiceprint recognition is an example of Biometrics i.e. it is a type of technology using which person's physiological/ behavioral characteristics can be measured.

The main focus of the research has explained about the feature extraction techniques. Speaker Recognition System makes use of a system based on comparison method is use to recognize the people from his/her voice. Prosodic features analysis of a speech signal has been elaborated in detail. In addition Gaussian mixture model generally used modeling technique explained in detail. The research has provided an elaborated Gaussian Mixture Model (GMM) and its component for speech signal. It has been revealed that Gaussian Mixture Model is very much appropriate for creating speech features model of a speaker. For speaker recognition, Gaussian mixture model is an essential appliance of statistical clustering. The task effortlessly performed by humans is not effortless for machine or computers such as voice recognition or face recognition. Speaker recognition technology produces a solution, using which the computers/machines out performs than humans. For achieving the research objective of speaker recognition, the research has made the following contributions.

- **Literature survey on speaker recognition:** An exhaustive review has been done on speaker recognition techniques during the initial phases of research work. As the outcome of literature survey it has been concluded that speaker recognition primary focused on the speaker identification, text-dependent, text independent etc. based speaker recognition system. The contribution has been published in [20] [46] [190]. In the last several decades research in speech and speaker recognition has been going on worldwide. Since speech is the basic and the most suitable form of communication to convey message among people. The development of speaker recognition system shows the interest of human in technology. It is as the first step towards natural human machine communication. Some considerable advances in speech along with speaker recognition are not likely to come solely from research in statistical pattern recognition and signal processing. Number of practical limitations has been encountered which hinder widespread deployment of application and services.

- **Analysis of various tools/models/techniques used at various level of Speaker Recognition System development:** Speaker recognition using prosodic and Gaussian mixture models as a technique to provide efficient and accurate solutions to problems of speaker recognition. A review is about the prosodic and about the GMM for speaker recognition technology has been performed. The studies show that mixture model analysis yields a large number of theorems, methods, applications and test procedures. There is much related theoretical work as well as research is available on Gaussian mixture applications. The research has explored the same only for automatic speaker recognition technology. The Detailed description is presented in chapter 2.

- **Voice Database creation:** Database of recorded male and female speakers from different channel has been developed. To build a voice database there are lots of medium available for recording the voice/speech. In the research, we have recorded the speech by using head phone, MATLAB program and some external/portable voice recording devices. To record voice a head phone or any other medium of recording is needed using which voice of speaker is recorded. In addition to Computer/Laptop, there is also some external audio/video devices required to store voice database. At the time of recording the speaker's speech of 2-3 minutes length has been recorded sometimes it is vary up to 5 minutes. Detailed description of the same is presented in chapter 4.

- **Development of framework for speaker recognition system:** A new approach is developed for speaker identification system. The developed framework aims to extract speech features using Prosodic feature extraction technique. Prosodic features achieve better recognition rate by considering supplementary information sources. The aim is to design a speaker identification system, and apply it to a speech of an unknown speaker (text-independent). By investigating the extracted features of the unknown speech and comparing them to the stored speaker models for each different speaker, an effort is made to identify the unknown speaker. Prosodic showed a significant improvement in automatic speaker recognition system performance. The Detailed description is presented in chapter 3.

- **Implementation of the Proposed Framework for Improving Speaker Recognition System Performance:** The proposed framework is implemented in chapter 4 and all the phases of the framework are discussed in detail. Each and every phase of the proposed framework has been implemented and tested using MATLAB as well as Praat software. Speaker models are created by using Gaussian mixture modeling technique, and stored for training and testing purpose. After creation of speaker models matching is performed and on the basis of match score decision is made that either speaker is accepted or rejected. In addition system performance is evaluated on the basis of equal error rate metric. The detailed description is shown in chapter 4.

- **Validation of Prosodic feature extraction technique:** It was validated that features based on Prosodic provided better performance compare to the conventional MFCC. For validation of the proposed approach an experimental tryout has been carried out. For voice sample collection first we created a voice database for speaker's (male & female) after that feature extracted from speech signal using different feature extraction methods such as MFCC, Prosodic .Using feature extraction method performance of a speaker recognition system was calculated. The Detailed description is presented in chapter 6.

7.3 Significance of the Work

Automatic speaker recognition is used for the purpose of authentication to improve security of an automatic system. In today's scenario it is more useful because of its voice based biometric technique. The world-wide collaboration and free exchange of ideas led to technology boost in the speaker recognition field as well. Applications of speaker recognition system provide prominent alternatives to biometrics such as finger prints, retina scans and face recognition. The key advantage of speaker recognition over these techniques is being its low costs, not harmful to human body and non-invasive etc. The proposed technique can be successfully implemented in the following areas:

- Forensics Department
- Remote access control security

- Web services, online Calling
- Personalization of services and customer relationship management
- Voice based biometric system
- Transactions authentication/Voice based banking
- Surveillance/criminal investigation etc.

7.4 Future Direction

In future work we wish to work with many more prosodic features with the fusion of some other types of speech features. Also we will try to work with prosodic features with another modelling technique. To till date many recent advances have been achieved in the field of speaker recognition but there are still many problems remains unsolved for which good solutions need to be found. These problems mainly arise from speaker variability, channel variability, recording condition, background noise etc. Also it is key point to find feature parameters that are more useful to improve performance of speaker recognition system. The aim of finding such speech features that are stable over time, unaffected to the variation of speaking manner (including speaking rate and level) and robust against variation against voice quality such as voice disguise or cold, is a very important task. There is also need to develop a method to deal with the problem of distortion such as telephone sets and channel and background noises. In addition, we will try to find out the reason behind rejected voice (unmatched voice), i.e. it is occurs in due to testing database or training database.

7.5 Conclusion

Today's commercial applications of speaker recognition systems have become a reality. ASR systems have been adopted in both government and financial sectors as a method to facilitate quick and secure authentication of individuals. For example: the Australian Government organization Centrelink uses speaker verification for the authentication of Welfare recipients using telephone transactions. In India ICICI bank provide voice based banking to their customer.

Now days when everything is going to digitalized, biometric techniques is one of the latest technologies used for the security systems? There is no longer required to carry debit card, credit card or remembering password or PIN. While physical traits of a

person used instead of all. In this thesis, we have firstly investigated approaches to improve robustness of systems. As the selection of appropriate speech feature and feature extraction technique are still growing research areas, the proposed research work has carried out these tasks. We have introduced extended prosodic features for improving performance of automatic speaker recognition system. This approach will bring surely significant improvement in system performance.

REFERENCES

REFERENCES

1. U. Sandouk, "Speaker Recognition Speaker Diarization and Identification", PhD Thesis, University of Manchester, School of Computer Science, pp. 14-101, 2012.
2. M. Savvides, "Introduction to Biometric Technologies and Applications", ECE & CyLab, Carnegie Mellon University. http://www.biometricscatalog.org/ biometrics/biometrics_101.pdf or biometrics_101.pdf
3. B. S. Atal, "Automatic Recognition of Speakers from their Voices", Proc. IEEE, vol. 64(4), pp. 460-475, 1976.
4. Q. Jin, "Robust Speaker Recognition", PhD Thesis, Language Technologies Institute School of Computer Science Carnegie Mellon University, Pittsburgh, pp. 23-177, 2007.
5. A. Rajsekhar G., "Real Time Speaker Recognition using MFCC and VQ", PhD Thesis, Department of Electronics & Communication Engineering, National Institute of Technology Rourkela, pp. 9-71, 2008.
6. J. Luettin, "Visual Speech and Speaker Recognition", PhD Thesis, Department of Computer Science University of Sheffield, pp. 16-156, May 1997.
7. T. Kinnunen, "Spectral Features for Automatic Text-Independent Speaker Recognition", Licentiate's Thesis, University of Joensuu, Department of Computer Science Joensuu, Finland, pp. pp. 1-151, December 21, 2003.
8. M. R. Srikanth, "Speaker Verification and Keyword Spotting Systems for Forensic Applications", PhD Thesis, Department of Computer Science and Engineering Indian Institute of Technology Madras, pp. 1-135, Dec. 2013.
9. M. El. Ayadi, Abdel-Karim S.O. Hassan, Ahmed Abdel-Naby, Omar A. Elgendy, "Text-independent Speaker Identification using Robust Statistics Estimation", Speech Communication vol-92, pp. 52–63, 2017.
10. S. Sarkar, Sreenivasa, Rao, K., "Stochastic Feature Compensation Methods for Speaker Verification in Noisy Environments" Appl. Soft Comput. 19, pp. 198–214, 2014.
11. G. Doddington, Liggett, W., Martin, A., Przybocki, M., and Reynolds, D, "Sheeps, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation", In Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998) (Sydney, Australia), 1998.
12. S. Furui, "Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques", ESCA Workshop on Speaker Characterization in Speech Technology Edinburgh, Scotland, UK, pp. 10-27, June 26-28, 1990.

13. L. P. Cordella, P. Foggia, C. Sansone, M. Vento, “A Real-Time Text-Independent Speaker Identification System”, Proceedings of the ICIAP, pp. 632, 2003.

14. B. Richard Wildermoth, “Text-Independent Speaker Recognition Using Source Based Features”, PhD Thesis, Griffith University Australia, pp. 1-101, Jan. 2001.

15. M. Breen, L. C. Dilley, J. Kraemer, and E. Gibson, “Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)", Corpus linguist. ling., vol. 8 (2), pp. 277-312, 2012.

16. A. Stolcke, “Higher-Level Features in Speaker Recognition”, Winter School on Speech and Audio Processing, IIT Kanpur, Jan. 2009.

17. N. Dehak, D. Pierre, and K. Patrick, “Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 15 (7), pp. 2095-2103, Sept. 2007.

18. Z. Huang, L. Chen and M. Harper, “An Open Source Prosodic Feature Extraction Tool”, School of Electrical and Computer Engineering Purdue University West Lafayette, pp. 2116-2121, 2006.

19. T. Kinnunen and Haizhou Li, “An Overview of Text-Independent Speaker Recognition: From Features to Supervectors”, Speech Communication vol.52, pp.12–40, 2010.

20. N. Singh and Khan R. A., “Underlying of Text Independent Speaker Recognition”, in IEEE Conference (ID: 37465) (10th INDIACom 2016 International Conference on Computing for Sustainable Global Development), held on 16th -18th March, 2016 at BVICAM, New Delhi, pp. 11-15, 2016.

21. D. Sierra Rodriguez, “Text-Independent Speaker Identification”, PhD Thesis, AGH University of Science and Technology Krakow, Faculty of Electrical Engineering, Automatics, Computer Science and Electronics, pp. 1-121, 2008.

22. M. Ghassemian and K. Strange, “Speaker Identification - Features, Models and Robustness”, Technical University of Denmark, DTU Informatics Kongens Lyngby, Denmark, pp. 1-118, 2009.

23. D. A. Reynolds, “An Overview of Automatic Speaker Recognition Technology”, MIT Lincoln Laboratory, Lexington, MA USA, @2002IEEE, pp. 4072-4075, 2002.

24. A. Majetniak, “Speaker Recognition using Universal Background Model on YOHO Database”, Aalborg University, The Faculties of Engineering, Science and Medicine Department of Electronic Systems, pp. 1-61, May 31, 2011.

25. B. Wubishet , “ Noise Robust Speaker Verification using SVM based GMM Supervector”, PhD Thesis, Addis Ababa University, School of Graduate

Studies Addis Ababa Institute of Technology, Electrical and Computer Engineering Department, pp. 1-102, Dec. 2012.

26. D. A. Reynolds, “Experimental Evaluation of Features for Robust Speaker Identification”, IEEE Trans. Speech Audio Processing, vol. SAP-2, No. 4, pp. 639-643, 1992.

27. S. Patra, “Robust Speaker Identification System”, PhD Thesis, Super Computer Education and Research Centre, Indian Institute of Science, Bangalore, pp. 1-131, Dec. 2007.

28. B. S. Atal, “Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and verification”, Journal Acoustical Society of America, vol. 55, no. 6, pp. 1304-1312, June 1974.

29. G. Gravier and G. Chollet, “Comparison of Normalization Techniques for Speaker Verification”, In Proc. on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pp. 97-100, 1998.

30. C. D. Shaver and John M. Acken, “The Development of Text-Independent Speaker Recognition Technology”, Journal of Electrical Engineering, pp. 1-8, 2014.

31. J. Gonzalez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. Toledano and J. Gonzalez-Rodriguez, “ATVS-UAM NIST SRE 2010 System”, in Proc. of FALA 2010, pp. 415-418, 2010.

32. N. Scheffer and R. Vogt, “On the use of Speaker Super Factors for Speaker Recognition”, In Proc. of ICASSP2010, pp.4410-4413, 2010.

33. R. Schwartz, W. Shen, J. Campbell and R. Granville, “Measuring Typicality of Speech Features in American English Dialects: Towards Likelihood Ratios in Speaker Recognition Casework”, 5th European Academy of Forensics Science, Glasgow, Scotland, Sept. 8, 2009

34. N. Chen, W. Shen, J. Campbell and P. Torres- Carrasquillo, "Informative Dialect Recognition Using Context- Dependent Pronunciation Modeling", ICASSP 2011, Prague Czech Republic, May 2011.

35. N. Singh and Khan R. A., “Extraction and Representation of Prosodic Features for Automatic Speaker Recognition Technology”, Fifth International Conference on AITMC (AIM-2015), Proceedings of Advanced in Engineering and Technology, Published by: Mc Graw Hill Education, 2015, pp.1-7, ISBN-10:93-85965-79-4.

36. C. Shaver and J. Acken, “Effects of Equipment Variation on Speaker Recognition Error Rates”, IEEE, ICASSP-2010, Dallas Texas, March 2010.

37. J. Ming, T. Hazen, J. Glass and D. Reynolds, “Robust Speaker Recognition in Noisy Conditions”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 5, July 2007.

38. T. Kohler, “The 2010 NIST Speaker Recognition Evaluation”, SLTC Newsletter, July 2010.

39. R. D. Kent and C. Read, “The Acoustic Analysis of Speech”, Singular Publishing Group, Inc., SanDiego, CA, 1992.

40. E. Vayrynen, “Emotion Recognition from Speech Using Prosodic Features”, PhD Thesis, University of Oulu Graduate School, University of Oulu, Faculty of Information Technology and Electrical Engineering, Department of Computer Science and Engineering, Infotech Oulu, pp. 1-92, 2014.

41. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Speaker and Session Variability in GMM-Based Speaker Verification,” IEEE Trans. Audio Speech and Language Processing, vol. 15, no. 4, May 2007.

42. S. Nakagawa, W. Zhang, and M. Takahashi, "Text-Independent Speaker Recognition by Combining Speaker-Specific GMM with Speaker Adapted Syllable-based HMM," Acoustics, Speech, and Signal Processing, Proceedings (ICASSP '04), IEEE International Conference, vol.1, pp. I-81-4, 17-21 May 2004.

43. S. Furui, “An overview of Speaker Recognition Technology, in Automatic Speech and Speaker Recognition”, C. H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Boston: Kluwer Academic, pp. 31-56, 1996.

44. C. Liu, P. Jyothi, H. Tang, V. Manohar, M. Hasegawa-Johnson, and S. Khudanpur, “Adapting ASR for under-Resourced Languages using Mismatched Transcriptions”, In Proc. ICASSP, 2016.

45. W. Michael John, “Learning Models of Speaker Variation”, PhD Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, pp. 1-63, July 12, 1996.

46. N. Singh, A. Agrawal and Khan R. A., “Automatic Speaker Recognition: Current Approaches and Progress in Last Six Decades”, Global Journal of Enterprise Information System. Vol. 9, Issue-3, July-September, pp. 38-45, ISSN: 0975-1432, 2017.

47. M. Faundez-Zanuy and E. Monte-Moreno, "State-of-the-art in Speaker Recognition", Aerospace and Electronic Systems Magazine, IEEE, vol.20, no.5, pp. 7-12, March 2005.

48. B. Martínez-Gonzalez, M. Jose Pardo, D. Julian Echeverry-Correa, Ruben San-Segundo, “Spatial Features Selection for Unsupervised Speaker Segmentation and Clustering”, Expert Systems with Applications, vol. 73, pp. 27–42, 2017.

49. D. Talkin, “A Robust Algorithm for Pitch Tracking (RAPT)”, Speech Coding and Synthesis, pp. 495-518, 1995.

50. P. Boersma, “Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound”, In Proc. the Institute of Phonetic Sciences, 1993.

51. D. A. Reynolds, “Automatic Speaker Recognition: Current Approaches and Future Trends” MIT Lincoln Laboratory, Lexington, MA USA, pp. 1-6, 2001.

52. S. Memon, “Automatic Speaker Recognition: Modeling, Feature Extraction and Effects of Clinical Environment”, PhD Thesis, School of Electrical and Computer Engineering Science, Engineering and Technology Portfolio RMIT University, pp. 1-242, June 2010.

53. R. Summerfield, T. Dunstone, C. Summerfield, “Speaker Verification in a Multi-Vendor Environment”, www.w3.org/2008/08/siv/Papers/Centrelink/w3c-sv_multivendor.pdf

54. K. A. Toh, “Fingerprint and Speaker Verification Decisions Fusion”, In Proceedings of 12th International Conference on Image Analysis and Processing, pp.626-631, 2003.

55. J. Kerin, “Biometric Passport Demand Likely”, 2004, www.news.com.au.

56. A. Michele Cavazza and C. Alberto, "Device for Speaker's Verification”, " http://www.google.com/patents/US4752958?hl=it&cl=en

57. International Banking (December 27, 2013). "Voice Biometric Technology in Banking | Barclays". Wealth.barclays.com. Retrieved February 21, 2016.

58. R. G. Hautamaki, Md. Sahidullah, V. Hautamaki and T. Kinnunen, “Acoustical and Perceptual Study of Voice Disguise by Age Modification in Speaker Verification”, Speech Communication, DOI: 10.1016/j.specom.2017.10.002, pp. 1-37, 2017.

59. K. Julia, "HSBC Rolls out Voice and Touch ID Security for Bank Customers | Business", The Guardian, Retrieved February 21, 2016.

60. "Speaker Identification". *Archived from* the original *on August 15, 2014*. Retrieved September 3, 2014.

61. N. Singh, Khan R. A. and Raj shree, “Applications of Speaker Recognition” Science Direct, Procedia Engineering, vol-38, pp. 3122-3126, 2012.

62. S. Furui, “50 years of Progress in Speech and Speaker Recognition”, Department of Computer Science, Tokyo Institute of Technology, pp. 1-9, 2004.

63. S. Furui, “Speech-to-text and Speech-to-Speech Summarization of Spontaneous Speech”, IEEE Trans., Speech & Audio Processing, vol. 12, I. 4, pp. 401-408, 2004.

64. Z. Shi-xiong, M. Man-wai and M. M. Helen, “Speaker Verification via High-Level Feature-based Phonetic-Class Pronunciation Modeling”, IEEE Transactions on Computers, vol. 56, No. 9, Sep 2007.

65. N. Singh, "A Study on Speech and Speaker Recognition Technology and Its Challenges", Proceedings of National Conference on Information Security Challenges. Lucknow: DIT, BBAU, year 2014, pp. 34-37.

66. D. Clark Shaver and J. M. Acken, “A Brief Review of Speaker Recognition Technology”, Electrical and Computer Engineering Faculty Publications and Presentations. 350. http://pdxscholar.library.pdx.edu/ece_fac/350, pp. 1-7, 2016

67. D. A. Reynolds, T.F. Quatieri, and R. B. Dunn, “Speaker Verification using Adapted Gaussian Mixture Models”, Digital Signal Processing, vol. 10, no. 1-3, pp. 19–41, 2000.

68. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Factor Analysis Simplified,” In Proc. ICASSP 2005, Philadelphia, PA, pp. 637–640, March 2005.

69. V. Wan and W. M. Campbell, “Support Vector Machines for Speaker Verification and Identification,” In IEEE International Workshop on Neural Networks for Signal Processing, Sydney, Australia, vol. 2, pp. 775–784, 2000.

70. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, “A Lognormal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition”, In Proc. EUROSPEECH, Rhodes, Greece, pp. 1391–1394, Sept. 1997.

71. A. Mansour, A. Mahmood and G. Muhammad, “ Speaker Recognition based on Arabic Phonemes”, Elsevier, Speech Communication, vol. 86, pp. 42–51, 2017

72. S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg A. Stolcke, H. Bratt, and R. R. Gadde, “Speaker Recognition using Prosodic and Lexical Features,” In Proc. IEEE ASRU, St-Thomas, pp. 19–24, Dec. 2003.

73. S. Kajarekar, L. Ferrer, K. Sonmez, J. Zheng, E. Shriberg, and A. Stolcke, “Modeling NERFs For Speaker Recognition,” in Proc. Odyssey 2004, Toledo, Spain, pp. 51–56, June 2004.

74. Y. Dauphin, H. de Vries and Y. Bengio “Equilibrated Adaptive Learning Rates for Non-Convex Optimization”, In Advances in Neural Information Processing Systems, pp. 1504–1512, 2015.

75. L. Ferrer, H. Bratt, S. Kajarekar, E. Shriberg, K. Sonmez, K. Stocke, and A. Venkataraman, “Modeling Duration Patterns For Speaker Recognition,” In Eurospeech, Geneva, pp. 2017–2020, 2003

76. K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, “Modeling Dynamic Prosodic Variation for Speaker Verification,” In Proc. ICSLP 1998, Sydney, Australia, pp. 2631–2634, 1998.

77. L. Sbattella, L. Colombo, C. Rinaldi, R. Tedesco, M. Matteucci, and A. Trivilini, “Extracting Emotions and Communication Styles from Prosody” , Springer-Verlag Berlin Heidelberg 2014, H.P. da Silva et al. (Eds.): Phy. CS 2014, LNCS 8908, pp. 21–42, 2014.

78. L. Ferrer, N. Scheffer and E. Shriberg, “A Comparison of Approaches for Modeling Prosodic Features in Speaker Recognition”, ©2010 IEEE, ICASSP 2010, pp. 4414-4417, 2010.

79. K. Sreenivasa Rao, “Role of Neural Network Models for Developing Speech Systems”, Sadhana, vol. 36, Part 5, pp. 783–836, Oct. 2011. @ Indian Academy of Sciences.

80. L. Kersta, “Voiceprint Identification”, Nature Magazine, 196, 1253, December 1962.

81. J. Lindh, “Handling the Voiceprint Issue”, FONETICK Proceedings, 2004.

82. R. Potter, G. Kopp and H. Green, “Technical Aspects of Visual Speech”, Bell Labs, New York, 1947.

83. Available At: http://www.cs.tut.fi/sgn/arg/intro/basics/seiskaspec.jpg

84. G. Doddington, “Speaker Recognition – Identifying People by their Voice”, Proceedings of IEEE, vol.73, 1651-1664, Nov. 1985.

85. L. Baum, T. Petrie, G. Soules, N. Weiss, “A Maximization Technique in the Statistical Analysis of Probabilistic Functions of Markov Chains”, Annals of Mathematical Statistics, vol. 41, No.1, 1970.

86. F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Chagnoleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz and D. Reynolds, “A Tutorial on Text-Independent Speaker Verification”, EURASIP Journal on Applied Signal Processing vol. 4, pp. 431-450, 2004.

87. J. Pelecanos and S. Sridharan, “Feature Warping for Robust Speaker Verification”, In Proc. IEEE Odyssey, ISCA Speaker Recognition Workshop, pp. 213–218, 2001.

88. T. Matsui, and S. Furui, “Comparison of Text-Independent Speaker Recognition using VQ-Distortion and Discrete/Continuous HMMs”, Proceedings of ICSLP, pp. 157-160, 1997.

89. G. Rodriguez and Joaquin, "Evaluating Automatic Speaker Recognition Systems: An Overview of the NIST Speaker Recognition Evaluations (1996-2014)", Loquens 1.1, 2014.

90. Biometrics, March 2015, Available: https://en.wikipedia.org/wiki/Biometrics.

91. N. Krishnamurthy, and J. H. L. Hansen, “Babble Noise: Modeling, Analysis, and Applications”, IEEE Trans. Audio Speech Lang. Process, vol.17 (7), pp. 1394–1407, 2009.

92. A. Jain, L. Hong and S. Pankanti, "Biometric Identification", Communications of the ACM vol. 43, pp. 91-98, 2000.

93. Speaker Recognition, March 2016, Available: https://en.wikipedia.org/wiki/Speaker Recognition,

94. P. Rose, "Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence", Elsevier, Computer Speech and Language vol. 20, pp. 159–191, 2006.

95. Speaker recognition, 2016, Available: TheFreeDictionary.com.

96. G. Michael, M. Mark P. and A. Lawrence Presley, "Recovering and Examining Computer Forensic Evidence", 2010, Available: https://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/oct2000/index.htm/computer.htm

97. D. Meuwly, "Forensic Individualization from Biometric Data", Science & Justice vol. 46 no. 4, pp. 205 – 213, 2006.

98. J. Bishop, "Using the Concepts of 'Forensic Linguistics,' 'Bleasure' and 'Motif' to Enhance Multimedia Forensic Evidence Collection", The 2014 International Conference on Security and Management SAM'14, Monte Carlo Resort in Las Vegas, Nevada USA, pp. 21-24, 2014.

99. R. Gonzalez H., T. Kinnunen, V. Hautamaki, and L. Anne-Maria, "Automatic versus Human Speaker Verification: The Case of Voice Mimicry", Speech Communication vol.72, pp. 13–31, 2015.

100. Biometrics Institute, Australia, 2016, Available: http://www.biometricsinstitute.org/pages/types-of-biometrics.html

101. D. Vincent, "Biometric-Security-Devices.com", year 2011-2016, Available: http://www.biometric-security-devices.com/types-of-biometric-devices.html

102. B. Aleksandra, "Biometric Authentication: Types of Biometric Identifiers", Bachelor's Thesis, Degree Programme in Business Information Technology University of Applied Sciences, pp. 1-56, 2013.

103. Pollack, Pickett and Sumby, "Experimental Phonetics", MSS Information Corporation, 1974, pp. 251–258, 1974

104. V. Lancker and Kreiman, "Familiar Voice Recognition: Patterns and Parameters", Part I: Recognition of Backward Voices" Journal of Phonetics, pp. 19–38, 1984.

105. Surveillance, August 2016, Available: https://en.wikipedia.org/wiki/Surveillance.

106. S. Soyuj K., M. Prasanna, S. R. Choubisa and Tarun, "Multimodal Biometric Person Authentication: A Review", IETE Technical Review vol. 29 no.2, pp. 54, 2012.

107. N. Singh, A. Agrawal and Khan R. A., "A Critical Review on Automatic Speaker Recognition", Science Journal of Circuits, Systems and Signal Processing. vol. 4 no. 2, pp. 14-17, July 2015.

108. E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," Speech Communication, vol. 46 no. 3-4, pp. 455–472, 2005.

109. G. Fant, "Acoustic theory of Speech Production", Mouton and Co., The Hague, Netherlands, 1970.

110. B. Healy Tukey, "The Quefrency Alanysis of the Time Series for Echos: Cepstrum, Pseudo-Auto-Covariance, Cross-Cepstrum, and Saphe Cracking" in Time Series Analysis, ch.15, pp. 209-243, 1963.

111. J. W. Cooley, and J. W. Tukey, "An Algorithm for the Machine Computation of Complex Fourier Series" Math Computation, vol. 19, Apr. 1965, pp.297-301.

112. A.M. Noll, "Cepstrum Pitch Determination", Journal of Acoustical Society of America, vol. 41, pp. 293-309, 1969.

113. A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech", IEEE, Trans. on Audio and Electroacousitcs, vol. 16:2, pp. 221-226, June 1968.

114. R. W Schafer, L. R. Rabiner, "Digital Representation of Speech", Invited Paper in Proceedings of the IEEE, Vol. 63:4, pp. 662-667, April 1975.

115. R. Vogt, and S. Sridharan, "Frame-weighted bayes Factor Scoring for Speaker Verification", In Proc. 10th Australian Int. Conf. on Speech Science & Technology (Sydney, Australia), pp. 404-409, 2004.

116. J. P. Campbell, "Speaker Recognition: A Tutorial", Proceeding of the IEEE, 85:1437–1462, Sept. 1997.

117. S. Furui, "Recent Advances in Speaker Recognition", Pattern Recognition Letters, 18:859–872, 1997.

118. R. Togneri and D. Pullella, "An overview of Speaker Identification: Accuracy and Robustness Issues", Circ. Syst. Mag. IEEE 11 (2), pp. 23–61, 2011.

119. D. A. Reynolds, "Automatic Speaker Recognition: Acoustics and Beyond", In Super SID project at JHU Summer Workshop, 2002. http://www.clsp.jhu.edu/ws2002/groups/supersid/.

120. S. Furui, "An Overview of Speaker Recognition Technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, pp. 1-9, April 1994.

121. H. Melin, "Speaker Verification in Telecommunication", Department of Speech, Music and Hearing, KTH, pp. 1-10, 1999.

122. S. Furui, “Speech and Speaker Recognition Evaluation”, Evaluation of Text and Speech Systems, Springer, pp. 1-27, 2007.

123. F. Weber, L. Manganaro, B. Peskin and E. Shriberg, “Using Prosodic and Lexical Information for Speaker Identification”, Proc. ICASSP, vol. 1, Orlando, pp. 141–144, 2002.

124. Available At: http://www.biometric-solutions.com/

125. Y. Ergun and V. V. Nabiyev, “A New Approach with Score-Level Fusion for the Classification of a Speaker Age and Gender”, Computers and Electrical Engineering, vol. 53, pp. 29–39, 2016.

126. A. Tolunay, “Text-Dependent Speaker Verification Implemented in MATLAB using MFCC and DTW”, PhD Thesis, Department of Electrical Engineering, Linkoping University, pp. 1-73, 2010.

127. X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, “Spoofing Detection from a Feature Representation Perspective,” In Processing of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2016.

128. A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, “Joint Speaker Verification and Antispoofing in the-Vector Space,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 821–832, 2015.

129. O. Othman Khalifa, S. Khan, Md. Rafiqul Islam, M. Faizal and D. Dol, “Text Independent Automatic Speaker Recognition”, 3rd International Conference on Electrical & Computer Engineering ICECE, Dhaka, Bangladesh, pp. 28-30, 2004.

130. J. A. Bachorowski and M.J. Owren, “Acoustic Correlates of Talker Sex and Individual Talker Identity are present in a Short Vowel Segment Produced in Running Speech”, Journal of Acoust. Soc. Am., vol. 106 no. 2, pp. 1054-1063, 1999.

131. N. Singh, Khan R. A. and Raj Shree. "Equal Error Rate and Audio Digitization and Sampling Rate for Speaker Recognition System." American Scientific Publishers. vol. 20 numbers 5-6, pp. 1085-1088, May 2014.

132. D.A. Reynolds and R.C. Rose, “Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models,” IEEE Trans. Speech Audio Processing, vol. 3 no. 1, pp. 72-83, 1995.

133. Stolcke, S. Kajarekar, L. Ferrer, and E. Shriberg, “Speaker Recognition with Session Variability Normalization Based on MLLR Adaptation Transforms”, IEEE Transactions on Audio, Speech, and Language Processing, 15(7), pp. 1987-1998, 2007.

134. A. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, “Modeling Prosodic Dynamics for Speaker Recognition”, In Proceedings of the IEEE

International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, pp. 788-91, 2003.

135. S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt and V. R. R. Gadde, “Speaker Recognition using Prosodic and Lexical Features”, In: Proceedings of the IEEE Speech Recognition and Understanding Workshop, St. Thomas, U.S. Virgin Islands, pp. 19–24, 2003.

136. E. Blaauw, “The Contribution of Prosodic Boundary Markers to the Perceptual difference between Read and Spontaneous Speech”, Speech Communication, vol. 14, pp. 359–375, 1994.

137. E. Shriberg, L. Ferrer, S. Kajareker, A. Venkataraman, “SVM Modeling of SNERF-grams for Speaker Recognition”, in: Proceedings of Inter Speech International Conference on Spoken Language Processing, JEJU Island, Korea, pp. 1-4, 2004.

138. L. Mary and B. Yegnanarayana, “Extraction and Representation of Prosodic Features for Language and Speaker Recognition”, Speech Communication, vol. 50, pp. 782–796, 2008.

139. E. Variani , X. Lei , E. McDermott , I. Lopez Moreno and J. Gonzalez-Dominguez , “Deep Neural Networks for small Footprint Text-Dependent Speaker Verification”, in: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference, pp. 4052–4056, 2014 .

140. F. Z. Thomas, “Speaker Recognition Systems: Paradigms and Challenges”, Center for Speech and Language Technologies, Tsinghua University Asia-Pacific signal and Information Processing Association, 2012-2013. Available: http://www.apsipa.org

141. J. Bartosek, “The use of Prosody in Speech Recognition Systems, Punctuation Detector for czech Speech", English, in Kraliky, Brno, CZ, 2010, pp. 24-27, 2010 ISBN: 978-80-214-4139-2.

142. A. Larcher, K. A. Lee, B. Ma, and Li. H., “Text-dependent Speaker Verification: Classifiers, Databases and RSR2015”, Speech Commun. 60, pp.56–77, 2014.

143. W. Rabah A. and E. Fuad Al-Saadi, “Text-Independent Speaker Identification in Noisy Environment using Singular Value Decomposition”, ICICS-PCM, Singapore, IEEE, pp. 1-5, 2003.

144. A. Fazel and S. Chakrabartty, “An overview of Statistical Pattern Recognition Techniques for Speaker Verification”, IEEE Circuits and System Magazine, pp. 62-81, 2011.

145. K.A. Lee, “The 2015 NIST Language Recognition Evaluation: the Shared View of I2R, Fantastic4 and SingaMS", in: Proc. Interspeech, pp. 3211-3215, 2016.

146. H. L. John, Hansen, and T. Hasan, “Speaker Recognition by Machines and Humans: A Tutorial Review”, IEEE Signal Processing Magazine, vol. 74, pp. 74-99, Nov. 2015.

147. B. Frederic, L. Mathan, “Second- Order Statistical Measures for Text-Independent Speaker Identification”, ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, pp. 51-54, April 1994.

148. P. Krishnamoorthy, “Speaker Recognition under Limited Data Condition by Noise Addition”, Expert Systems with Applications, Elsevier, Vol. 38, pp. 13487–13490, 2011.

149. D. A. Reynolds, “Speaker Identification and Verification using Gaussian Mixture Speaker Models”, Speech Communication, vol.17, pp. 91-108, 1995.

150. R. Dunn, T. Quatieri, D. A. Reynolds and J. Campbell, “Speaker Recognition from Coded Speech in Matched and Mismatched Conditions”, In Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001) (Crete, Greece, 2001), pp. 115-120, 2001.

151. B. P. Yuhas, M. H. Goldstein and T. J. Sejnowski, “Integration of Acoustic and Visual Speech Signals using Neural Nets”, IEEE Communication, Mag. pp. 65-71, Nov. 1989.

152. N. Balakrishnan, “Improved Text-Independent Speaker Recognition using Gaussian Mixture Probabilities”, A Report in Candidacy for the Degree of Master of Science (M.S.) Department of Electrical and Computer Engineering Carnegie Mellon University, pp. 1-46, May 2005.

153. K. Kim and M. Young Kim, “Robust Speaker Recognition against Background Noise in an Enhanced Multi-Condition Domain”, IEEE Transactions on Consumer Electronics, vol. 56, 3, pp. 1684-1688, Aug. 2010.

154. D. A. Reynolds, “Gaussian Mixture Models”, Encyclopedia of Biometrics, pp. 659-663, 2009.

155. R. Rajeshwara Rao, A. Prasad and Ch. Kedari Rao, "Robust Features for Automatic Text-Independent Speaker Recognition Using Gaussian Mixture Model", International Journal of Soft Computing and Engineering (IJSCE), vol. 1, pp. 330-335, Nov. 2011.

156. Q. D’Almeida, Frederico, A. O. Francisco, Nascimento, Pedro A. Berger, and L. M. da Silva. "Automatic Speaker Recognition with Multi-resolution Gaussian Mixture Models (MR-GMMs)", The International Journal of Forensic Computer Science, pp.9-21.

157. L. Liang, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian Mixture Models for Speech Recognition", Signal Processing Letters, IEEE, vol. 18, no. 7, pp. 419-422, July 2011.

158. S. O. Sadjadi, J. Pelecanos and S. Ganapathy, “The IBM Speaker Recognition System Recent Advances and Error Analysis”, In: Proc. Interspeech, pp.3633–3637, 2016.

159. M.W. Mak and H. B. Yu, “A Study of Voice Activity Detection Techniques for NIST Speaker Recognition Evaluations”, Comput. Speech Lang., vol. 28(1), pp. 295–313, 2014.

160. O. Glembek, J. Ma, P. Matejka, B. Zhang, O. Plchot, L. Burget and S. Matsoukas, “Domain Adaptation via within-class Covariance Correction in i-vector based Speaker Recognition Systems”, In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4032–4036, 2014.

161. Y. Tian, Z. Chen and F. Yin, “Distributed Iterated Extended Kalman Filter for Speaker Tracking in Microphone Array Networks”, Applied Acoustics, vol. 118, pp. 50–57, 2017.

162. G. Ye, B. Mak and M. Mak, “Fast GMM Computation for Speaker Verification using Scalar Quantization and Discrete Densities”, In: Proc. Int. Speech Commun. Association (Interspeech), pp. 2327–2330, 2009.

163. X. Li, M. Mak and S. Kung, “Robust Speaker Verification over the Telephone by Feature Recuperation”, In Proc. 2001 Int. Symposium on Intelligent Multimedia, Video, and Speech Processing (Hong Kong, 2001), pp. 433–436, 2001.

164. J. F. Bonastre, F. Bimbot, L.J. Boe, J.P. Campbell, D.A. Reynolds, and I. Magrin-Chagnolleau, “Person Authentication by Voice: A Need for Caution,” Proc. Of Eurospeech, ISCA, Geneva, Switzerland, pp. 33-36, 1-4 Sept. 2003.

165. F. Soong and A. Rosenberg, “On the use of Instantaneous and Transitional Spectral Information in Speaker Recognition”, IEEE Trans. on Acoustics, Speech and Signal Processing vol. 36 (6), pp. 871–879, 1998.

166. M. Zilovic, R. Ramachandran and R. Mammone, “Speaker Identification based on the use of Robust Cepstral Features Obtained from Pole-zero Transfer functions”, IEEE Trans. on Speech and Audio Processing vol. 6 (3) pp. 260–267, 1998.

167. W. Majewski, and G. Mazur-Majewska, “Speech Signal Parametrization for Speaker Recognition under Voice Disguise Conditions”, In Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999) (Budapest, Hungary), pp. 1227–1230, 1999.

168. L. D. Van, “Speaker Verification Systems and Security Considerations”, In Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003) (Geneva, Switzerland), pp. 1661–1664, 2003.

169. C. Bishop, “Pattern Recognition and Machine Learning”, s.l. : Springer Science Business Media, pp. 435-441, 2006.

170. Available at: https://www.google.com/patents/EP2048656B1?cl=en

171. M. Alfredo et.al., “Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models”, Journal of Information Security, vol. 3, pp. 335-340, 2012.

172. W. A. Rabah and E. Fuad Al-Saadi, “Robust Text-independent Speaker Recognition with Short Utterance in Noisy Environment Using SVD as a Matching Measure”, J. King Saud Univ., Comp. & Info. Sci., vol. 17, pp. 23-41, 2004.

173. T. H. Kinnunen, “Optimizing Spectral Feature Based Text-Independent Speaker Recognition”, PhD Thesis, University of Joensuu Computer Science Dissertations 12, pp. 1-156, 2005.

174. Available At: http://www.biometric.com/

175. A. Afolabi, A. Williams and O. Dotun, “Development of a Text-dependent Speaker Identification Security System”, Res. J. Appl. Sci. 2 (6), pp. 677–684, 2007.

176. R. Nakatsu, J. Nicholson, and N. Tosa, “Emotion Recognition and its Application to Computer agents with Spontaneous Interactive Capabilities”, Knowledge based systems, vol. 13, pp. 497–504, 2000.

177. Morrison, G. Stewart, and H. Selby, “Forensic voice comparison”, Thomson Reuters, 2010.

178. S. Ganapathy et.al., “Robust Feature Extraction using Modulation Filtering of Autoregressive Models”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 8, pp. 1285-1295, Aug. 2014.

179. T. Hesham, “A High-performance Text-independent Speaker Identification of Arabic Speakers using a CHMM-based Approach” Alexandria Engineering Journal, vol. 50, pp. 43–47, 2011.

180. C. Chih-Hung et. al. “A Statistical Out-of-Speaker Detection Approach for Smart Home Voice-Control Scenario of Protective Warming Care on FPGA”, ASE BD & SI, Kaohsiung, Taiwan, pp. 1-4, Oct. 2015.

181. P. Emmanual et. al., “Phonatory Signature of the Deaf Child”, ESCA Workshop on Automatic Speaker Recognition”, Identification and Verification, Martigny, Switzerland, pp. 201-204, April 1994.

182. A. Lozano-Diez, A. Silnova, P. Matejka, O. Glembek, O. Plchot, J. Pesan, L. Burget and J. Gonzalez-Rodriguez, “Analysis and Optimization of Bottleneck Features for Speaker Recognition”, In: Proceedings of Odyssey, The Speaker and Language Recognition Workshop, pp. 21–24, 2016.

183. S. Xuanjing et. al., “A Speaker Recognition Algorithm Based on Factor Analysis”, 7th International Congress on Image and Signal Processing, 978-1-4799-5835-1/14 ©2014 IEEE, pp. 897-901, 2014.

184. S.M. Anzar et. al., “Efficient Online and Offline Template Update Mechanisms for Speaker Recognition”, Computers and Electrical Engineering (Elsevier) vol. 50, pp. 10–25, 2016.

185. I. Sanchez-Cortina et. al., “Speaker-adapted Confidence Measures for Speech Recognition of Video Lectures”, Computer Speech and Language (Elsevier), vol. 37, pp. 11–23, 2016.

186. Available at: http://www.biometric-solutions.com/performance-of-biometrics.html

187. S. K. Achintya, Md. Sahidullah, T. Zheng-Hua and T. Kinnunen, “Improving Speaker Verification Performance in Presence of Spoofing Attacks Using Out-of-Domain Spoofed Data”, Conference paper, pp. 1-6, Aug. 2017.

188. K. Abbas and M. H. Mohammad, “A PLDA Approach for Language and Text-independent Speaker Recognition”, Computer Speech & Language vol.45, pp. 457-474, 2017.

189. S. Yang and W. De Liang, “Robust Speaker Identification Using Auditory Features and Computational Auditory Scene Analysis”, Department of Computer Science and Engineering, Center for Cognitive Science, The Ohio State University, Columbus, ICASSP, 1-4244-1484-9/08, IEEE, pp.1589-1592, 2008

190. N. Singh, A. Agrawal khan R. A., “Gaussian Mixture Model: A Better Modeling Technique for Speaker Recognition”, Proceedings of (DIAL-2017) National Conference on Digital India-Altering Landscape, Tracking the Journey to Digitally Empowered New India, 9th -10th Dec 2017, Organized by CSI Lucknow Chapter, pp. 14-18, 2017. ISBN: 978-93-5291-226-1.

191. D. Garcia-Romero and A. McCree, “Supervised Domain Adaptation for i-vector based Speaker Recognition”, In: Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4047–4051, 2014.

192. P. Matejka, O. Glembek, O. Novotny, O. Plchot, F. Grezl, L. Burget and J. H. Cernocky, “Analysis of DNN Approaches to Speaker Identification”, In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5100–5104, 2016.

193. Woodland, M. J. F. Gales and P. C., “Mean and Variance Adaptation within the MLLR Framework”, Computer Speech and Language, vol. 10, 4, pp. 249–264, 1996.

194. Available At : FrameworkDef/What%20is%20framework_%20-%20Definition%20from%20WhatIs.com.html

195. J. H. Kunzel and A. Paul, "Forensic Automatic Speaker Recognition with Degraded and Enhanced Speech", Journal of the Audio Engineering Society, vol. 62 (4), pp. 244-253, 2014.

196. P. Rose, “Forensic Speaker Identification”, Taylor & Francis, London, 2002.

197. J. Wolf, “Efficient Acoustic Parameters for Speaker Recognition” Journal of the Acoustical Society of America vol. 51, No. 6, pp.2044-2056, 1972.

198. T. F. Quatieri, D. A. Reynolds, and G. C. O’Leary, “Estimation of Handset Nonlinearity with Application to Speaker Recognition”, IEEE Transactions on Speech and Audio Processing, pp. 567–584, 2000.

199. L. R. Rabiner and B. H. Juang, “Fundamentals of Speech Recognition”, Prentice Hall, NJ, 1993.

200. S. Pillay, “Voice Biometrics under Mismatched Noise Conditions”, PhD Thesis, “University of Hertfordshire”, pp. 19-26, 2010.

201. Y. Lei, L. Burget, L. Ferrer, M. Graciarena and N. Scheffer, “Towards Noise Robust Speaker Recognition Using Probabilistic Linear Discriminant Analysis”, In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4253–4256, 2012.

202. X. Sun and Y. Miyanaga, “Dynamic time Warping for Speech Recognition with Training part to Reduce the Computation”, International Symposium on Signals, Circuits and Systems, (ISSCS 2013), Iasi, Romania, pp. 1-4, July 11-12, 2013, DOI 10.1109/ISSCS.2013.6651195.

203. P. Motlicek, S. Dey, S. Madikeri, L. Burget, “Employment of Subspace Gaussian Mixture Models in Speaker recognition”, In: International Conference on Acoustics, Speech and Signal Processing, pp. 4445–4449, 2015.

204. R. Saeidi, T. Kinnunen, H. Sadegh Mohammadi, R. Rodman and P. Franti, “Joint Frame and Gaussian Selection for Text-Independent Speaker Verification”, In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 4530–4533, 2010.

205. S. Jelil, R. K. Das, R. Sinha, S. M. Prasanna, “Speaker Verification using Gaussian Posteriorgrams on Fixed Phrase Short Utterances”, In: Interspeech, pp. 1042–1046, 2015.

206. Available At: http://zone.ni.com/reference/en-XX/help/371361H-01/lvanlsconcepts/char_smoothing_windows/

207. A. E. Rosenberg and S. Parthasarathy, “Speaker Background Models for Connected Digit Password Speaker Verification”, In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, USA, 1996.

208. H. Gish and M. Schmidt, “Text Independent Speaker Identification”, IEEE Signal Processing Magazine, vol. 11 (4), pp. 18-32, 1994.

209. S. Cumani, “Fast Scoring of full Posterior PLDA Models”, IEEE/ACM Trans. Audio Speech Lang. Process, vol. 23(11), pp. 2036–2045, 2015.

210. C. Longworth, “Kernel Methods for Text-Independent Speaker Verification”, PhD Thesis, Department of Engineering, Cambridge University and Christ's College, pp. 38-56, 2010.

211. V. C. Sharada and S. C. Mahesh, “Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition”, Procedia Computer Science, vol. 58, pp. 272 – 279, 2015.

212. Tur, E. Shriberg, A. Stolcke, and S. Kajarekar,” Duration and Pronunciation Conditioned Lexical Modeling for Speaker Recognition”, Proc. Eurospeech, pp. 2049-2052, 2007.

213. S. Verboven and M. Hubert, “Libra: A MATLAB Library for Robust Analysis”, Chemometr. Intell. Lab. Syst, vol. 75 (2), pp. 127–136, 2005.

214. H. Aronowitz, “Inter Dataset Variability Compensation for Speaker Recognition”, In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4002–4006, 2014.

215. Available at: Overall Speaker Models, pp. 73-95.

216. B. Yegnanarayana and S. Kishore, “AANN: An Alternative to GMM for Pattern Recognition”, Neural Networks, vol. 15, pp. 459-469, 2002.

217. T. Kemp, M. Schmidt, M. Westphal and A. Waibel, “Strategies for Automatic Segmentation of Audio Data”, In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul: IEEE, pp. 1423–1426, 2000.

218. C. Knapp and G. C. Carter, “The Generalized Correlation Method for Estimation of Time Delay”, IEEE Trans Acoust. Speech Signal Process, vol. 24(4), pp. 320–327, 1976.

219. H. R. S. Mohammadi and R. Saeidi, “Efficient Implementation of GMM based Speaker Verification using Sorted Gaussian Mixture Model”, Proc. Eur. Signal Process. Conf. (EUSIPCO), pp. 4–8, 2006.

220. A. Kanagasundaram, D. Dean and S. Sridharan, “Improving out-domain PLDA Speaker Verification using Unsupervised Inter-dataset Variability Compensation Approach”, In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4654–4658, 2015.

221. K. Asawa, V. Verma and A. Agrawal, “Recognition of Vocal Emotions from Acoustic Profile”, In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics, 2012.

222. L. Ferrer, “Prosodic Features Extraction”, Technical report, SRI, 2002.

223. H. Mattias, E. Jens and B. Tomas, “Automatically Extracted F0 Features as Acoustic Correlates of Prosodic Boundaries”, Proceedings, FONETIK, Dept. of Linguistics, Stockholm University, pp. 1-4, 2004.

224. H. Nakasone, “Automated Speaker Recognition in Real World Conditions: Con-trolling the Uncontrollable”, In Proc. 8th European Conf. on Speech Communication and Technology (Eurospeech 2003) (Geneva, Switzerland), pp. 697-700, 2003.

225. H. Fujisaki and K. Hiros, “Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese”, In journal of acoustical society of Japan (E), vol. 5 (4), pp. 233-241, 1984.

226. P. Boersma and D. Weenink, “Praat: Doing Phonetics by Computer [Computer program]”, Version 5.4.09, retrieved 15 June 2015 from http://www.praat.org/.

227. M. Brookes, “Voicebox: Speech Processing Toolbox for MATLAB Software”, 2006, available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

228. L. Anolli and R. Ciceri, “The Voice of Emotions”, Angeli, Milano, 1997.

229. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, “Front-end Factor Analysis for Speaker Verification”, IEEE Trans. Audio Speech Lang. Process, vol. 19 (4), pp. 788–798, 2011.

230. E. Shriberg and A. Stolcke, “Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing”, In International Conference on Speech Prosody, 2004.

231. D. Hirst and R. Espesser, “Automatic Modeling of Fundamental Frequency using a Quadratic Spline Function”, Travaux del’Institut de Phonétique d’Aix, vol. 15, pp. 75–85, 1993 [http://www.icp.inpg.fr/~loeven/Praat/momel_english.html]

232. A. Leemann and M. J. Kolly, “Speaker-invariant Supra-segmental Temporal Features in Normal and Disguised Speech”, Speech Communication vol. 75, pp. 97-122, 2015.

233. R. Djemili, R. Bourouba and Korba, “A Speech Signal based Gender Identification System using Four Classifiers”, In: Multimedia computing and systems (ICMCS), 2012 International Conference on. IEEE, pp. 184–187, 2012.

234. N. G. Ward and W. Tsukahara, “Prosodic Features which cue Backchannel Responses”, in English and Japanese. J. Pragmat, vol. 32, pp. 1177–1207, 2000.

235. X. Sun, “A Study on Efficient Robust Speech Recognition with Stochastic Dynamic Time Warping”, PhD Thesis, Graduate School of Information Science and Technology Hokkaido University Sapporo, Hokkaido, Japan, pp. 1-113, July 17, 2014.

236. Available At: http://web.science.mq.edu.au/~cassidy/comp449/html/ch09.html

237. W.H. Equitz, “A New Vector Quantization Clustering Algorithm”, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 37(10), pp. 1568-1575, Oct. 1989.

238. D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny and E. Lleida, “Unscented Transform for i-vector-based Noisy Speaker Recognition”, In : Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4042–4046, 2014.

239. NIST 2016, “Speaker Recognition Evaluation” Plan, pp. 1-8, August 4, 2016, http://www.nist.gov/itl/iad/mig/sre16.cfm

240. Ing. Jan Bartosek, “Prosody Utilization in Continuous Speech Recognition” , PhD Thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Circuit Theory, pp. 1-165, Aug. 2016.

241. A. S. Safi, “Text-Independent, Automatic Speaker Recognition System Evaluation With Males Speaking Both Arabic And English”, B.S., King Saud University, Faculty of the Graduate School of the University of Colorado, pp. 1-60, 2015.

242. Z. Yan, et. al. “Deception Detecting from Speech Signal using Relevance Vector Machine and Non-Linear Dynamics Features”, Elsevier Science Direct, Neuro Computing , vol. 151, pp.1042–1052, 2015.

243. R. H. Bolt, F. S. Cooper, E. E. David, Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens, “Speaker Identification by Speech Spectrograms: A Scientists’ View of Its Reliability for Legal Purposes”, J. Acoust. Soc. Amer., vol. 47 (2), pp. 597–612, 1970.

244. DOI:www/validation/Verificationandvalidation/Wikipedia/free/encyclopedia.html

245. M. S. Martis, “Validation of Simulation Based Models: A Theoretical Outlook”, The Electronic Journal of Business Research Methods vol. 4 (1), pp.39 -46, 2006.

246. AIAA, “Guide for the Verification and Validation of Computational Fluid Dynamics Simulations”, [online], American Institute of Aeronautics and Astronautics, AIAA-G-077- , 1998

247. C. Muller, “Speaker Classification I - Fundamentals, Features, and Methods”, vol. 4343. Springer Berlin/Heidelberg, 2007.

248. B. Ziotko, et. al. "Hybrid Wavelet-Fourier-HMM Speaker Recognition", Department of Electronics, AGH University of Science and Technology Krak, pp. 19-27, 2014

249. J. D. Sterman, “Appropriate Summary Statistics for Evaluating the Historic Fit of System Dynamics Models”, Dynamica, vol 10, pp. 51-66, 1984.

250. J.P.C. Kleijnen, “Validation of Models: Statistical Techniques and Data Availability”, Proceedings of the 1999 Winter Simulation Conference, P. A. Farrington, H. B. Nembhard, D. T. Sturrock and G. W Evans, (eds.), pp.647-654, 1999.

251. J.D. Sterman, Business Dynamics, “Systems Thinking and Modeling for a Complex World”, IRWIN McGraw-Hill, pp.845, 2000.

252. DMSO (Defense Modeling and Simulation Office). “The Principles of Verification, Validation, and Accreditation”, [online], Verification, Validation, and Accreditation Recommended Practices Guide, www.dmso.mil/public, U.S. Department of Defense, Office of the Director of Defense Research and Engineering, Nov. 1996.

253. D. A. Reynolds, "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models", The Lincoln Laboratory Journal, vol. 8 (2), pp. 173-192, 1995.

254. S. S. Stevens and J. Volkmann, “The Relation of Pitch to Frequency: A Revised Scale”, The American Journal of Psychology, vol. 53 (3), pp. 329-353, 1940.

255. DOI: Student's%20t%20Test%20(For%20Independent%20Samples).html

256. DOI: www.excel-easy.com/ /t-Test

257. R. Gray, "Vector Quantization", IEEE Magazine on Acoustics Speech and signal Processing, vol. 1, pp. 4-29, 1984.

258. F.K. Soong, et. al., “A Vector Quantization Approach to Speaker Recognition”, AT & T Technical Journal, vol.66 (2), pp.14-26, 1987.

259. S. K. Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, “On the Vulnerability of Speaker Verification to Realistic Voice Spoofing”, In Proc. of Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6, 2015.

GLOSSARY

GLOSSARY

Speaker Identification: Correctly identify the speaker, out of a certain set, of a certain speech.

Speaker Modeling: Different techniques used to build a model to retain enough information about the speaker identity.

Speaker Recognition SR: Any task related to automatically deal with the speaker identity.

Speaker Segmentation: Automatically identify turn points of speakers in a large stream.

Speaker Verification: Correctly verify that two segments are uttered by the same speaker.

SRE: Speaker Recognition Evaluation

Text-Dependent SR: Any task of SR that the speakers are restricted to a certain speech.

Text-Independent SR: Any task of SR that does not restrict the speech.

DNA Features: Speaker specific features extracted using the deep network architecture.

EER: Equal Error Rate, a special point on the ROC curve, where the miss ratio is equal to the false alarm ratio.

Filter Bank: A set of filters to cover the whole range of frequency found in the signal.

Formants: The higher frequency boundaries for a voiced part of speech.

Fourier Transformation: A method to decompose a real life signal into a number of sinusoids.

Frequency: The number of times the signal repeats itself in a time unit.

Fundamental Frequency: The frequency at which the vocal cords vibrate.

Gaussian Mixture Model: A mixture of more than one normal distribution, combined with different weights.

Klatt's formant synthesizer: A source/filter model based synthesizer proposed by Klatt.

Liner Predictive Coding: Compression technique using linear combinations and residual.

Mel-Scale: A non-linear scale based on human frequency perception.

NIST: National Institute for Science and Technology

Nyquist Frequency: The highest sampling frequency that can faithfully preserve the original signal.

Period: The time a periodic signal needs to repeat itself.

Periodic Signal: A signal that repeat itself every period T.

Pitch: The human perception of the fundamental frequency.

ROC Curves: Receiver Operating Characteristic, a plot to show the effect of different parameter values in a parameterized method.

Signal Filtering: Affecting certain frequency domains of the signal.

Signal Windowing: Extracting parts of the signal in the time domain.

Source/filter model: A simple method to model the human voice production by using a source

(base frequency generator) and a filter (filter bank to shape the final voice).

Speaker Clustering: Group different speech segments according the speaker.

Speaker Detection: Correctly flag the speeches, out of a set of speech, of a certain speaker.

Speaker Diarization: Detect who is speaking when on a stream.

Average: A difference equation to approximate a low-pass filter.

Beam Forming: An algorithm to align of many streams of the same recording.

Butterworth Filter: A filter written in the different equation format.

Cepstrum domain: A sample domain which is obtained when applying frequency analysis to the frequency domain.

Difference Equation: An approximate to a filter as a combination of previous input/output samples.

Differentiator: A difference equation to approximate a high-pass filter

Dirac Impulse: A signal consisting of one infinitely high impulse.

Divergence: A measure of the distance between two statistical models.