BENCHMARK FOR SPEAKER IDENTIFICATION UNDER ELECTRONIC VOCAL DISGUISE USING MEL FREQUENCY CEPSTRAL COEFFICENTS

Barry Ramya Mahathi

Register No: 11SLP004

A Dissertation Submitted in Part Fulfillment of Final Year
Master of Science (Speech-Language Pathology)
University of Mysore, Mysore.

ALL INDIA INSTITUTE OF SPEECH AND HEARING
MANASAGANGOTHRI
MYSORE – 570 006
May, 2013

DEDICATED TO

Amma,

Grand Parent's, &

Savithri Mam

CERTIFICATE

This is to certify that this dissertation entitled ***"Benchmark for speaker identification under electronic vocal disguise using Mel frequency Cepstral Coefficients"*** is the bonafide work submitted in part fulfillment for the Degree of Masters of Science (Speech-Language Pathology) with Registration No. 11SLP004. This has been carried out under the guidance of a faculty of this institute and has not been submitted earlier to any other University for the award of any other Diploma or Degree.

Mysore
May, 2013

Prof. S.R. Savithri
Director
All India Institute of Speech & Hearing
Manasagangothri, Mysore – 57006

CERTIFICATE

This is to certify that this dissertation entitled ***"Benchmark for speaker identification under electronic vocal disguise using Mel frequency Cepstral Coefficients"*** has been prepared under my supervision and guidance. It is also certified that this has not been submitted earlier in any other University for award of any Diploma or Degree.

Guide

Mysore
May, 2013

Prof. S.R. Savithri
Director
All India Institute of Speech & Hearing
Manasagangothri, Mysore – 57006

DECLARATION

This is to certify that this dissertation entitled ***"Benchmark for speaker identification under electronic vocal disguise using Mel frequency Cepstral Coefficients"*** is the result of my own study under the guidance of Prof. S. R. Savithri, Director, All India Institute of Speech and Hearing, Mysore, and has not been submitted earlier in any other University for the award of any Diploma or Degree.

Mysore, **Register No. 11SLP004**
May, 2013

ACKNOWLEDGEMENTS

I thank Almighty God for being with me always and giving me strength and courage to face all the challenges in life. Thank you God for leading me to the right path always.

I express my deepest gratitude and heartfelt thanks to my guide and director of AIISH Professor S. R. Savithri, for her constant guidance, support and patience throughout the study. Mam'm you are my inspiration and I am so lucky to be your student.

I thank my dearest Amma, wonderful women on this earth who dedicated entire life to the well being of me love u so much amma.

I thank my grandparent's Satya kesava for their love, affection, encouragement and financial support throughout my education without you I may not be here.

I thank Dr.Geetha Mam HOD of speech sciences department for her support and advice to complete the Dissertation on time.

I thank Dr.T.V. Anananthapadmanabha for his guidance and support in completing the Dissertation.

I thank Dr.Jaya kumar, Mr. Raja Sudhakar and Ms.Sahana for their cooperation, guidance and technical help throughout the study.

I thank all the staff in the department for their kind cooperation and support throughout the study.

I thank my one and only little sister Manjeera for her encouragement and entertainment but I miss the time what we were before love you darling.

I thank all my batch mates for their kind understanding, encouragement and for boosting up the confidence levels in me all through the masters program.

I thank all the inmates especially my batch mates (Jyothsna, Thulasi, Ceana, Shilpa, Ankita, Tanu) and Pari Didi in Kapila Ladies Hostel for making my stay comfortable and i had very special and sweet memories with you all people which I couldn't forget in lifetime.

I sincerely thank all the subjects in the study for their willingness to participate in the study and for their kind cooperation.

I would like to thank the Librarian and all the staff in the library for their on time help.

Finally I thank all my professors, lecturers, clinical supervisors, seniors, and my batch mates for constant support and encouragement throughout the masters program. Am so lucky to be a student of AIISH.

Table of contents

Chapter	Title	Page No.
	Table of contents	i
	List of tables	ii – iii
	List of figures	iv
I	Introduction	1-10
II	Review of literature	11- 55
III	Method	56 - 61
IV	Results	62 - 78
V	Discussion	79 - 80
VI	Summary and conclusions	81 - 87
	References	88 - 92

List of Tables

Sl. No.	Title	Page No.
1.	Illustration of Euclidean distance within and between subjects (Least Euclidian distance in bold).	60
2.	Diagonal matrix for normal condition (MFCC) for HPI of /a:/.	63
3.	Diagonal matrix for normal condition (MFCC) for LPI of /a:/.	63
4.	Diagonal matrix for normal condition (LPCC) for HPI of /a:/.	64
5.	Diagonal matrix for normal condition (LPCC) for LPI of /a:/.	64
6.	Diagonal matrix for normal condition (MFCC) for HPI of /i:/.	64
7.	Diagonal matrix for normal condition (MFCC) for LPI of /i:/.	65
8.	Diagonal matrix for normal condition (LPCC) for HPI of /i:/.	65
9.	Diagonal matrix for normal condition (LPCC) for LPI of /i:/.	65
10.	Diagonal matrix for normal condition (MFCC) for HPI of /u:/.	66
11.	Diagonal matrix for normal condition (MFCC) for LPI of /u:/.	66
12.	Diagonal matrix for normal condition (LPCC) for HPI of /u:/.	66
13.	Diagonal matrix for normal condition (LPCC) for LPI of /u:/.	67
14.	Percent correct identification for Normal condition.	67
15.	Diagonal matrix for female disguise condition (MFCC) for HPI of /a:/.	68
16.	Diagonal matrix for female disguise condition (MFCC) for LPI of /a:/.	68
17.	Diagonal matrix for female disguise condition (LPCC) for HPI of /a:/.	68
18.	Diagonal matrix for female disguise condition (LPCC) for LPI /a:/.	69
19.	Diagonal matrix for female disguise condition (MFCC) for HPI of /i:/.	69
20.	Diagonal matrix for female disguise condition (MFCC) for LPI of /i:/.	69
21.	Diagonal matrix for female disguise condition (LPCC) for HPI of /i:/.	70
22.	Diagonal matrix for female disguise condition (LPCC) for LPI of /i:/.	70
23.	Diagonal matrix for female disguise condition (MFCC) for HPI of /u:/.	70

24. Diagonal matrix for female disguise condition (MFCC) for LPI of /u:/. 71

25. Diagonal matrix for female disguise condition (LPCC) for HPI of /u:/. 71

26. Diagonal matrix for female disguise condition (LPCC) for LPI of /u:/. 71

27. Percent correct identification for female disguise condition. 72

28. Diagonal matrix for child disguise condition (MFCC) for HPI of /a:/. 73

29. Diagonal matrix for child disguise condition (MFCC) for LPI of /a:/. 73

30. Diagonal matrix for child disguise condition (LPCC) for HPI of /a:/. 73

31. Diagonal matrix for child disguise condition (LPCC) for LPI of/a:/. 74

32. Diagonal matrix for child disguise condition (MFCC) for HPI of /i:/. 74

33. Diagonal matrix for child disguise condition (MFCC) for LPI of /i:/. 74

34. Diagonal matrix for child disguise condition (LPCC) for HPI of /i:/. 75

35. Diagonal matrix for child disguise condition (LPCC) for LPI of /i:/. 75

36. Diagonal matrix for child disguise condition (MFCC) for HPI of /u:/. 75

37. Diagonal matrix for child disguise condition (MFCC) for LPI of /u:/. 76

38. Diagonal matrix for child disguise condition (LPCC) for HPI of /u:/. 76

39. Diagonal matrix for child disguise condition (LPCC) for LPI of /u: /. 76

40. Percent correct identification for child disguise condition. 77

41. Diagonal Matrix of Normal’s Vs Female Disguise condition for three vowels on MFCC. 77

42. Diagonal Matrix of Normal’s Vs Female Disguise condition for three vowels on LPCC. 77

43. Diagonal Matrix of Normal’s Vs Child Disguise condition for three vowels on MFCC. 78

44. Diagonal Matrix of Normal’s Vs Child Disguise condition for three vowels on LPCC. 78

45. Percent correct identification scores. 80

46. Percent correct identification scores. 87

List of figures

Sl. No.	Title	Page No
1.	Schematic representation of speaker recognition.	12
2.	Schematic representation of speaker identification.	13
3.	Schematic representation of speaker verification.	14
4.	Classification of errors.	16
5.	Mel filtering [Milner, 2003].	44
6.	Mel Frequency spaced filter banks.	45
7.	Euclidean distance and graphic comparison of feature vectors (LPCC and MFCC).	58-59

CHAPTER I

Introduction

The identification of people by their voices is a common practice in day-to-day life. We Identify persons by listening to their voices, over a phone line, radio, among other devices. If the person is familiar to us, we can identify her/him by the tone of the voice, the style of speaking, and so on. If we do not know her/him, we can still infer some characteristics like gender, age, emotional state and language, among others.

It is a worldwide growing tendency that perpetrators are inclined to disguise their voices in order to conceal their identities, especially in cases of threatening calls, extortion, kidnapping and even emergency police calls. Voice disguise is defined as a deliberate action of a speaker who wants to change his voice for the purpose of falsifying and concealing identity. Now-a-days many possibilities are available for speakers to change and distort their voices and to confuse recognition by the human ear to automatic system. For example whisper, falsetto, foreign accent, change of speaking rate, imitation, pinched nostril and object in mouth are all favourite techniques for the perpetrators.

In recent days crime rate is increasing rapidly mainly through telecommunication means especially due to increased usage of mobile phones. Simultaneously misuse of using mobile phones is also increased in many cases like bomb threat, ransom demand, sexual abuse, and hoax emergency call; in these cases **voice** is the only source available for analysis. Thus, the Forensic Speaker Identification came into existence.

Forensic Speaker Identification is seeking an expert opinion in the legal process as to whether two or more speech samples are of the same person. According to Rose (1992) Speaker recognition can be speaker identification and speaker verification. Speaker identification is deciding if a speaker belongs to a group of known speaker population. ***Speaker Identification*** is carried out by a combination of auditory and acoustic methods (and, where appropriate, some text and discourse analysis methods) and provides an opinion as to whether a particular voice, for example recorded making a telephone call, or participating in a conversation recorded by a recording device, is that of a particular known person. ***Speaker verification*** is verifying the identity claim of the speaker. Or it is defined as deciding if a speaker is whom he claims to be. This is different than the speaker identification problem, which is deciding if a speaker is a specific person or is among a group of persons. The speaker verification is also termed as voice verification, speaker authentication, voice authentification, talker authentication, and talker verification.

Speaker recognition methods can be divided into *text-independent* and *text-dependent* methods. In a text-independent system, speaker models capture characteristics of somebody's speech which show up *irrespective of what one is saying*. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her *speaking one or more specific phrases* (Rabiner, 1993). Every technology of speaker recognition, identification, and verification, whether text-independent or text-dependent, has its own advantages and disadvantages and may require different treatments and techniques. The choice of which technology to use is application-specific. At the highest level, all speaker recognition systems contain two main modules - *feature extraction* and *feature matching.*

Individual`s identity verification is an essential requirement for controlling access to protected resources. Personal identity is usually claimed by presenting a unique personal possession such as a key, a badge, or a password. However, these can be lost or stolen. Further a simple identity claim is not sufficient if the potential for loss is great and the penalty for false identification is severe. Hence verification of that claimed identity is necessary. This can be attempted by examining an individual's ***biometric features***, such as finger prints, hand geometry, or retinal pattern, or by examining certain features derived from the individual's unique activity such as ***speech*** or hand writing. In each case, the features are compared with previously stored features for the person whose identity is being claimed. If this comparison is favourable based on decision criterion, then the claimed identity is verified. In the present era of widely used telephone, mobile phone, radio, and tape recorder communication, the only information available to investigators may consist of a single voice recording, generally made during a telephone/mobile phone conversation. Among these methods, identity verification based on a person's voice has special advantages for practical deployment. Speech is our most natural means of communication and therefore user acceptance of the system would be very high. Apart from access, speaker identification is also used in forensic cases. Therefore, there is a pressing need on the part of police and magistrates for establishment of legal proof of identity from measurements of the voice. In view of this, the considerable interest in obtaining reliable techniques for speaker identification and in using these as the basis for such proof is easily understood.

Speaker Identification may be requested for a number of different criminal offences, such as making genuine or hoax emergency service calls to the police, ambulance or fire brigade, making threatening or harassing telephone calls, blackmail or extortion demands, taking part in criminal conspiracies such as those involving the importation, trafficking or manufacture of illegal drugs, or conspiring to traffic in people, arms, currency, and cultural artifacts. Speaker Identification may also be required in civil cases or for the media. These cases include calls to radio stations, local or other government authorities, insurance companies, or recorded conversations, rallies or meetings.

In speaker identification, the speech sample in question and control may suffer from the problems of noisy and poor quality recordings, vocal disguise, non contemporary, different text, different language and also electronic scrambling such as Voice synthesizers, Text to Speech converter and the almost limitless potential applications of speech processing in modern communication systems and networking such as Voice Over Internet Protocol (VOIP). There are several factors which affect speaker identification task. They are as follows:

Uniqueness: The identification task might involve an open set of trials. Specifically, the unknown must be detected from within a large to very large population of 'possibilities'. But this can be overcome to some extent so, that we can reduce the number of possibilities by taking in to consideration, the gender, dialect, language, some common phrases used and style of speaking by the speaker. However, it becomes very difficult to identify a speaker by his/her voice, especially when they are talking in an environment which distorts or masks their utterances (channel distortions) or when they are excited or stressed (speech distortions).

System distortion includes several kinds of signal degradation. One is reduced frequency response, i.e., the signal pass band can be limited when someone talks over a telephone line or mobile phone, poor quality tape recorders are used to 'store' the utterances and / or microphones of limited capability are employed. In these cases, the important information about the talker is lost and these elements are not usually retrievable. Such limited signal pass band can reduce the number of helpful speaker specific acoustic factors. Second, noise can create a particularly debilitating type of system distortion as it tends to make the talker`s voice and, therefore, can obscure elements needed for identification. Examples of noise are those created by wind, motors, fans, automobile movement and clothing friction. The noise itself may be intermittent or steady state saw tooth or thermal and so on. Third, any kind of frequency or harmonic distortion can also make the task of identification more difficult. Examples include intermittent short circuits, variable frequency response, and harmonic distortion and so on.

The speaker themselves can be the source of many types of distortions ***- Speaker distortion***. Fear, anxiety or stress can occur when the perpetrator is speaking during the commission of crime. They often will degrade identification as the speech shifts triggered by these emotions can markedly change one or more of the parameters within the speech signal. The effects of ingested drugs or alcohol; and even a temporary health state such as a cold can affect the speech. The suspect may sometimes attempt to disguise their voice. All these affect the speaker identification process horrendously.

The vocal disguise can be classified as deliberate and non-deliberate or *electronic* and *non-electronic*. Electronic disguise refers to the use of electronic devices to alter the voice. Non-electronic one is to distort the voice by changing speakers’ vocal configuration and phonation manner, such as whisper, falsetto, foreign accent, change of speaking rate, imitation, pinched nostril and object in mouth.

The correct identification rate of normal voices can be degraded by the voice variations from background noise, different transmission channels, extreme emotions, illnesses, etc. If the voice is disguised deliberately the identification would become more difficult and even impossible. Therefore, it is necessary to study the effect of disguised voice on forensic speaker identification.

Researchers have used formant frequencies , fundamental frequency, F0 contour, Linear Prediction coefficients (Atal, 1974; Imperl, Kacic & Hovert, 1997), Cepstral Coefficients (Jakkhar, 2009; Medha, 2010; Sreevidya, 2010) and Mel Frequency Cepstral coefficients (Plumpe, Quateri & Reynolds, 1999; Hassan, Jamil & Rahman, 2004; Chandrika, 2010; Tiwari, 2010) to identify speaker. However, the Cepstral Coefficients and the Mel Frequency Cepstral Coefficients have been found to be more effective in speaker identification compared to other features. Hence, the present study will be focusing on usefulness of linear prediction cepstral coefficients and Mel frequency cepstral coefficients (MFCC) on speaker recognition.

Atal (1974) examined several different parameters using linear prediction model for their effectiveness for automatic recognition of speakers from their voices. The predictor coefficients, as the impulse response function, the autocorrelation function, the area function and the cepstrum function were used as input to an automatic speaker recognition system. The speech data consisted of six repetitions of 60 utterances by ten speakers. Result showed that cepstrum was found to be the most effective parameter, providing an identification accuracy of 70% for speech 50 ms in duration, which increased to more than 98% for duration of 0.5s Using the same speech data, verification accuracy was found to be approximately 83% for duration of 50 ms increasing to 95% for duration of 1sec. The cepstrum is used as the inverse Fourier transform of the log magnitude Fourier spectrum. It is used to separate the transfer function and the excitation signal which exists in the low quefrency and high quefrency, respectively. The coefficients that make up the resulting cepstrum are known as the cepstral coefficients.

In other studies (Jakkar, 2009; Medha, 2010; & Sreevidya, 2010) cepstrum was used for speaker identification. The maximum percent correct identification obtained using Cepstrum was 80% (Medha, 2010) and 80% (Sreevidya, 2010) in Indian languages.

Some experiments were conducted by Reich, Moll, & Curtis (1976) to find out the effect of vocal disguises upon speaker identification. Reich (1976) described an experiment involving the effects of selected vocal disguises upon spectrographic speaker identification. The results of this experiment suggest that certain vocal disguises markedly interfere with spectrographic speaker identification. The reduction in speaker identification performance ranged from 14.17% (slow rate) to 35.00% (free disguise). These experimental data obviously contradict Kersta's (1962) claim that spectrographic speaker identification is essentially unaffected by attempts at disguising one's voice. The mean performance level (56.67% correct) on the undisguised task was considerably poorer than the data for similar experimental conditions (approximately 80%) (Tosi, Oyer, Lashbrook, Pedrey, Nichol & Nash, 1972). In general, results of this experiment show that nasal and slow rate were the least effective disguises, while free disguise was the most effective on the spectrographic speaker identification. The exclusion of low confidence decisions produced significantly higher correct percentages. It was also found that stimulus words containing nasal phonemes (i.e., me, on, and) were considered quite useful for spectrographic speaker identification. Reich et al, (1976) found that the inclusion of disguised speech samples in the spectrographic matching tasks significantly interfered with speaker identification performance and had a significant effect on the types of errors made by the examiners. Specifically, the errors of false identification increased, accompanied by a proportional decrease in the errors of false elimination. Hence investigation failed to substantiate prior claims (Kersta, 1962; Anon, 1965) that spectrographic speaker identification is unaffected by attempts at vocal disguise.

Reich, & Duke (1979) describe another experiment involving the effects of selected vocal disguises upon speaker identification by listening. The results of this experiment suggested that certain vocal disguises markedly interfere with speaker identification by listening. The reduction in speaker identification performance by vocal disguise ranged from naïve listeners was 22.0% (slow rate) to 32.9% (nasal) and sophisticated listeners was 11.3% (hoarse) to 20.3% (nasal). In general, results of this experiment show that nasal disguise (naïve and sophisticated listeners) was the most effective, while slow rate disguise (naïve listeners) and hoarse disguise (sophisticated listeners) were the least effective disguises on the speaker identification by listening. Further, nasal disguise was the most effective disguise in speaker identification by listening experiment (Reich et al., 1979). In contrast, the nasal disguise was the least effective in a previous spectrographic matching experiment (Reich et al., 1976). Similarly, the power spectra of nasal consonants (Glenn & Kleiner, 1968) and coarticulated nasal spectra (Su; Li and Fu, 1974) seem to provide strong cues for the machine matching of speakers.

Reich (1981) examined the ability of naïve and sophisticated listeners to detect extemporaneous disguise in the male voice. Both naive and sophisticated listeners were able to detect the presence of selected disguises with a high degree of accuracy and reliability.

Also, the effects of vocal disguises markedly interfere with spectrographic speaker identification as well as speaker identification by listening. The nasal and slow rate were the least effective disguises, while free disguise was the most effective disguise upon the spectrographic speaker identification, and nasal disguise (naïve and sophisticated listeners) was the most effective, slow rate disguise (naïve listeners) and hoarse disguise (sophisticated listeners) were the least effective disguises upon the speaker identification by listening. Both naive and sophisticated listeners were able to detect the presence of selected vocal disguises with a high degree of accuracy and reliability.

However, till date there are limited studies on electronic vocal disguise. As on to date, several electronic vocal disguises are available in the market in various forms like pocket recorders, mobile handsets and electronic toys. This may increase the crime rate in the society. With the introduction of electronic disguises on mobile phones speaker identification is still more difficult as the culprit can select any electronic disguise and speak to demand ransom. In this context, the present study examined speaker identification in electronic vocal disguise using Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC). Specifically, in MFCC a power spectra was extracted with a filter bank uniformly spaced on the log Mel scale which is converted to a time domain to obtain the cepstra. In LPCC a power spectra was extracted based on LPC which is converted to time domain to get the cepstra.

Thus, the aim of the study was to establish ***Benchmark for speaker identification under electronic vocal disguise using Mel frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients. (LPCC)*** The objectives of the study were to provide benchmarks for (a) Mel-frequency cepstral coefficients for Telugu* vowels in electronic vocal disguise condition, and (b) linear prediction cepstral coefficients for Telugu vowels in electronic vocal disguise condition.

* Telugu is hypothetically classified as a Dravidian language with heavy Indo-Aryan influence and is native to the Indian subcontinent. It is the official language of Andhra Pradesh, it is also one of the twenty-two scheduled languages of the Republic of India and was conferred the status of a classical language by the government if India. The mother tongue of the majority of people of Andhra Pradesh, it is also spoken in neighbouring states like Chhattisgarh, Karnataka, Maharashtra, Orissa and Tamil Nadu. Telugu is the third most-spoken language in India (74 million native speakers according to the 2001 census) and is 15th in the Ethnologue list of most-spoken languages worldwide. (www.//htpp//.telugulanguage.com).

CHAPTER II

Review of literature

Expert opinion is increasingly being sought in the legal process as to whether two or more recordings of speech are from the same speaker or not. This is usually termed forensic speaker identification, or forensic speaker recognition. Forensic speaker identification can be very effective, contributing to both conviction and elimination of suspects. The aim of speaker identification is, not surprisingly, identification: 'to identify an unknown voice as one or none of a set of known voice' (Naik, 1994).

The voice identification technique was first adopted by the Michigan State Police in 1966 and introduced in the American court in the mid 1960's. Such method was used widely in different states including California, Florida and New York since then. However, different admission standards and interpretation methods were used among courts resulting in a lack of consistency (McDermott & Owen, 1996). Forensic voice identification had already been used in various crime cases, including murder, bomb threats, rape, political corruption and kidnapping. Some witnesses of these cases could see the criminals but some could not, for example, the voices were heard over a telephone line or when the witness was blindfolded.

The concept of recognizing a person through their voice can be useful as security control when accessing confidential information areas, access to remote computers, voice dialling, banking by telephone, telephone shopping, database access services, information services, voice mail or as PIN code for your ATM.

Forensic speaker identification can often be found classified as a kind of speaker recognition (Nolan, 1997). Hecker (1971) suggests that speaker recognition is "any decision-making process that uses the speaker-dependent features of the speech signal," and Atal (1976) offers the formulation "any decision-making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance." Another aspect wherein Atal's characterization is not totally correct for forensic speaker identification is that "it strongly suggests that an unambiguous, categorical outcome is expected: the person is either determined to be or determined not to be the speaker of a given utterance. In the forensic case the outcome should be a ratio of probabilities. Despite these shortcomings, it is clearly still helpful to persevere with the idea of forensic speaker identification as a kind of speaker recognition" (Rose, 1990). Figure 1 shows the schematic representation of speaker recognition.

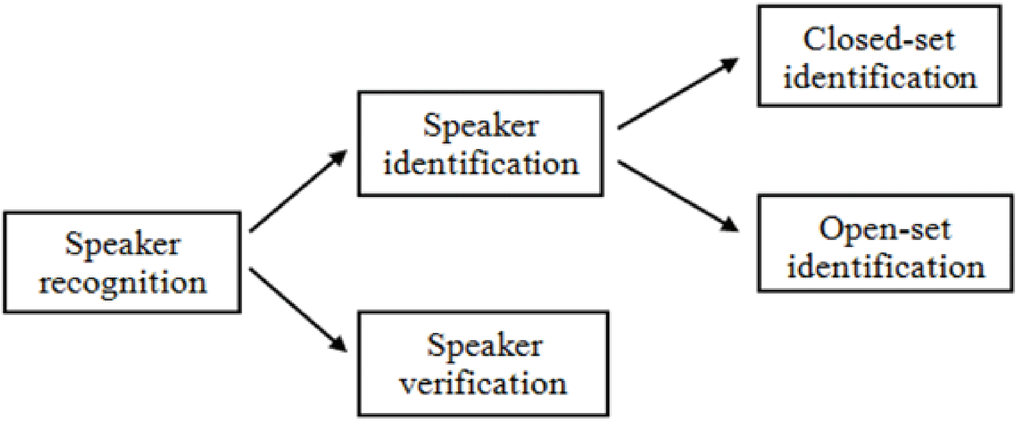


Figure 1: Schematic representation of speaker recognition.

There are two main classes of speaker recognition task, identification and verification (Furui, 1994; Nolan, 1997). The distinction between them rests on the type of question that is asked and on the nature of the decision-making task involved to answer that question.

The aim of speaker identification is ‘to identify an unknown voice as one or none of a set of known voice’ (Naik, 1994). One has a speech sample from an unknown speaker, and a set of speech samples from different speakers the identity of whom is known. The task is to compare the sample from the unknown speaker with the known set of samples, and determine whether it was produced by any of the known speaker (Nolan, 1983). Figure 2 shows a schematic representation of simple speaker identification. The speaker identification experiment is represented with a reference set of 50 known speaker samples. In Figure 2, the unknown sample on the left is compared with that from known speaker 1(A), then known speaker 2 (B), and so on. The question mark represents the question: are these two speech sample from the same speaker? If it is decided that the unknown sample is the same as one of the known speaker, say known speaker 4, then that identifies the speaker of the unknown sample as D.

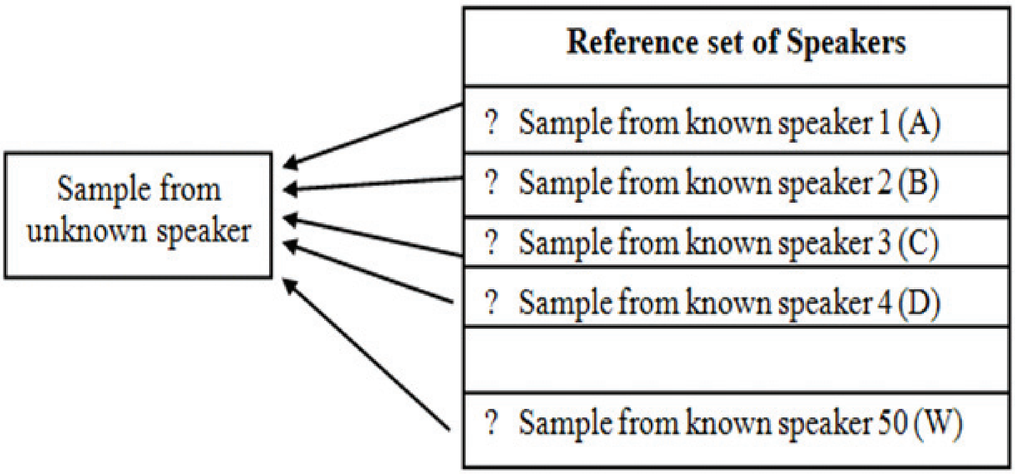

Reference set of Speakers
? Sample from known speaker 1 (A)
? Sample from known speaker 2 (B)
? Sample from known speaker 3 (C)
? Sample from known speaker 4 (D)
? Sample from known speaker 50 (W)
Sample from unknown speaker

Figure 2: Schematic representation of speaker identification.

Speaker verification is the other common task in speaker recognition. This is where ‘an identity claim from an individual is accepted or rejected by comparing a sample of his speech against a stored reference sample by the individual whose identity he is claiming' (Nolan, 1983). The schematic representation of speaker verification is shown in Figure 3. Here speaker D wants to access and verified. The system has samples of speaker D`s voice in storage, which it retrieves and compares with that of the sample rendered by speaker D. If the two voice samples are judged similar enough, speaker D`s claim is verified and he is given access.

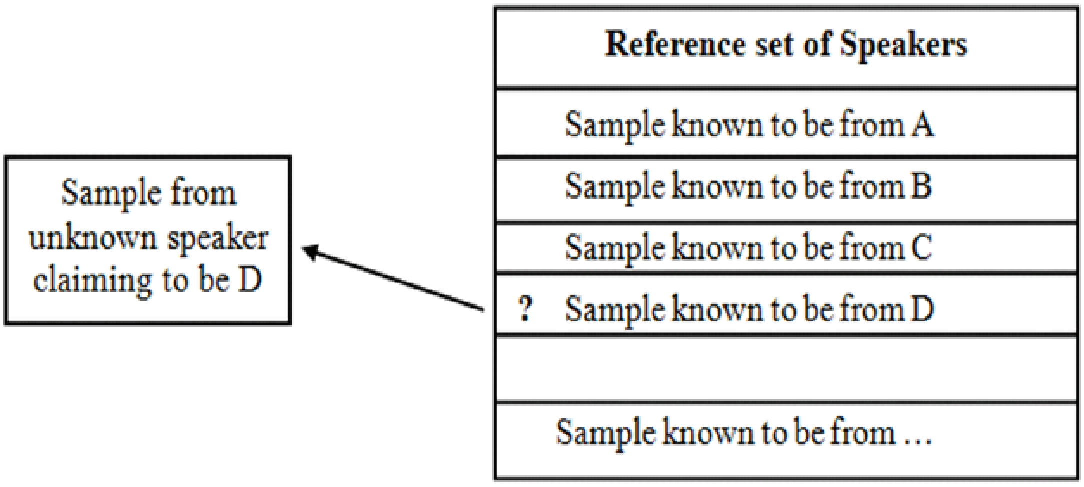

Sample from unknown speaker claiming to be D
Reference set of Speakers
Sample known to be from A
Sample known to be from B
Sample known to be from C
? Sample known to be from D
Sample known to be from …

Figure 3: Schematic representation of speaker verification.

In speaker identification, the reference set of known speakers can be of two types: closed or open. A closed reference set means that it is known that the owner of the unknown voice is one of the known speakers. An open set means that it is not known whether the owner of the unknown voice is present in the reference set or not. Closed set identification is usually a much easier task than open set identification. Since it is known that the unknown speaker is one of the reference set, the closed set identification task lies in (a) estimating the distance between the unknown speaker and each of the known reference speakers, and (b) picking the known speaker who is separated by the smallest distance from the unknown speaker. The pair of sample separated by the smallest distance is then assumed to be from the same speaker (Nolan, 1983). *Because the nearest known speaker is automatically selected in a closed set identification, no threshold is needed*. Both closed and open sets can occur in forensic case-work, although the latter, where one does not know if the putative offender is among the suspects or not, is usually more common. Since the task usually becomes very much simpler with a closed set, the distinction between open and closed set tasks is an important one in forensic speaker identification.

In speaker identification, only two type of decision are possible. Either the unknown sample is correctly identified or it is not. Verification is more complicated, with four types of decision. The decision can be correct in two ways: the speaker is correctly identified as being who they say they are, or not being who they say they are. And it can be incorrect in two ways: the identity claim of the speaker can be incorrectly rejected (the speaker is who they say they are but rejected), or incorrectly accepted (the speaker is an impostor but is nevertheless accepted).

In the open set speaker identification task three types of errors are possible. Figure 4 shows the schematic representation of classification of errors.

(a) Error A: A match did exist but the examiner selected the wrong one (false identification).

(b) Error B: A match did exist but the examiner failed to recognize it (false elimination).

(c) Error C: A match did not exist although the examiner selected one (false identification).

In the closed set speaker identification, since a match always exist, only one kind of error is possible: false identification or wrong identification. This error from closed set identification is labelled Error D.

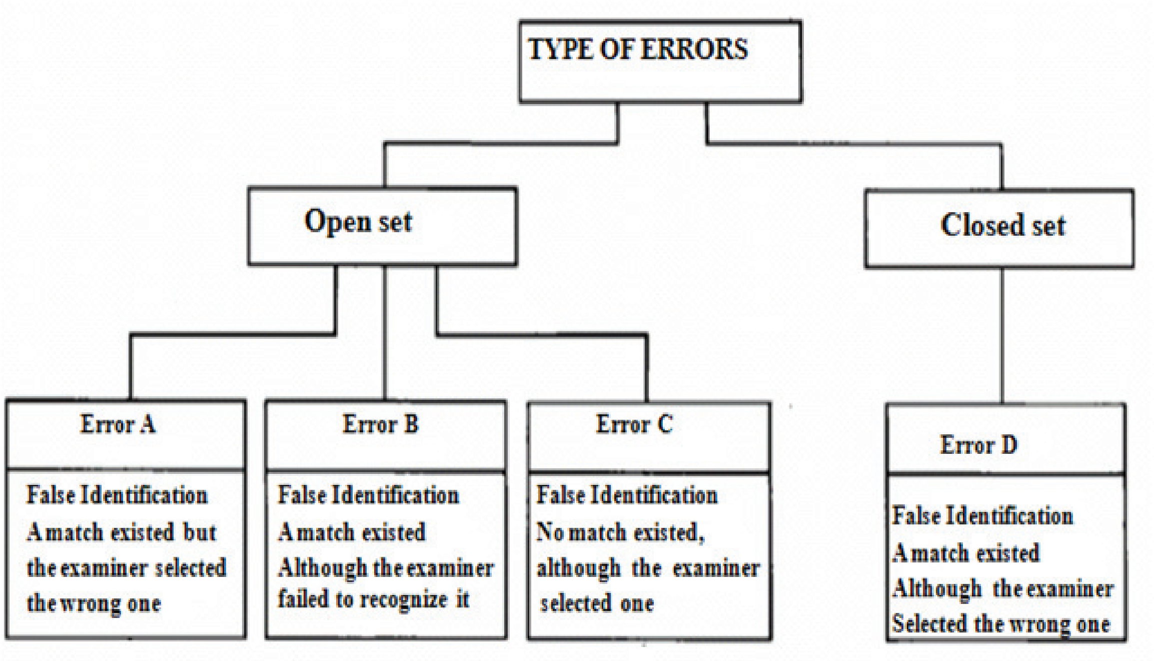

TYPE OF ERRORS
Open set
Closed set
Error A
False Identification
A match existed but the examiner selected the wrong one
Error B
False Identification
A match existed
Although the examiner failed to recognize it
Error C
False Identification
No match existed, although the examiner selected one
Error D
False Identification
A match existed
Although the examiner Selected the wrong one

Figure 4: Classification of errors.

Methods of speaker identification

Hecker (1971) classifies the methods of speaker identification into three general categories-(a)Speaker identification by listening (subjective method), (b) Speaker identification by visual examination of spectrograms (subjective method), and (c) Speaker identification by machine (objective method).

All have demonstrated some success in the laboratory but none have been particularly successful under field like conditions. Of these approaches, the third method (semi automatic and automatic) appears to be the most promising for the future, primarily because (1) specific parameters within the speech signal can be selected and analyzed serially or simultaneously, (2) the selected vectors may be used in various combinations, and (3) subjective analysis by human is eliminated.

Several studies have been reported on speaker identification by listening method. In studies of McGehee (1937), listeners attempted to select a single target voice from a set of five male voices after delays that ranged from 1 day to 5 months. The correct identification scores declined from 83% after 1 day to 80.8% after 1week, 68.5% after 2 weeks, 57% after l month, and to 13% after 5 months. Thompson (1985) used male voices in a six-voice line up task in which listeners rated each voice as to whether it was the voice they had heard 1 week previously. They could also respond that the voice heard previously was not in the line up or that they were not sure whether it was in the line up. However, the listener were not given the option of saying the voice heard previously was in the line up more than once. Thus, from the viewpoint of the listeners, the experiment was an open-set task, but not an independent-judgment task. Such a task can be considered an open-set, multiple-choice task with a decision threshold by the listener. The result were 62.1% correct identifications, 22.1% incorrect identifications, and 15.8% “not in line up" or "not sure if in line up" response.

Hecker (1971) reported that speaker recognition by listening appears to be the most accurate and reliable method at that time. Speaker authentication and identification were examined by Stevens (1968), for two different methods of presentation of the speech material: (a) speech samples presented aurally through headphones, and (b) speech samples presented visually as conventional intensity-frequency-time patterns, or spectrograms. They carried out two kinds of experiments- a series of closed tests in which there was a library of samples from eight speakers, and test utterances were known to be produced by one of the speakers; and a series of open tests in which the same library of eight speakers was used, but test utterances may or may not have been produced by one of the speakers. They reported that aural identification of talkers based on utterances of single words or phrases is more accurate than identification from the spectrograms and average error rate obtained by listening was 6% than visual 21% for the closed set identification. These scores depend upon the talker, the subject, and the phonetic content and duration of the speech material. For the open visual tests, appreciable numbers of false acceptances (incorrect authentications) were made. The results suggest procedures that might be used to minimize error scores in practical situations.

Schwartz (1968) reported speaker identification of gender using voiceless fricatives (/s/, /ʃ/, /f/, /θ/). Nine females and nine males recorded the four fricatives in isolation. The stimuli were randomized and presented via loudspeaker to ten listeners for gender identification. The results indicated that the listeners could identify the gender of the speakers from the isolated production of /s/ and /ʃ/, but could not from the /f/ and /θ/ production. Subsequent spectrographic analysis of the /s/ and /ʃ/ stimuli revealed that the female spectra tended generally to be higher and parallel in frequency compared to that in male. Ingemann (1968) support the above results and reported that listeners are often able to identify the sex of a speaker from hearing voiceless fricatives in isolation and sex was better identified on fricative /h/.

Schwartz (1968) & Ingemann (1968) employed isolated voiceless fricatives as auditory stimuli and they found that listeners could accurately identify speaker gender from these stimuli, especially from /h/, /s/, and /ʃ/. The laryngeal fundamental was not available to the listeners because of the voiceless condition of the consonants; these findings indicate that accurate speaker gender identification is possible from vocal tract resonance information alone.

Schwartz & Rine (1968) investigated the ability of listeners to identify speaker gender from two whispered vowels (/i/ and /α/). They found 100% correct identification for /α/ and 95% correct identification for /i/, despite the absence of the laryngeal fundamental.

Coleman (1971) employed /i/, /u/, and a prose passage to study the speaker gender identification abilities of his subjects. All stimuli were produced at the same vocal fundamental frequency (85 Hz) by means of an electrolarynx. Coleman discovered that the listeners are capable of accurately recognizing the gender of the speaker, even when the fundamental frequency remained constant for all speakers. In a later experiment, Coleman (1973) presented recordings of 40 speakers' normal (voiced) productions of a prose passage to a group of listeners, and he found that "... listeners were basing their judgments of the degree of maleness or femaleness in the voice on the frequency of the laryngeal fundamental"

Lass (1976) investigated the relative importance of the speaker's laryngeal fundamental frequency and vocal tract resonance characteristics in speaker sex identification tasks. Six sustained isolated vowels (/i/, /ε/, /æ/, /α/, /o/, and /u/) were recorded by 20 speakers, 10 males and 10 females, in a normal and whispered manner. A total of three master tapes (voiced, whispered, and filtered) were constructed from these recordings. The filtered tape involved 255 Hz low-pass filtering of the voiced tape. The tapes were played to 15 listeners for speaker sex identification judgments and confidence ratings of their evaluations. Results of their judgments indicate that, of the 1800 identifications made for each tape (20 speakers X

6 vowels X 15 listeners), 96 % were correct for the voiced tape, 91% were correct for the filtered tape, and 75% were correct for the whispered tape. Moreover, the listeners were most confident in their judgments on the voiced tape, followed by the filtered tape, and showed the least amount of confidence on the whispered tape. These findings indicate that the laryngeal fundamental frequency appears to be a more important acoustic cue in speaker sex identification tasks than the resonance characteristics of the speaker.

Speaker identification by listening only, one of the methods discussed is, far from being 100% accurate. It is an entirely subjective method; an expert witness using only this method would be unable to justify his conclusions in a court of law (Hecker, 1971).

Speaker identification by visual examination of spectrograms (subjective method): In the mid 1940's, the scientists of the Bell Telephone Laboratories in USA developed the first sound spectrograph (the Sonagraph), a visual record of speech including frequency, intensity and time (McDermott & Owen, 1996). In the Fifties, Lawrence Kersta, an engineer from the Bell Telephone Laboratories, developed "voiceprint identification" (Hollien, 2002). Studies using the spectrograph were carried out in the 1950s and 1960s in USA (Hollien, 2002).

In 1946, Potter, Kopp & Green suggested of using the spectrogram as a method of speaker recognition. They portray talker dependent features in addition to the phonetic variations. The term *Voiceprint* was introduced by Lawrence Kersta (1960); he studied if the patterns on sonograms exhibited features which could be used to identify speakers. He published a paper on "voice identification" in which he initiated an erroneous idea that there is a close relationship between finger print and voice print. Kersta's identification method human observers visually matching spectrogram and to duplicate his investigation with what we believe are methodological and analytical improvements.

Kersta (1962) examined the "voiceprint" using spectrograms taken from five clue words spoken in isolation using 12 talkers and closed test identification. High school girls were trained for 5 days to identify talkers from spectrograms on the basis of eight "unique acoustic cues." A 5x4, 9x4, or 12x4 matrixes of spectrograms, was presented to the observer whose task was to group the spectrogram in piles representing the individual talkers. Results of the study show high rate of identification accuracy that were inversely related to the number of talkers. For 5, 9 and 12 talkers, identification rate were 99.6%, 99.2% and 99.0%, respectively and for words spoken in isolation the correct rates were higher for the "bar prints" than for the "contour prints".

However, similar results are not obtained by other researches. The correct identification scores reported by Kersta are outstandingly high, 99%-100%, for short words spoken either in isolation or in context, as compared to (a) 81%-87%, for short words spoken in isolation, reported by Bricker & Pruzansky (1966), (b) 89% for short words taken from context, reported by Pruzansky (1963), and (c) 84%-92%, for short words spoken in isolation, reported by Pollack, Pickett, & Sumby (1954).

Some methods yielded virtually 100% correct identification rates when the test stimuli were identical sentence and somewhat lower rates when the test stimuli were short or monosyllabic words spoken in isolation. In practical situation, the stimuli available for comparison are words spoken in different context. Young & Campbell (1967) using three words spoken by five speakers and 10 examiners with spectrogram reported correct identification rate 37.3%, and 78.4% in context and isolation respectively. The results were interpreted to indicate that different contexts decrease the identification ability of observers because the shorter stimulus durations of words in context decreases the amount of acoustic information available for matching, and the different spectrographic portrayals introduced by different phonetic contexts outweighs any intra-talker consistency.

Stevens (1968) compared aural with the visual examination of spectrograms using a set of eight talkers and a series of identification tests. The average error rate for listening was 6% and for visual was 21%. He observed that the mean error rate decreased from approximately 33.0% to 18.0 % as the duration of the speech sample increased from monosyllabic words to phrases and sentences. They also concluded that for visual identification, longer utterances increase the probability of correct identification.

Considering the above studies, some move towards speaker identification is obvious. However, a general procedure is not known or accepted.

Bolt (1970) reported that speech spectrograms, when used for voice identification, are not analogous to finger prints, primarily because of fundamental differences in the source of patterns and differences in their interpretation. To asses' reliability of voice identification under practical conditions, whether by experts or explicit procedures are not yet been made, and requirements for such studies are not outlined. Hecker (1971) reported that speaker recognition by visual comparison of spectrograms is coming into use in criminology, but the validity of this method is still in question.

Findings of a large scale study (Tosi, 1972) were published in which attempts were made to more closely imitate law enforcement conditions, but only spectral comparisons were made (no aural). A two-year experiment on voice identification through visual inspection of spectrograms was performed with the twofold goal of checking Kersta's (1962) claims in this matter and testing models including variables related to forensic tasks. The 250 speakers used in this experiment were randomly selected from a homogeneous population of 25000 males speaking general American English, all students at Michigan State University. A total of 34996 experimental trials of identification were performed by 29 trained examiners. Each trial involved 10 to 40 known voices, in various conditions: With closed and open trials, contemporary and non-contemporary spectrograms, nine or six clue words spoken in isolation, in a fixed context and in a random context, etc. The examiners were forced to reach a positive decision (identification or elimination) in each instance, taking an average time of 15 minutes. Their decisions were based solely on inspection of spectrograms; listening to the identification by voices was excluded from this experiment. The examiners graded their self-confidence in their judgments on a 4-point scale (1 and 2, uncertain; 3 and 4, certain). Results of this experiment confirmed Kersta's experimental data, which involved only closed trials of contemporary spectrograms and clue words spoken in isolation. Experimental trials of this study, correlated with forensic models (open trials, fixed and random contexts, non-contemporary spectrograms), yielded an error of approximately 6% false identifications and approximately 13% false eliminations. The examiners judged approximately 60% of their wrong answers and 20% of their right answers as "uncertain." This suggests that if the examiners had been able to express no opinion when in doubt, only 74% of the total number of tasks would have had a positive answer, with approximately 2% errors of false identification and 5% errors of false elimination. Main differences of conditions that could exist between models and real cases were as follows:

(1) Population of known voices: In forensic cases, the catalog of known voices could theoretically include millions of samples. In the present practical situations that police must handle. In these cases the catalog of known voices is open, true, but limited to a few suspected persons. Therefore, it seems reasonable to disregard size of the population of known voices as a differential characteristic that could hamper extrapolation of results from the present experiment to real cases.

(2) Availability of time and responsibility of the examiners: In real cases, a professional examiner may devote all the time necessary to reach a conclusion. In addition, he is aware of the consequences that a wrong decision could mean to his professional status as well as the consequences to the speaker whom he might erroneously identify. Availability of time and responsibility between experimental and professional examiners might help to improve the accuracy of the professional examiners.

(3) Type of decisions examiners are urged to reach in each trial: In the statistical models, the examiners were forced to reach a positive conclusion in each trial, even if they were uncertain of the correct response. In real forensic cases, the professional examiner is permitted to make the following alternative decision- (a) Positive identification; (b) Positive elimination; (c) Possibility that the unknown speaker is one of the suspected persons, but more evidence is necessary in order to reach a positive identification; (d) Possibility that the unknown speaker is none of the available suspected persons, but more evidence is necessary to reach a positive elimination; and (e) Unable to reach any conclusion with the available voice samples. These possibilities of alternative decisions could confer an extremely high reliability to the positive identifications or eliminations.

(4) Availability of clues: In the experimental models of this study, only spectrograms of nine or six clue words were available to the examiners for visual inspection. Rather, a professional examiner is entitled to request as many samples as he deems necessary to reach a positive conclusion. In real forensic cases the professional examiner must necessarily listen first to the unknown and known voices while processing the spectrograms for visual comparison. A combination of methods of voice recognition by listening and by visual enhances the accuracy of voice identifications.

In summary, Tosi (1972) suggest that the conditions a professional examiner encounters performing voice identifications will tend to decrease rather than increase the percentage of error observed in the present experiment.

Hazen (1973) reported that for reduced population, error rates were higher for closed tests (12.86% and 57.14%) than for open tests (11.91% and 52.38%), but were almost five times as great for the different context condition (57.14% and 52.38%) than for the same context condition (12.86% and 11.91%). Hollien (1974) comments on spectrographic speaker identification, it now appears that the controversy about "voiceprints" is doing the judicial system and the relevant scientific community a considerable disservice. Final perspective of the letter is to urge responsible investigators interested in the problem to focus their research activities on the development of methods. That will provide efficient and objective ways to identify individuals from their speech, especially in the forensic situation. All these may be possible under undisguised voice. However, with vocal disguise the situation may be different. Reich (1976) reported that the examiners were able to match speakers with a moderate degree of accuracy (56.67%) when there was no attempt to vocally disguise either utterance. In spectrographic speaker identification nasal and slow rate were the least effective disguises, while free disguise was the most effective.

Most of the speaker identifications are conducted in laboratory condition. The results may differ in actual conditions. A survey of 2000 voice identification comparisons made by Federal Bureau of Investigation (FBI) examiners (Koenig 1986) was used to determine the observed error rate of the spectrographic voice identification technique under actual forensic conditions. The survey revealed that decisions were made in 34.8% of the comparisons with a 0.31% false identification error rate and a 0.53% false elimination error rate. These error rates are expected to represent the minimum error rates under actual forensic conditions.

Bolt et. al. (1973) wrote a letter to the editor on reviews of recent research on speaker identification by comparisons of speech spectrograms by human observers. Various factors affecting the reliability of identification were discussed, particularly those that would be present in practical forensic situations and concluded that identification under practical conditions has not been scientifically established.

Speaker identification by machine (objective method)

In the years following identification by the aural mode, voice processing technology became quite popular and the simplest approach used was to generate and examine amplitude and frequency, time matrices of speech samples. The other approach was to extract speaker dependent parameter from the signals and analyze them by machines. The objective methods include Semi-automatic method, and Automatic method. In the semi-automatic method, there is extensive involvement of the examiner with the computer, whereas in the automatic method, this contact is limited.

Semi-automatic method: The examiner selects unknown and known samples (similar phonemes, syllables, words and phrase) from speech samples, which have to be compared, i.e. computer processes these samples, extracts parameters and analyzes them according to a particular program. The interpretation is made by the examiner.

Automatic method: In this the computer does all the work and the participation of the examiner is minimal. For the purpose of automatic identification, special algorithms are used which differ based on the phonetic context. This method is used very often in forensic sciences but factors such as noise and distortion factors of voice and other samples need to be controlled. In such case a combination of subjective and objective methods should be used.

Some studies related to speaker identification by machine published in earlier years are summarized below.

Glenn & Kleiner (1968), describe a method of automatic speaker identification based on the physiology of the vocal apparatus and essentially independent of the spoken message has been developed. Power spectra produced during nasal phonation are transformed and statistically matched. Initially, the population of 30 speakers was divided into three subclasses, each containing 10 speakers. Subclass l contained 10 male speakers, Subclass 2 contained 10 females' speakers, and Subclass 3 contained an additional 10 male speakers. For each speaker, all 10 samples of the spectrum of /n/ from the test set were averaged to form a test vector. The test vectors were compared, with the stored speaker reference vectors for the appropriate subclass. The values of the cosine of the angle between the reference and the test vectors are correlation values between the test vector for a given speaker and the reference vector for each speaker in the subclass. The maximum correlation value for each test vector is used and 97% over all correct identification was attained. Next, the effect of a larger population was tested by correlating each speaker's averaged test data with the reference vectors for all 30 speakers and an average identification accuracy of 93 % was reached. Finally, the effect of averaging speaker samples was tested as follows. The same speaker reference vectors based on all 10 training samples were used. However, the test data were subjected to varying degrees of averaging. First, single-speaker samples were correlated with the 30 speaker reference vectors. The average identification accuracy for all 300 such samples (10 per speaker) was 43%. Then, averages of two speaker samples from the test data were taken as test vectors. The average identification accuracy for 150 such vectors was 62%. Next averages of five speaker samples from the test data were taken as test vectors. The average identification accuracy for 60 such vectors was 82%.

In this experiment involving the identification of individual speakers out of a population of 10 speakers, an average identification accuracy of 97% was obtained. With an experimental population of 30 speakers, identification accuracy was 93%. The results of the experiments support the hypothesis that the power spectrum of acoustic radiation produced during nasal phonation provides a strong cue to speaker identity. The procedure developed to exploit this information provides a basis for automatic speaker identification without detailed knowledge of the message spoken.

Automatic speaker verification was accomplished by Luck (1969) using cepstral measurement to characterize short segments in each of the first two vowels of the standard test phrase" My code is." The length of the word "my" and the speaker's pitch were used as additional parameters. The verification decision is treated as a two-class problem, the speaker being either the authorized speaker or an impostor.

Reference data is used only for the authorized speaker. The decision is based on the test sample's distance to the nearest reference sample. Data is presented to show that, if reference samples are collected over a period of many days, then verification is possible more than two months later, whereas, if reference data is collected at one sitting, verification is highly inaccurate as little as 1 h later. Four authorized speakers and 30 impostors were examined, with error rates obtained from 6% to 13%. Impostors attempting to mimic the authorized speaker could not improve their ability to deceive the system significantly.

It has been observed by many who have seen the system in operation that greater accuracy would be obtained if a final decision were based on a series of two or three repetitions of the test phrase. This is to say that increased accuracy depends on increasing the information available to the decision mechanism. One might increase the available information, for example, by (a) changing the decision rule so as to make more efficient use of the data contained in the 34-dimensional vector; (b) increasing the dimensionality of the vector by using additional coefficients from the Fourier transform of the log spectrum; (c) defining additional analysis segments to be used in the decision process; (d) seeking new types of measurements that contain more compact information about the speaker; or (e) providing reference data on a few impostors. It should be mentioned in closing that, while it appears difficult for an impostor to change his voice to fool the system, it is a trivial matter for the reference speaker to alter his voice if he wishes to be identified as an impostor. Thus, although the system is reasonably tolerant of the normal variations in the reference speaker's voice, the data presented is necessarily based on the assumption that he wishes to have his identity verified.

Meltzer & Lehiste (1972) investigated the relative quality of synthetic speech. They selected three speaker one man, one women and one child. They recorded a set of 10 monophthong English vowels stimulated by each speaker. Ten vowels were synthesized on a Glace-Holmes synthesizer using the spectrograms of each speaker. Formant values for men, women, and children were combined with the respective fundamental frequencies 9 different combinations for each of the 10 vowels was synthesized. The 150 stimuli were presented to 60 trained listeners for both vowel and speaker identification. The overall vowel and speaker identification score for the normal set were 79.46% and 90.03% respectively, and for synthesized set were 50.87% and 69.73%, respectively. The differences from the normal set (−28.59 and −20.30%) constitute an evaluation measure for the performance of the synthesizer.

Wolf (1972) describes an investigation of an efficient approach to selecting such parameters, which are motivated by known relations between the voice signal and vocal-tract shapes and gestures. In a scheme for the mechanical recognition of speakers it, is desirable to use acoustic parameters that are closely related to voice characteristics that distinguish speakers. This study describes an investigation of an efficient approach to selecting such parameters, which are motivated by known relations between the voice signal and vocal-tract shapes and gestures. Rather than general measurements over the extent of an utterance, only significant features of selected segments are used. A simulation of a speaker recognition system as performed by manually locating speech events within utterances and using parameters measured at these locations to classify the speakers. Useful parameters were found in F0, features of vowel and nasal consonant spectra, estimation of glottal source spectrum slope, word duration, and voice onset time. These parameters were tested in speaker recognition paradigms using simple linear classification procedures. When only 17 such parameters were used, no errors were made in speaker identification from a set of 21 adult male speakers. Under the same condition, speaker verification errors of the order of 2% were also obtained.

Atal (1972) examined the temporal variations of pitch in speech as a speaker identifying characteristics. The pitch data was obtained from 60 utterances, consisting of six repetitions of the same sentence, spoken by 10 speakers. The pitch data for each utterance was represented by a 20-dimensional vector in the Karhunen-Loeve coordinate system. The 20-dimensional vectors representing the pitch contours were linearly transformed so that the ratio of inter-speaker to intra-speaker variance in the transformed space was maximized. The speaker corresponding to the reference vector with the smallest distance was considered as correct identification. The percentage of correct identifications was reported 97% and suggested that temporal variations of pitch could be used effectively for automatic speaker recognition.

In another experiment Atal (1974) examined several different parameters using linear prediction model for their effectiveness for automatic recognition of speakers from their voices. He determined twelve predictor coefficients approximately once every 50 ms from speech sampled at 10 kHz. The predictor coefficients, as the impulse response function, the autocorrelation function, the area function, and the cepstrum function were used as input to an automatic speaker-recognition system. The speech data consisted of 60 utterances, consisting of six repetitions of the same sentence spoken by 10 speakers. The identification decision was based on the distance of the test sample vector from the reference vector for different speakers in the population; the speaker corresponding to the reference vector with the smallest distance was judged to be the unknown speaker. In verification, the speaker was verified if the distance between the test sample vector and the reference vector for the claimed speaker was less than a fixed threshold. He reported that the cepstrum was found to be the most effective parameter, providing an identification accuracy of 70% for speech 50 ms in duration, which increased to more than 98% for a duration of 0.5 sec. Using the same speech data, the verification accuracy was found to be approximately 83% for a duration of 50 ms, increasing to 98% for a duration of 1sec.

Doddington (1974) developed the speaker verification system using six spectral/time matrices located within a test phrase with corresponding matrices defined during training. Evaluation was performed over a data set including 50 "known" speakers and 70 "casual impostors" including 20% female speakers in each session. Five different phrases (including "We were away a year ago") were collected in each session. Each matrix is 0.1 sec long and is precisely located by scanning the test phrase for a best match with the reference matrix. Known speakers gave 100 sessions; Impostors; 20. Data collection spanned 3.5 months. First 50 sessions of each known speaker's data were used for training, last 50 for test; 0.6% of the phrases yielded unusable data. Substitute phrase from that session was used if phrases yielded unusable data (two substitutions allowed, maximum). All impostor acceptance rates were determined for 2% true speaker rejection. A single fixed threshold was used for all speakers. Impostor acceptance rates were 2.5% for one phrase, 0.25% for two phrases, and 0.08% for three phrases. Five percent of known speaker data was labelled by the speakers as "not normal" because of respiratory ailments, etc. This data yielded a 4.5% reject rate for one phrase. Two professional mimics were employed to attempt to defeat the system. Each chose the five subjects he thought he could most easily mimic. Interactive trials with immediate feedback were of no apparent aid. Successful impersonation of about 5.5% for one phrase was achieved.

No successful attempts for three phrases could be constructed from the mimic data. Reject rate for known speakers was plotted versus session number, at a nominal reject rate of 10%.Initial and final reject rates of 5% and 15%, respectively, indicate the necessity of adaptation in a practical system.

Hollien (1977) carried out a study in order to evaluate the LTS discriminative function relative to large populations, different languages, and speaker/ system distortions. These issues were studied in two separate experiments. Intra-speaker and inter-speaker variations in long-term speech spectra constituted the experimental measures for both. In the first experiment, LTS was applied to two relatively large populations of American and Polish college students; the studies were carried out both for unlimited and restricted pass bands, the second condition simulating telephone transmissions. Laboratory simulation of field conditions was the focus of the second experiment. The distortions may result from system characteristics or may be speaker generated. The distortions can be caused by ambient/intermittent noise, competing signals, restricted pass band, and similar conditions. Examples of speaker induced distortions include emotional states, conditions of health, stress, and disguise. The identification was based on comparisons with normal speech production. As with the first experiment, the effects of full and limited pass bands were studied. The results should not be generalized directly to practical and/or applied situations in the speaker identification area.

In this study two experiments were carried out in which long-term spectra were extracted from controlled speech samples in order to study the effectiveness of that technique as a cue for speaker identification. In the first study, power spectra were computed separately for groups of 50 American and 50 Polish male speakers under full band and pass band conditions; an n-dimensional Euclidean distance technique was used to permit identifications. The procedure resulted in high levels of speaker identification for these large groups especially under the full band conditions. In a second experiment, the same approach was employed in order to discover if it was resistant to the effects of variation in speech production at least under laboratory conditions. Talkers were 25 adult American males; three different speaker conditions were studied: (a) normal speech, (b) speech during stress, and (c) disguised speech. The results demonstrated high levels of correct speaker identification for normal speech, slightly reduced scores for speech during stress and markedly reduced correct identifications for disguised speech. It would appear that long-term speech spectra can be utilized to identify individuals from their speech even in relatively large groups when they are speaking normally or under stress (of the type studied); LTS does not appear to be an effective technique when voice disguise is employed. While this approach was utilized only in controlled laboratory experiments, it is suggested that it may have some merit for use in applied situations or as one of the features in a multiple-vector approach.

The results of this research suggest several conclusions. First, it may be concluded that n- dimensional Euclidean distance among long-term speech spectra may be successfully utilized as criteria for speaker identification, at least under laboratory conditions. Moreover, this method exhibits a number of advantages: (a) It is relatively simple to carry out; (b) it eliminates such crucial factors as the time-alignment problem; (c) the data generated for the identifications do not depend on the overall power level of the speech samples used; and (d) the process does not depend on human and, hence, subjective judgments. Finally, it appears that distortions created by limited pass band and stress as these two factors are defined in these experiments have only minimal effects on the sensitivity of the LTS vector as a speaker identification cue.

On the other hand, this method does not appear to be a viable one when talkers disguise their speech at least, when the LTS vector is used alone as an identification technique. Moreover, the multiple and interactive effects of two or more distorting parameters appear to degrade the process by more than the sum of the individual effects and, in such cases, the identification levels quickly become unacceptable.

In short, as with so many other approaches to the problem of speaker identification, the LTS technique constitutes a reasonable robust tool in the laboratory but its efficiency is quickly reduced when distorting effects of the type found in the more realistic environment impinge on the process. On the other hand, it is quite possible that an LTS vector can be utilized successfully as one of several (vectors) in a multifactor scheme of automatic speaker identification.

Johnson (1977) attempted to utilize LTS in a situation that more closely parallels actual field application. Three “crimes” involving telephoned messages were simulated; each of the telephone calls was recorded simultaneously on both a reel-to-reel tape recorder via direct hook up and on a cassette recorder via a suction cup tap. All subjects (talkers and “suspects”) were volunteers drawn from a group of cooperating law enforcement agents. Two sets of twelve suspects each were recorded for the first two “crimes”; suspects for the third crime were drawn from a pool consisting of the two previous sets of individuals. All evaluations were conducted on the basis of a closed set paradigm. Sixteen and eight second samples from each set of unknowns and suspects were recorded and then subjected to power spectra analysis ; the resultant data sets were analyzed by discriminate analysis , a pattern matching statistical technique. Preliminary results indicate that the LTS method does not perform as well as a speaker identification cue under forensic conditions as it does in the laboratory. Since system frequency response is of substantial importance to the long-term spectra technique, the observed degradation in LTS performance appears to be due to limitations in equipment and in the communication channels utilized in the research.

Furui (1978) examined this effect on two kinds of speaker recognition; one used the time pattern of both the fundamental frequency and log-area-ratio parameters and the other used several kinds of statistical features derived from them. Results of speaker recognition experiments revealed that the long-term variation effects have a great influence on both recognition methods, but are more evident in recognition using statistical parameters. In order to reduce the error rate after a long interval, it is desirable to collect learning samples of each speaker over a long period and measure the weighted distance based on the long-term variability of the feature parameters. When the learning samples are collected over a short period, it is effective to apply spectral equalization using the spectrum averaged over all the voiced portions of the input speech. By this method, an accuracy of 95% can be obtained in speaker verification even after five years using statistical parameters of a spoken word.

In summary, Glenn & Kleiner (1968) describe an experiment involving identification based on the spectrum of nasal sounds in different environments in test and reference data. If just one speaker sample was correlated with the thirty reference vector, a correct identification rate of 43% was obtained. This rose to 93% if the average of 10 speaker samples was used for correlation and further to 97% if the relevant population of speakers was reduced to 10. These results indicate that quite accurate speaker identification can be achieved on the basis of spectral information taken from individual segment of an utterance, in this case nasal. It is noted by the authors that no account was taken of the phonetic environment of the nasals. If the test had been restricted to exponents of /n/ in a single environment, or if the effect of coarticulation could somehow have been factored out, it might be expected that within-speaker variation would have been reduced and as a result some of the errors eliminated.

Luck (1969) reported that 34-dimensional measurement vector on an independent set of data is the best for speaker identification process and data taken at one sitting will not be representative of the variations in a speaker's voice over a longer period of time. The possibility that impostor could improve their score with extensive practice but it is encouraging that the system held up as well as it did under the first attempt at mimicking. It appears difficult for an impostor to change his voice to fool the system; impostors attempting to mimic the authorized speaker could not improve their ability to deceive the system significantly.

Wolf (1972) measured fundamental frequency at a number of points in utterances, and found these measurements to be among the most efficient at disguising speakers. Wolf (1972) also found two nasal spectral parameters, one from /m/ and one from /n/, this time extracted from read sentences, to be ranked second and third among a number of segmental parameters. An average identification error of 1.5% was achieved for 210 "utterances" by the 21 speakers with only nine parameters if parameters was increased to 17, zero identification error was achieved.

The study conducted by Doddington et. al. (1974) to develop the speaker verification system using of six spectral/time matrices located within a test phrase with corresponding matrices defined during training. Each matrix is 0.1 sec long and is precisely located by scanning the test phrase for a best match with the reference matrix. All impostor acceptance rates were determined for 2% true speaker rejection.

Hollien & Majewski (1977) who achieved less good identification for American English than the Polish speakers concluded that the power of the long term spectrum as an identification tool might be ‘somewhat language dependent’. Identification were computed from 80-10000Hz long term spectra, and also band limited (315-3150Hz) versions simulating telephone transmissions. With the full bandwidth, identification dropped from 1005 for normal speech to 92% under stress or to 20% under disguise; with the limited bandwidth, from 88% to 685 under stress or 32% under disguise. LTS does not appear to be an effective technique when voice disguise is employed. The LTS technique constitutes a reasonable robust tool in the laboratory but its efficiency is quickly reduced when distorting effects of the type found in the more realistic environment impinge on the process. It is quite possible that an LTS vector can be utilized successfully as one of several (vectors) in a multifactor scheme of automatic speaker identification.

Johnson et. al. (1977) reported that preliminary results indicate that the LTS method does not perform as well as a speaker identification cue under forensic conditions as it does in the laboratory. Since system frequency response is of substantial importance to the long-term spectra technique, the observed degradation in LTS performance appears to be due to limitations in equipment and in the communication channels utilized in the research.

Pamela (2002) investigated the reliability of voiceprints by extracting acoustic parameters in the speech samples. Six normal Hindi speaking male subjects in the age range of 20-25 years participated in the study. Twenty-nine bisyllabic meaningful

Hindi words with 16 plosives, five nasals, four affricates and four fricatives in the word-medial position formed the material. Subject read the words five times. All recordings were audio-recorded and stored onto the computer memory. F_2, F_2 transition duration, onset of frication noise, onset of burst in stop consonants, closer duration and duration of phonemes were measured from wideband spectrograms (VSS-SSL). Percent of time a parameter was the same within and between subjects was noted. The results indicated no significant difference in F_2, onset of burst and frication noise, F_3 transition duration, closure duration, and phoneme duration between subjects. However, the results indicated high intra-subject variability. High intra-subject variability for F_2 transition duration, onset of burst, closer duration, retroflex and F_2 of high vowels was observed. Low inter-subject variability and high intra-subject variability for phoneme duration was observed indicating that this could be considered as one of the parameters for speaker verification. The results indicated that more than 67% of measures were different across subjects and 61% of measures were different within subjects. It was suggested that two speech samples can be considered to be of the same speaker when not more than 61% of the measures are different and two speech samples can be considered to be from different speakers when more than 67% of the measures are different. Probably this was the first time in India, an attempt to establish benchmarking was done.

Thus, semi-automatic speaker identification (SAUSI) included attempts to use nasal spectra, 34-dimensional vector, F0 at different points of utterances, Spectral/time matrices, and Long-term spectra and LTS vectors. However, no parameter is found to be 100% efficient across conditions and disguise. The future should tell us about an effective SAUSI.

In the previous studies speaker identification by listening (subjective method), visual examination of spectrograms (subjective method) and machine (objective method) was done. None of the studies provide 100% correct identification in forensic situation.

For manual forensic speaker identification (FSI) the correct identification rate by normal voices can be degraded by the voice variations from great background noise, different transmission channels, extreme emotions, illnesses, etc. if the voice is disguised deliberately the identification would become more difficult and even impossible. Therefore, it is necessary to study the effect of disguised voice on FSI.

Reich (1976) described an experiment involving the effects of selected vocal disguises upon spectrographic speaker identification. The results of this experiment suggest that certain vocal disguises markedly interfere with spectrographic speaker identification. The reduction in speaker identification performance ranged from 14.17% (slow rate) to 35.00% (free disguise). These experimental data obviously contradict Kersta's (1962) claim that spectrographic speaker identification is essentially unaffected by attempts at disguising one's voice. The mean performance level (56.67% correct) on the undisguised task was considerably poorer than the data for similar experimental conditions (approximately 80%) Tosi, Oyer, Lashbrook, Pedrey, Nichol & Nash (1972).

In general, results of this experiment show that nasal and slow rate were the least effective disguises, while free disguise was the most effective on the spectrographic speaker identification. The exclusion of low confidence decisions produces significantly higher correct percentages. It is readily apparent that stimulus words containing nasal phonemes (i.e., me, on, and) were considered quit useful for spectrographic speaker identification. Reich, Moll, & Curtis (1976) found that the inclusion of disguised speech samples in the spectrographic matching tasks significantly interfered with speaker identification performance and had a significant effect on the types of errors made by the examiners. Specifically, the errors of false identification increased, accompanied by a proportional decrease in the errors of false elimination. Hence investigation failed to substantiate prior claims (Kersta, 1962; Anon., 1965) that spectrographic speaker identification is unaffected by attempts at vocal disguise.

Reich & Duke. (1979) describe another experiment involving the effects of selected vocal disguises upon speaker identification by listening. The results of this experiment suggested that certain vocal disguises markedly interfere with speaker identification by listening. The reduction in speaker identification performance by vocal disguise ranged from naïve listeners was 22.0% (slow rate) to 32.9% (nasal) and sophisticated listeners was 11.3% (hoarse) to 20.3% (nasal). In general, results of this experiment show that nasal disguise (naïve and sophisticated listeners) was the most effective, while slow rate disguise (naïve listeners) and hoarse disguise (sophisticated listeners) were the least effective disguises on the speaker identification by listening.

The nasal disguise, for example, was the most effective disguise in speaker identification by listening experiment (Reich & Duke, 1979). In contrast, the nasal disguise was the least effective in a previous spectrographic matching experiment (Reich, Moll & Curtis, 1976). Similarly, the power spectra of nasal consonants (Glenn and Kleiner, 1968) and coarticulated nasal spectra (Su; Li and Fu,, 1974) seem to provide strong cues for the machine matching of speakers. It is interesting to know the listeners in the present study were unable to successfully utilize these seemingly speaker dependent cues. The free (i.e., extemporaneous) disguise proved to be very effective in both the spectrographic matching experiment (Reich, Moll & Curtis, 1976) and the present listening experiment.

There are few disguises, but first it is important to determine if the talker is attempting to alter, or not alter, his or her speaking mode. Reich (1981) examined the ability of naïve and sophisticated listeners to detect extemporaneous disguise in the male voice. Both naive and sophisticated listeners were able to detect the presence of selected disguises with a high degree of accuracy and reliability.

Thus, the effects of certain vocal disguises markedly interfere with spectrographic speaker identification as well as speaker identification by listening. The nasal and slow rate were the least effective disguises, while free disguise was the most effective disguise upon the spectrographic speaker identification, and nasal disguise (naïve and sophisticated listeners) was the most effective, while slow rate disguise (naïve listeners) and hoarse disguise (sophisticated listeners) were the least effective disguises upon the speaker identification by listening. Both naive and sophisticated listeners were able to detect the presence of selected vocal disguises with a high degree of accuracy and reliability.

Hecker (1968) conducted an experiment to determined the effect of induced stress on speaker identification and reported that listeners could identify the stressful responses of some subjects with better than 90% accuracy and of others only at chance level. The test phrases from contrasting responses were analyzed with respect to level and fundamental frequency, and spectrograms of these test phrases were examined. The results of this experiment indicate that task-induced stress can produce a number of characteristic changes in the acoustic speech signal. Most of these changes are attributable to modification in the amplitude, frequency, and detailed waveform of the glottal pulses. Other changes results from differences in articulation.

Several authors done experiments on disguise voice condition in order to identify the speaker even though they didn't get the positive results still studies are going on, but till date there were no studies on electronic vocal disguise , present day several electronic vocal disguises are available in the market in various forms like pocket recorders, mobile handsets, electronic toys. This may increase the crime rate in the society. As for now in India till date there is no information about the forensic disguises. In this context conditions the present study may be useful in providing data for forensic speaker identification.

Mel Frequency Cepstral Coefficients (MFCC)

Psychophysical studies of the frequency resolving power of the human ear has motivated modelling the non-linear sensitivity of human ear to different frequencies [Holmes & Holmes, 2001]. The selective frequency response of the basilar membrane (hair spacing) acts as a bank of band pass filters equally spaced in the Bark scale. Figure 1 shows the linear spacing between 100 Hz to 1 kHz and the logarithmic spacing above 1 kHz further reduces dimensionality of frame/vector of speech. The low-frequency components of the magnitude spectrum are ignored and the useful frequency band lies between 64 Hz and half of the actual sampling frequency. This band is divided into 23 channels equidistant in Mel frequency domain.

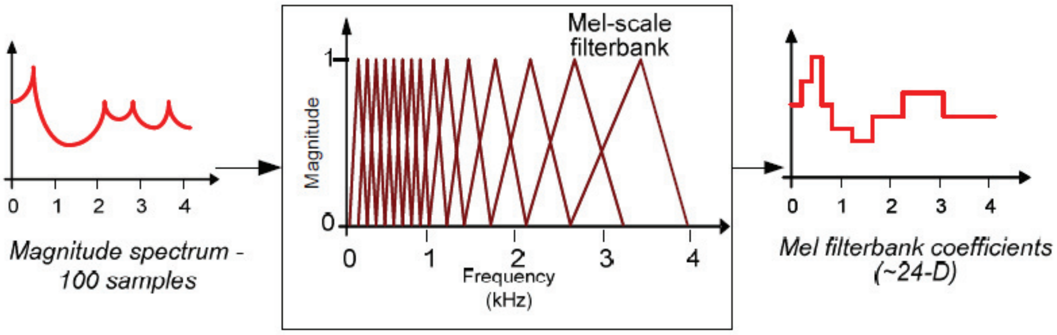

0
1
2
3
4
Magnitude spectrum - 100 samples
Mel-scale filterbank
Magnitude
1
0
0
1
2
3
4
Frequency (kHz)
0
1
2
3
4
Mel filterbank coefficients (~24-D)

Figure 5: Mel filtering [Milner, 2003].

There exists a wide range of possibilities for parameter extraction from speech frames. Mel-Frequency Cepstral Coefficients (MFCC). MFCC's are based on the known variation of the human ears critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies. In addition, rather than the speech waveforms themselves, MFCC's are shown to be less susceptible to the variation of the speaker's voice and surrounding environment. Initially, Fast Fourier Transformation (FFT) of a speech sample is extracted which is converted to Mel frequency. Cepstral coefficients are extracted on Mel frequencies.

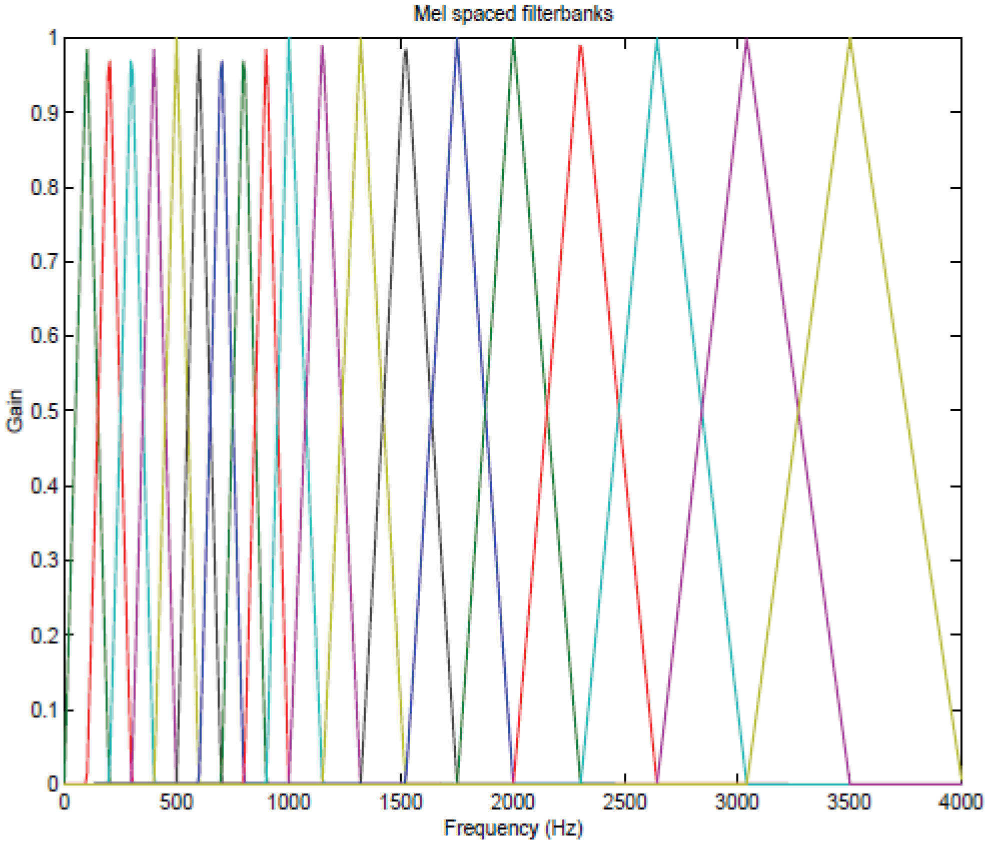


Figure 6: Mel Frequency spaced filter banks.

Glenn & Kleiner (1968) aimed to find out the validity of the hypothesis that each of the speakers produces a unique and identifiable power spectrum during nasal phonation in recognition experiments. The result obtained showed 97% of identification accuracy with /n/ nasal sound for the entire experiment.

Su, Li and Fu (1974) used an approach based on the statistical properties of nasal spectra to quantitatively study the co-articulation of nasal consonant with the vowels following them in isolated /həCVd/ utterances. Nasal spectra have been obtained from two adult males and two adult females and the result were statistically quantified by Euclidian distance. The result showed that the co-articulation between /m/ followed by vowel is much greater than that between /n/ followed by vowel. This study also showed that co-articulation was found to give more reliable clues than nasal spectra alone for the speaker identification.

Lakshmi (2009) investigated speaker identification in twenty Kannada speaking normal male subjects within the age range of 21-40 yrs were taken in her study. Nine commonly occurring meaningful Kannada words containing the long vowels /a:/, /i:/, /u:/ were taken and embedded in 3 Kannada sentences and formant feature vectors were extracted. Correct percent identification (when five speakers were considered for F2=F1) for vowel /i:/ was 70%, for vowel /a:/ 40%, and for /u:/ it had very poor percentage both in field and lab condition.

Wolf (1972) describes an investigation of an efficient approach to selecting such parameters, which are motivated by known relations between the voice signal and vocal-tract shapes and gestures. In a scheme for the mechanical recognition of speakers it, is desirable to use acoustic parameters that are closely related to voice characteristics that distinguish speakers. This study describes an investigation of an efficient approach to selecting such parameters, which are motivated by known relations between the voice signal and vocal-tract shapes and gestures. Rather than general measurements over the extent of an utterance, only significant features of selected segments are used. A simulation of a speaker recognition system as performed by manually locating speech events within utterances and using parameters measured at these locations to classify the speakers. Useful parameters were found in F0, features of vowel and nasal consonant spectra, estimation of glottal source spectrum slope, word duration, and voice onset time. These parameters were tested in speaker recognition paradigms using simple linear classification procedures. When only 17 such parameters were used, no errors were made in speaker identification from a set of 21 adult male speakers. Under the same condition, speaker verification errors of the order of 2% were also obtained.

Luck (1969) conducted an experiment on automatic speaker verification by using cepstral measurements to characterize short segments in each of the first two vowels of the standard test phrase “My code is." The length of the word "my" and the speaker's pitch were used as additional parameters. The verification decision was treated as a two-class problem, the speaker being either the authorized speaker or an impostor. Reference data was used only for the authorized speaker. The decision was based on the test sample's distance to the nearest reference sample. Data was presented to show that, if reference samples were collected over a period of many days, then verification was impossible more than two months later, whereas, if reference data was collected at one sitting, verification was highly inaccurate as little as 1 h later. Four authorized speakers and 30 impostors error rates obtained ranged from 6% to 13%. Impostors attempting to mimic the authorized speaker could not improve their ability to deceive the system significantly. This study was an automatic approach for speaker identification with only standard utterance “my code is”. But in forensic cases single utterance may not be encountered.

Bernasconi (1990) conducted a study of the effectiveness of instantaneous and transitional spectral information for text-dependent speaker verification. Instantaneous information, represented by contours of cepstral and normalized cepstral coefficients, describes the variation of the short-time speech spectra. Transitional information, represented by first-order orthogonal coefficients which result from the orthogonal polynomials expansion of the cepstral contours, describes the speed of variation of the speech spectra. Cepstral and first-order orthogonal coefficients vectors were employed separately as well as joined into a single feature vector, in classical DTW-based verification algorithms. Investigations on a population of 22 speakers (high-quality speech) showed that the elimination of the time-invariant spectral components from the speech features, taking place when performing cepstral normalization or computing first-order orthogonal coefficients, brings a substantial reliability improvement. Furthermore, transitional information is practically as effective as instantaneous information, whereas combining both kinds of information does not lead to further improvement.

Atal (1974) examined several different parameters using linear prediction model for their effectiveness for automatic recognition of speakers from their voices. He determined twelve predictor coefficients approximately once every 50 ms from speech sampled at 10 kHz. The predictor coefficients, as the impulse response function, the autocorrelation function, the area function, and the cepstrum function were used as input to an automatic speaker-recognition system. The speech data consisted of 60 utterances, consisting of six repetitions of the same sentence spoken by 10 speakers. The identification decision was based on the distance of the test sample vector from the reference vector for different speakers in the population; the speaker corresponding to the reference vector with the smallest distance was judged to be the unknown speaker. In verification, the speaker was verified if the distance between the test sample vector and the reference vector for the claimed speaker was less than a fixed threshold. He reported that the cepstrum was found to be the most effective parameter, providing an identification accuracy of 70% for speech 50 ms in duration, which increased to more than 98% for a duration of 0.5 sec. Using the same speech data, the verification accuracy was found to be approximately 83% for a duration of 50 ms, increasing to 98% for a duration of 1sec.

Jakkar (2009) studied speaker identification in twenty male Hindi speaking normal subjects age range from 21 to 38 years were considered. Nine commonly occurring, meaningful Hindi words containing the long vowels /a:/, /i:/, and /u:/ in the word medial position were considered. The direct recording was done by using an Olympus voice recorder WS-100. The mobile phone recording was done by making a call using the same network (Airtel). All the subjects used Nokia 2626 mobile hand set in usual position and the voice was recorded automatically by Nokia N72 mobile phone when call was received at the other end. Cepstrum was extracted using fast Fourier technique, two cepstral coefficients were determined namely, quefrency and amplitude. The benchmark for speaker identification using cepstrum was 88.33% (live Vs live), 81.67% (mobile Vs mobile) for speaker population of 20 in Hindi language.

Medha (2010) considered twenty Hindi speaking normal subjects (ten males and ten females) within in the age range of 25-40 years study. Nine commonly occurring meaningful Hindi words containing the long vowels /a:/, /i:/, /u:/ were taken in embedded in 6 Hindi sentences in initial, medial, final position. Cepstral coefficients were extracted. Percent correct identification for females in /a:/ 40%, /i:/ 40%, /u:/ 20% and for males /a:/ 80%, /i:/ 80%, /u:/ 20%.

Srividya (2010) used cepstrum in thirty Kannada speaking normal subjects within the age range of 35-55 years. Three long vowels /a:/, /i:/ and /u:/ embedded in three Kannada sentences were selected as stimuli. The sentences selected have six repetitions of long vowels in different context. Subjects were instructed to read the sentences twice in a normal speaking rate. Two repetitions of each of three sentences were recorded using Olympus –WS digital recorder at speaker's workplace to suit the realistic forensic situation. The recordings were done in a single session. One out of two repetitions of the tokens containing the long vowels perceived correctly as target vowel was selected. Hence for each speaker there were 18 tokens (3X6 = 18). Totally there were 180 tokens for analysis. CSL-4500 (Kay Pentax, New Jersey) was used for extracting F_2 and F_3. Results indicated higher percent correct identifications for vowel /u:/ (70%) and at chance identification (50 % identification each) for vowels /a:/ and /i:/.

Quatieri, Jankowski & Reynolds, (1994) measured onset times of resonant energy pulses are measured with the high-resolution Teager operator and used as features in the Reynolds Gaussian-mixture speaker identification algorithm. Feature sets are constructed with primary pitch and secondary pulse locations derived from low and high speech formants. Preliminary testing was performed with a confusable 40-speaker subset from the NTIMIT (telephone channel) database. Speaker identification improved from 55 to 70% correct classification when the full sets of new resonant energy-based features were added as an independent stream to conventional mel-cepstra.

Plumpe, Quatieri & Reynolds (1999) used an automatic technique for estimating and modeling the glottal flow derivative source waveform from speech, and applying the model parameters to speaker identification, is presented. The estimate of the glottal flow derivative is decomposed into coarse structure, representing the general flow shape, and fine structure, comprising aspiration and other perturbations in the flow, from which model parameters are obtained. The glottal flow derivative is estimated using an inverse filter determined within a time interval of vocal-fold closure that is identified through differences in formant frequency modulation during the open and closed phases of the glottal cycle. This formant motion is predicted by Ananthapadmanabha and Fant to be a result of time-varying and nonlinear source/vocal tract coupling within a glottal cycle. The glottal flow derivative estimate is modeled using the Liljencrants-Fant model to capture its coarse structure, while the fine structure of the flow derivative is represented through energy and perturbation measures. The model parameters are used in a Gaussian mixture model speaker identification (SID) system. Both coarse- and fine-structure glottal features are shown to contain significant speaker-dependent information. For a large TIMIT data subset, averaging over male and female SID scores, the coarse-structure parameters achieve about 60\% accuracy, the fine-structure parameters give about 40% accuracy, and their combination yields about 70% correct identification. Finally, in preliminary experiments on the counterpart telephone-degraded NTIMIT database, about a 5% error reduction in SID scores is obtained when source features are combined with traditional mel-cepstral measures.

Kinnunen(2003) indicated that the Mel-frequency cepstral coefficients (MFCC) is the most evident example of a feature set that is extensively used in speaker recognition, but originally developed for speech recognition purposes. When MFCC feature extractor is used in speaker recognition system, one makes an implicit assumption that the human hearing mechanism is the optimal speaker recognizer. Authors aimed to find the critical parameters that affect the performance and tried to give some general guidelines about the analysis parameters. He conducted experiments on two speech corpora using vector quantization (VQ) speaker modelling. The corpora are a 100 speaker subset of the American English TIMIT corpus, and a Finnish corpus consisting of 110 speakers. Although noise robustness is an important issue in real applications, it is outside the scope of this thesis. The author’s main attempt is to gain at least some understanding what is individual in the speech spectrum. The results indicate that in addition to the smooth spectral shape, a significant amount of speaker information is included in the *spectral details*, as opposed to speech recognition where the smooth spectral shape plays more important role.

Hasan, Jamil, Rabbani, & Rahman (2004) used MFCCs for feature extraction and vector quantization in security system based in speaker identification. Database consists of 21speakers, which includes 13 males and 8 female speakers. The system has been implemented in Matlab 6.1 on windows XP platform. Study shows 57.14% speaker identification for code book size of 1, 100% speaker identification for code book size of 16. Study reveals MFCC technique has been applied for speaker identification.

Mao, Cao, Murat & Tong (2006) did a study where they used linear predictive coding (LPC) parameter and Mel frequency cepstrum coefficient (MFCC) for speaker identification. Firstly, MFCC was used as the parameter and then Lempel-Ziv Complexity was combined with MFCC as parameters. The text-dependent recognition rate of 50 speakers increased from 42% to 80% and the text-independent recognition rate of 50 speakers increased from 60% to 72%. This test shows that Lempel-Ziv complexity, as a new parameter, can be applied to speaker identification.

Wang, Ohtsuka, & Nakagawa (2009) proposed a method that integrated the phase information with MFCC on a speaker identification method, and a preliminary experiment was performed. In this paper, they propose a new modified feature parameter obtained from the original phase information, and evaluated it by using speech database consisting of normal, fast and slow speaking modes. The speaker identification experiments were performed using NTT database which consists of sentences uttered by 35 Japanese speakers (22 males and 13 females) on five sessions over ten months. Each speaker uttered only 5 training utterances at a normal speaking mode (about 20 seconds in total). The proposed new phase information was more robust than the original phase information for all speaking modes. By integrating the new phase information with the MFCC, the speaker identification error rate was remarkably reduced for normal, fast and slow speaking rates in comparison with a standard MFCC-based method .The experiments show that the phase information is also very useful for the speaker verification.

Chandrika (2010) compared the performance of speaker verification system using MFCCs when recording is done with mobile handsets over a cellular network as against digital recording. Ten subjects who participated in the study were provided with words containing long vowels /a: /, /i: / and /u: /. Speakers were provided with CDMA handset (Reliance, LG). A call was made to the speaker's handset from another CDMA Reliance LG handset with recording opinion held by the experimenter. MFCC values were extracted from the speech samples obtained. The average MFCC vector over the entire segment was extracted using MATLAB coding. The formula for linear frequency to Mel frequency transformation used was constant times log (1+ f/700). Results revealed that the overall performance of speaker verification system using MFCCs is about 80% for the data base considered. The overall performance of speaker recognition is about 90% to 95% for vowel /i/. The results obtained also showed improved performance in speaker recognition for vowels /i/. The accuracy of performance for vowel /i/ is marginally better compared to vowel /a/ and /u/.

Tiwari (2010) used MFCC to extract, characterize and recognize the information about speaker identity using MFCC with different number of filters. Results showed 85% of efficiency using MFCC with 32 filters in speaker recognition task.

Ramya (2011) used Mel frequency Cepstral coefficients (MFCC) for speaker identification in their study the results indicated the percent correct identification was above chance level for electronic vocal disguise for females. Interestingly the vowel /u: / had (96.66%), /a: / 93.33 %, and /i: / 93.33%.

The review indicates that the effects of vocal disguises markedly interfere with spectrographic speaker identification as well as speaker identification by listening. The nasal and slow rate were the least effective disguises, while free disguise was the most effective disguise upon the spectrographic speaker identification, and nasal disguise (naïve and sophisticated listeners) was the most effective, while slow rate disguise (naïve listeners) and hoarse disguise (sophisticated listeners) were the least effective disguises upon the speaker identification by listening. Both naive and sophisticated listeners were able to detect the presence of selected vocal disguises with a high degree of accuracy and reliability.

However, till date there are very limited studies on electronic vocal disguise. As on to date, several electronic vocal disguises are available in the market in various forms like pocket recorders, mobile handsets and electronic toys. This may increase the crime rate in the society. With the introduction of electronic disguises on mobile phones speaker identification is still more difficult as the culprit can select any electronic disguise and speak to demand ransom. In this context, the present study examined speaker identification in electronic vocal disguise using Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC). Specifically, in MFCC a power spectrum was extracted with a filter bank uniformly spaced on the log Mel scale which is converted to a time domain to obtain the cepstrum. In LPCC a power spectra was extracted based on LPC which is converted to time domain to get the cepstrum.

Thus, the aim of the study was to establish ***Benchmark for speaker identification under electronic vocal disguise using Mel frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients. (LPCC)*** The objectives of the study were to provide benchmarks for (a) Mel-frequency cepstral coefficients for Telugu vowels in electronic vocal disguise condition, and (b) linear prediction cepstral coefficients for Telugu vowels in electronic vocal disguise condition.

CHAPTER III

Method

Participants: Ten Telugu speaking normal male subjects in the age range of 20-35 years participated in this study. The inclusion criteria of the speakers were as follows:

(a) No history of speech, language, hearing problem,

(b) Normal oral structure,

(c) No other associated psychological or neurological problem, and

(d) Reasonably free from cold or other respiratory illness at the time of recording.

Material: Commonly occurring forensically related Telugu meaningful words with long vowel /a:/, /i:/, /u:/ embedded in five sentences formed the material. Each vowel occurred thrice in 3 different words. The sentences taken were as follows:

1. /p**u:**ʤ a inṭiki inka: r**a:**le:ḓu kaḓa:? /

2. /m**i:** p**a:**pani kidnap chæs**a:**mu/

3. /po:l**i:**sulaku cheap**a:**ro: ʤ**a:**graṱa/

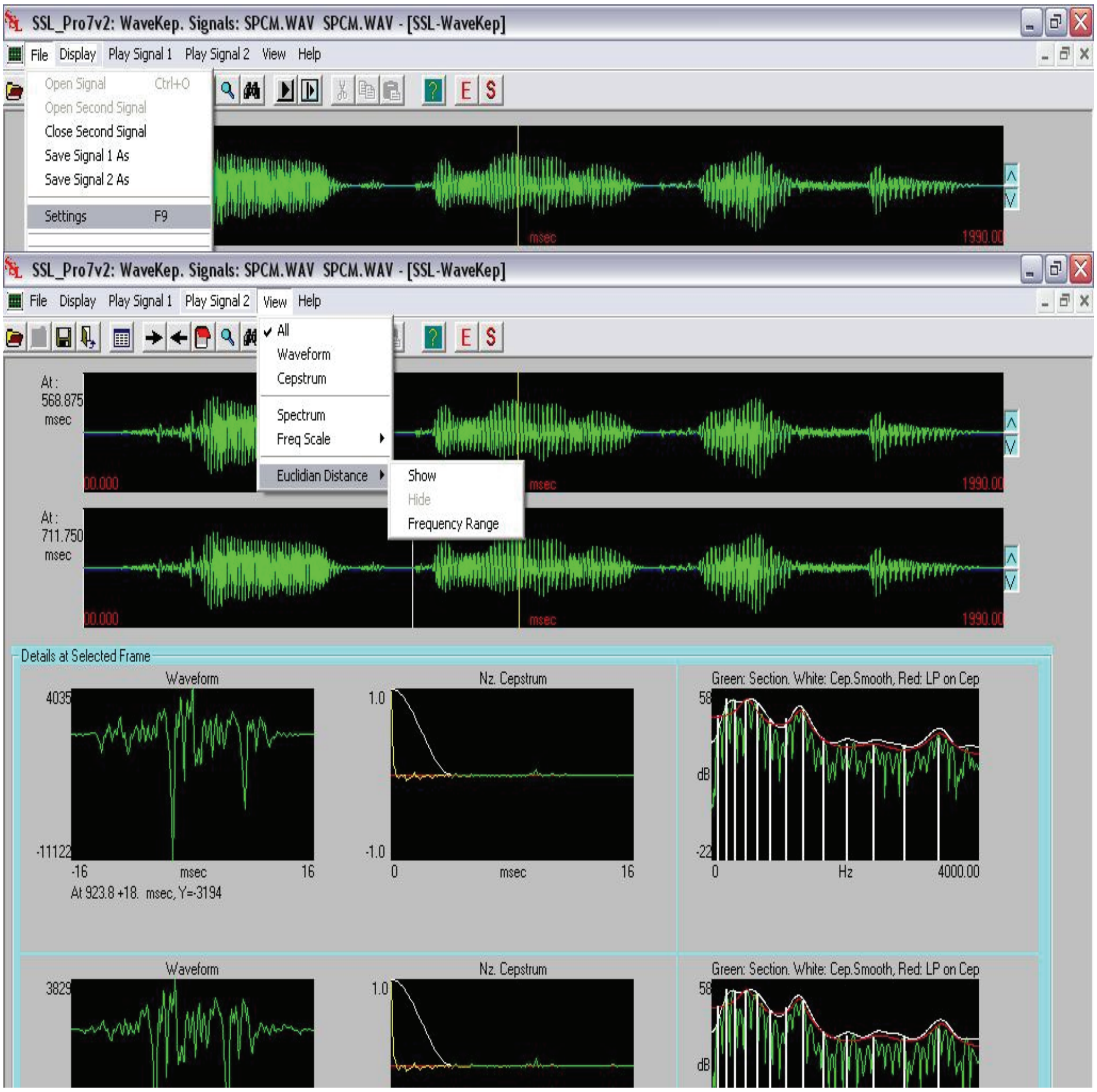
4. / mi:ru ko:ṭi r**u:**pajalu mæmu chepina cho:tiki ṱ**i:**sukuranḍi/

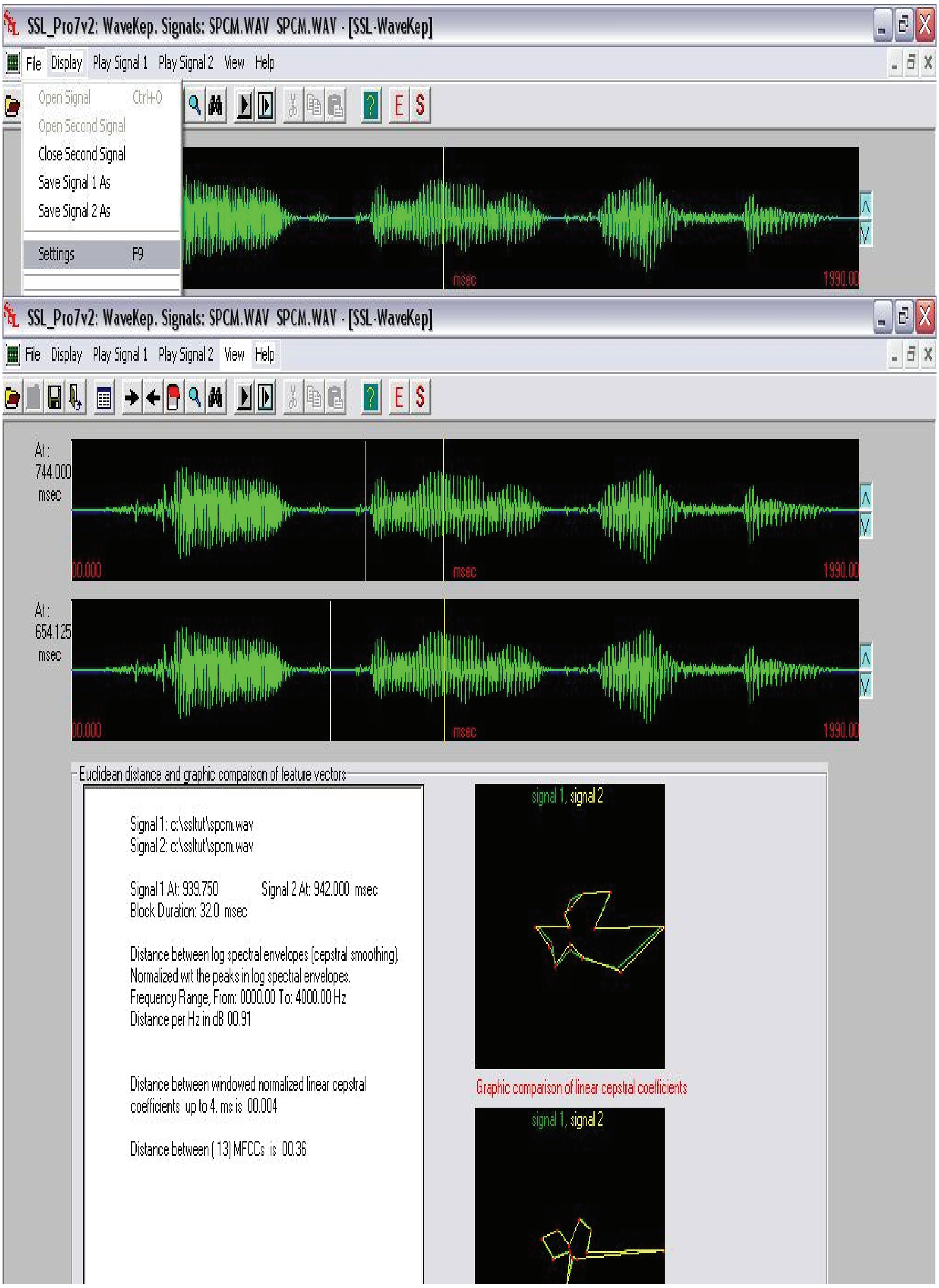
5. /m**u:**du ganṭalaku pho:nu chæsṱ**a:**mu/

Procedure: Subjects were informed about the study and their written consent was obtained. The written material was provided to the subjects and they were familiarized with the sentences. In the study two mobiles phones were used - Micromax X250 and Nokia 7210 supernova. Micromax X250 has a special option called “**MAGIC VOICE**” in which several disguise conditions exist. In the study only two electronic disguises - children voice and female voice were considered. Male subjects were instructed to speak the sentences in normal, children voice, and female voice. Subjects had the mobile MicromaxX250 and the experimenter had the NOKIA 7210 Supernova mobile. Subjects were instructed to switch on to the mode MAGIC VOICE in the mobile and to make a call to other mobile model Nokia7210 supernova and to speak the sentences thrice in 2 disguise conditions one after another by taking a 30 sec interval between calls. The speech samples were recorded and stored in the microchip of the Nokia7210 supernova mobile.

The recorded samples were transferred on to the computer memory. All the samples were converted from .amr file extension to .wav file and from stereo to mono using you-tube downloader, Adobe Audition and Praat software (Boersma & Weenink 2009). Converted samples were stored in separate folders for each subject and for each disguise condition in the system.

Analysis: SSL Pro.V4 Software (Voice and Speech Systems, Bangalore, India) and SSL WORK BENCH (Voice and Speech Systems, Bangalore, India) was used for analysis. The files were opened in Wavekep of SSL Pro V4. Second repetition of each recording was considered as test set and the third repetition was considered as training set. The second repetition of a word of the one subject was displayed as first file and the third repetition of the same word of same/different subject/ disguise was considered as the second file. The cursor was positioned at the zero cross of the complex wave in the steady state of the vowel in both the files. Cepstrum and Mel frequency cepstrum/ linear prediction of cepstral coefficients were extracted at the point of the cursor. Using the option ‘Show Euclidian distance’, Euclidian distance between linear Cepstral coefficients and between 13 MFCCs of the first and the second files was extracted and noted down. The same procedure was continued for zero crossings of 5 complex waves in the steady state of the vowel/s. Figure 7 shows two wave files with LPCC and MFCC coefficient values.


SSL_Pro7v2: WaveKep. Signals: SPCM.WAV SPCM.WAV - [SSL-WaveKep]
File Display Play Signal 1 Play Signal 2 View Help
Open Signal Ctrl+O
Open Second Signal
Close Second Signal
Save Signal 1 As
Save Signal 2 As
Settings F9
msec
1990.00
SSL_Pro7v2: WaveKep. Signals: SPCM.WAV SPCM.WAV - [SSL-WaveKep]
File Display Play Signal 1 Play Signal 2 View Help
✓ All
Waveform
Cepstrum
Spectrum
Freq Scale
Euclidian Distance
Show
Hide
Frequency Range
At :
568.875
msec
00.000
msec
1990.00
At :
711.750
msec
00.000
msec
1990.00
Details at Selected Frame
Waveform
4035
-11122
-16
msec
16
At 923.8 +18. msec, Y=-3194
Nz. Cepstrum
1.0
-1.0
0
msec
16
Green: Section, White: Cep.Smooth, Red: LP on Cep
58
dB
-22
0
Hz
4000.00
Waveform
3829
Nz. Cepstrum
1.0
Green: Section, White: Cep.Smooth, Red: LP on Cep
58
dB



```
Signal 1: c:\sslut\spcm.wav
Signal 2: c:\sslut\spcm.wav

Signal 1 At: 939.750        Signal 2 At: 942.000  msec
Block Duration: 32.0  msec

Distance between log spectral envelopes (cepstral smoothing).
Normalized wrt the peaks in log spectral envelopes.
Frequency Range, From: 0000.00 To: 4000.00 Hz
Distance per Hz in dB 00.91


Distance between windowed normalized linear cepstral
coefficients  up to 4. ms is  00.004

Distance between ( 13) MFCCs  is  00.36
```

Figure 7: Euclidean distance and graphic comparison of feature vectors (LPCC and MFCC).

The unknown subject (second repetition) was labelled as US1, US2, US3, US4, US5, US6, US7, US8, US9, and US10, and the known subjects (third repetition) were labelled as KS1, KS2, KS3, KS4, KS5, KS6, KS7, KS8, KS9, and KS10. The Euclidian distance within and between subjects were noted down and the subjects having the least Euclidian distance was considered as same. An illustration is shown in table 1.

Screen 1	**Screen 2**	**Euclidean distance**
Second repetition of US1	**Third repetition of US1**	**0.01**
	KS1	0.012
	KS2	0.78
	KS3	0.15
	KS4	0.43
	KS5	0.52
	KS6	0.31
	KS7	0.03
	KS8	0.61
	KS9	0.78
	KS10	0.61

Table 1: Illustration of Euclidean distance within and between subjects (Least Euclidian distance in bold).

If the distance between the unknown and corresponding known speaker was less than any other known speaker the identification was deemed to have been correct. If the distance between the unknown and the corresponding known speaker was more than any other known speaker then the speaker was deemed to be falsely identified. The percent correct identification was calculated using the following formula:

$$\text{Percent correct identification} = \frac{\text{Number of correct identification}}{\text{Number of total identifications}} \times 100$$

Along with SSL Pro.V4 the recently developed software (Voice and Speech Systems, Bangalore). This software was used to test the performance of distance based, semiautomatic speaker recognition system, which is vocabulary (phone) dependent. Initially the file was specified using a notepad and .dbs file that is extension of the notepad file was created. Followed by this samples for analysis were segmented. As soon as all files were segmented the software opened another window to train the samples randomly. After training, MFCC and LPCC were selected and the sample for identification was tested. Finally the software automatically generated the speaker identification threshold. This data was stored and the same procedure was repeated at least for 5 times by randomizing the training samples and the speaker identification thresholds were noted for the highest score and the lowest score.

In this study, all the speech samples are contemporary, as all the recordings of the same person were carried out in the same session. Closed set speaker identification tasks were performed, in which the examiner was aware that the ‘unknown speaker’ is one among the ‘known’ speakers.

CHAPTER IV

Results

Results of the study will be discussed under following headings:

1) MFCC and LPCC of male speakers in normal condition (Male voice Vs Male voice)
2) MFCC and LPCC of male speakers in female disguise condition (Female disguise Vs Female Disguise)
3) MFCC and LPCC of male speakers in child disguise condition (Child Disguise Vs Child Disguise)
4) Comparison of MFCC and LPCC in normal and female disguise condition,(Male voice Vs Female disguise)
5) Comparison of MFCC and LPCC in normal and child disguise condition (Male voice Vs Child disguise)

1) **MFCC and LPCC of male speakers in normal condition:** The results are discussed in two conditions for three long vowels /a: /, /i: /, /u: /. The two conditions are highest percent identification (HPI) and the lowest percent identification (LPI) for 10 speakers for MFCC and LPCC. The highest percent identification score for MFCC for long vowels /a: /, /i: /, /u: / was 90%, 100%, 60% and the lowest percent identification score when the training samples were randomized was 70%, 90%, 40%. The highest percent identification score for LPCC for long vowels /a: /, /i: /, /u: / was 80%, 80%, 60% and the lowest percent identification score was 40%, 60%, 10% when the training samples were randomized. The results indicated that the percent correct identification was better for MFCC when compared to LPCC. Tables 2 to 12 show the diagonal matrix of Euclidian distance for 3 vowels for two

conditions (highest and lowest for both MFCC and LPCC for long vowels /a:/, /i:/ and /u:/) and correct and false identifications are indicated by green and red colours in bold. Table 14 shows the overall percent correct identification for male speakers (Normal condition). 1, 2 etc. in the first row and column represent subject number.

	1	2	3	4	5	6	7	8	9	10
1	**4.025**	6.901	11.508	13.751	4.522	4.62	11.035	9.911	8.793	10.031
2	8.967	**4.905**	6.149	7.211	10.181	10.386	7.028	4.929	6.913	6.185
3	9.821	6.584	**3.585**	7.654	10.438	11.076	5.328	7.186	5.826	4.585
4	13.627	8.987	7.436	**5.227**	14.381	14.676	8.753	8.106	9.788	8.388
5	4.58	8.133	12.273	14.643	**4.274**	4.867	11.957	11.035	8.963	10.933
6	4.103	7.824	11.679	14.256	**3.816**	5.228	11.45	10.911	7.909	10.379
7	10.644	7.604	6.045	8.014	10.877	11.178	**4.385**	7.427	7.313	4.509
8	11.258	7.122	6.94	7.407	12.203	11.908	7.193	**5.037**	9.805	6.512
9	7.387	6.04	6.479	9.271	7.52	8.7	6.444	8.179	**3.733**	5.805
10	9.387	6.697	5.661	7.999	9.83	10.164	4.341	6.457	7.41	**3.936**

Table 2: Diagonal matrix for normal condition (MFCC) for HPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	**4.414**	8.881	10.659	13.274	5.173	4.435	10.292	11.965	7.968	9.671
2	8.57	**5.575**	7.019	8.858	9.33	8.586	8.008	8.454	5.891	7.517
3	11.31	6.691	**4.112**	7.828	11.231	10.391	5.902	8.031	5.035	5.234
4	14.332	7.393	6.859	**4.944**	14.83	14.205	8.087	7.873	8.093	8.099
5	4.408	10.466	11.925	14.598	4.884	**3.401**	11.487	13.641	8.587	11.181
6	4.325	10.304	12.188	14.587	4.533	**3.79**	11.69	13.303	9.102	11.111
7	11.565	6.684	5.517	7.698	11.291	10.758	**3.82**	7.419	5.552	4.742
8	10.181	**4.693**	6.75	7.415	10.656	10.389	7.261	5.719	7.095	6.15
9	8.503	8.025	7.519	10.679	8.795	7.289	7.618	11.162	**3.947**	8.166
10	10.266	5.566	4.678	7.657	10.282	9.444	**3.074**	6.754	4.364	3.759

Table 3: Diagonal matrix for normal condition (MFCC) for LPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	**3.849**	6.655	18.826	19.288	4.124	4.204	16.224	13.792	12.264	14.536
2	10.33	**8.129**	14.308	14.75	10.266	10.255	12.19	10.626	9.997	10.925
3	15.526	12.903	**4.486**	8.365	15.584	15.518	4.864	15.737	6.745	5.89
4	19.967	16.955	12.402	**10.864**	19.718	19.652	12.289	14.94	14.113	12
5	**5.964**	7.68	17.878	18.741	5.994	5.994	15.148	14.611	11.223	13.758
6	5.621	7.495	18.281	19.097	5.397	**5.381**	15.575	14.387	11.785	14.216
7	17.106	14.38	8.57	10.819	17.145	17.126	**6.543**	16.671	9.47	7.661
8	21.669	19.002	16.606	**15.35**	21.545	21.527	17.115	17.348	18.42	16.371
9	11.287	9.989	9.367	11.996	11.996	11.97	7.689	15.509	**5.951**	7.365
10	16.165	13.376	7.924	9.055	16.231	16.196	6.551	15.189	8.24	**6.461**

Table 4: Diagonal matrix for normal condition (LPCC) for HPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	**3.829**	8.881	17.03	20.796	4.69	4.367	17.405	15.628	10.954	16.859
2	9.248	8.784	11.586	16.044	10.029	8.616	12.58	**11.073**	6.944	11.568
3	17.618	15.727	**4.506**	10.378	18.04	16.178	9.222	9.479	7.501	6.968
4	19.905	16.396	**12.608**	13.553	20.371	18.743	14.893	12.924	13.281	12.995
5	5.112	9.448	15.643	19.499	6.066	**5.087**	16.24	14.938	9.997	15.614
6	**4.447**	9.46	17.497	20.98	5.192	4.89	17.864	16.115	11.863	17.317
7	16.332	14.525	**5.141**	11.056	16.749	14.841	6.73	8.874	5.949	5.482
8	20.557	16.12	19.711	17.407	21.095	20.35	19.34	**14.803**	19.433	17.74
9	12.742	13.055	8.361	14.457	13.357	11.584	10.55	11.312	**5.21**	9.334
10	15.084	13.887	8.008	11.491	15.712	14.015	8.92	9.669	**7.074**	7.536

Table 5: Diagonal matrix for normal condition (LPCC) for LPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	**3.858**	9.876	8.329	10.155	6.078	11.15	14.409	7.641	11.159	10.518
2	10.082	**4.435**	6.369	6.784	12.838	14.645	14.687	9.566	10.434	5.762
3	9.175	4.876	**4.625**	5.881	12.355	12.792	12.974	7.433	8.53	5.083
4	9.827	8.099	5.55	**4.791**	13.393	12.373	12.343	6.697	7.319	6.957
5	6.001	11.773	10.469	11.937	**3.841**	9.695	12.879	9.155	11.793	12.18
6	8.796	11.001	9.604	10.654	8.919	**8.701**	11.324	9.809	9.861	11.654
7	15.561	15.022	13.222	13.196	15.431	9.748	**5.219**	12.936	10.334	13.517
8	6.374	8.802	5.77	6.807	10.545	11.511	12.607	**4.134**	8.191	7.183
9	11.269	13.11	8.995	9.926	13.714	7.265	9.126	8.247	**4.767**	11.712
10	10.466	7.286	5.663	6.01	13.782	13.296	12.279	6.9	8.199	**5.119**

Table 6: Diagonal matrix for normal condition (MFCC) for HPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	**4.093**	9.458	8.436	10.696	5.521	8.673	14.318	7.767	10.857	10.943
2	10.458	**4.343**	6.212	7.146	12.316	11.288	14.356	10.271	11.085	6.063
3	8.956	4.913	**4.125**	5.901	11.617	9.477	12.722	7.422	8.465	6.083
4	9.501	7.697	5.551	**4.8**	12.255	9.026	12.069	7.162	7.201	7.902
5	6.282	11.989	11.007	13.018	**3.563**	7.886	12.655	9.415	11.581	12.891
6	10.699	13.567	11.905	13.584	10.534	**8.393**	11.821	11.689	10.933	14.435
7	15.529	15.211	13.184	13.292	15.596	9.905	**5.234**	13.189	9.756	13.584
8	6.41	7.942	5.491	6.652	9.949	8.562	12.179	**4.219**	7.494	7.787
9	11.526	12.395	9.145	9.774	13.837	7.21	9.791	8.91	**4.399**	12.365
10	10.473	7.548	**5.623**	5.776	13.538	10.732	12.697	6.898	8.059	6.726

Table 7: Diagonal matrix for normal condition (MFCC) for LPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	**8.204**	31.654	35.972	38.745	10.73	12.661	28.91	27.964	26.72	39.876
2	43.113	24.241	26.545	23.56	46.747	37.479	30.727	29.155	30.264	**20.417**
3	43.715	17.403	**12.044**	12.888	47.985	39.95	33.552	20.094	26.844	13.653
4	37.297	25.774	23.637	**16.407**	40.517	34.095	28.769	23.358	21.435	19.938
5	5.275	35.606	40.277	41.961	**2.941**	13.09	29.044	32.309	29.164	42.998
6	14.623	34.07	38.485	39.019	15.417	**14.332**	26.488	31.005	24.846	40.007
7	25.903	30.379	35.522	32.79	27.225	21.159	**13.337**	30.488	21.119	31.378
8	29.723	23.424	20.602	18.69	32.938	28.38	27.577	**16.823**	17.8	23.336
9	33.123	23.178	22.13	19.472	36.334	29.077	26.326	19.998	**17.394**	21.792
10	43.943	25.917	23.189	**15.594**	46.622	40.005	31.047	25.961	25.516	17.671

Table 8: Diagonal matrix for normal condition (LPCC) for HPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	**6.825**	36.489	40.368	39.602	9.059	12.933	28.83	23.723	26.665	43.715
2	39.181	**21.822**	24.957	29.748	42.062	37.079	30.614	29.241	32.854	27.069
3	40.294	15.25	**12.894**	16.489	43.204	39.496	32.084	18.985	26.438	18.485
4	37.954	20.679	24.148	19.467	40.573	36.707	29.473	24.029	25.9	**18.436**
5	5.978	41.054	45.878	42.714	**2.112**	10.742	29.697	28.554	29.318	46.769
6	**16.295**	35.433	40.926	37.845	18.113	17.157	25.837	26.271	23.562	41.392
7	25.549	30.479	38.052	33.773	26.49	22.454	**13.382**	27.798	22.429	33.884
8	32.85	19.741	21.006	17.387	35.265	32.574	28.581	**14.732**	20.546	22.828
9	34.633	20.355	23.304	**19.469**	36.94	32.228	26.229	20.037	20.654	21.626
10	42.455	20.194	23.188	**18.133**	44.821	41.198	31.593	27.276	29.046	18.516

Table 9: Diagonal matrix for normal condition (LPCC) for LPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	**3.413**	5.318	6.433	5.567	6.761	6.647	6.428	6.856	9.828	6.015
2	11.021	**7.595**	12.524	10.66	10.717	10.306	11.147	9.509	12.43	9.378
3	9.09	8.799	6.899	7.716	7.285	5.727	**6.662**	6.792	6.882	8.165
4	5.08	5.483	7.158	**3.204**	5.973	4.213	5.953	5.42	7.954	5.476
5	**4.732**	5.874	8.37	6.151	7.062	7.101	7.948	8.21	10.823	6.538
6	7.884	8.345	6.985	5.812	5.946	5.276	6.066	5.963	**5.021**	5.982
7	9.524	8.507	8.107	7.961	7.656	7.424	6.191	**5.118**	7.624	7.134
8	5.483	5.801	5.142	5.04	6.445	4.851	3.76	**2.566**	6.277	5.579
9	8.381	9.118	5.367	6.36	7.551	6.697	6.332	6.258	**4.248**	7.149
10	5	4.956	7.324	5.011	4.875	5.156	5.771	5.576	8.636	**4.36**

Table 10: Diagonal matrix for normal condition (MFCC) for HPI of /u:/.

	1	2	3	4	5	6	7	8	9	10
1	**4.064**	5.937	6.54	6.456	4.69	4.884	5.625	7.946	9.839	6.708
2	10.72	**6.939**	12.644	10.938	10.687	9.511	10.265	10.251	12.862	8.008
3	7.893	8.867	7.016	6.828	8.777	7.782	6.664	**5.983**	7.634	8.575
4	4.203	5.664	7.042	3.816	4.808	**3.354**	5.003	5.524	8.137	6.074
5	4.941	6.228	8.404	7.245	**4.374**	5.079	6.975	8.919	10.774	7.348
6	6.804	8.24	6.804	5.345	7.066	5.757	5.91	5.456	**4.57**	6.763
7	8.818	8.188	8.175	7.09	9.558	8.15	6.37	**5.514**	7.81	6.175
8	5.389	5.937	5.239	4.482	6.663	4.88	**3.6**	3.726	6.618	5.233
9	7.426	9.001	5.037	6.237	8.523	6.744	6.47	6.618	**4.469**	7.714
10	4.439	4.969	7.4	5.355	**3.72**	3.894	4.932	6.13	8.553	4.863

Table 11: Diagonal matrix for normal condition (MFCC) for LPI of /u:/.

	1	2	3	4	5	6	7	8	9	10
1	5.604	18.373	10.391	3.67	**2.947**	6.296	12.969	7.068	5.999	13.295
2	25.503	**21.791**	29.987	26.215	28.403	25.489	23.886	25.481	28.07	25.288
3	10.765	18.649	**7.778**	11.599	13.728	11.718	14.101	11.166	9.574	17.347
4	5.431	17.118	10.553	3.452	**2.967**	5.845	11.291	6.372	6.066	11.746
5	7.012	19.501	11.231	4.729	**2.906**	7.464	14.17	7.673	7.408	14.099
6	**6.838**	16.327	9.581	7.96	9.061	7.367	12.005	8.395	7.806	13.915
7	22.682	18.475	24.478	23.295	24.654	22.735	**17.75**	22.257	23.208	19.429
8	5.214	15.848	9.122	5.295	7.157	4.983	10.071	**4.219**	6.864	11.884
9	8.221	18.75	6.623	8.697	9.755	8.345	13.69	9.402	**5.906**	15.379
10	7.021	18.048	10.713	5.487	**4.744**	7.211	12.365	8.221	6.686	12.738

Table 12: Diagonal matrix for normal condition (LPCC) for HPI of /u:/.

	1	2	3	4	5	6	7	8	9	10
1	**2.479**	3.939	7.626	2.997	3.978	6.27	5.238	6.249	7.924	5.97
2	35.113	33.518	37.272	33.788	33.361	33.102	33.552	33.178	34.224	**32.958**
3	15.077	15.202	11.001	13.206	12.437	12.91	12.122	12.685	**9.734**	12.476
4	**1.879**	3.98	7.611	3.183	4.432	6.825	5.455	6.246	8.312	5.799
5	**3.729**	5.877	8.082	5.755	6.7	8.16	7.265	7.894	9.318	8.133
6	10.305	10.324	10.296	8.935	**8.073**	9.374	9.28	9.684	9.804	9.258
7	29.727	29.107	30.236	27.853	27.501	27.651	27.269	26.551	29.369	**25.74**
8	6.024	5.057	8.929	4.665	4.906	**3.154**	3.753	4.09	6.792	6.414
9	9.221	10.385	**5.811**	8.07	7.93	9.181	8.33	9.037	6.996	8.462
10	**3.908**	5.148	8.884	4.831	5.609	7.583	6.604	7.152	8.858	6.432

Table 13: Diagonal matrix for normal condition (LPCC) for LPI of /u:/.

		/a:/	**/i:/**	**/u:/**
MFCC	Highest PCI	90%	100%	60%
	Lowest PCI	70%	90%	40%
LPCC	Highest PCI	80%	80%	60%
	Lowest PCI	40%	60%	10%

Table 14: Percent correct identification for Normal condition.

2) **MFCC and LPCC in male speakers with female disguise condition:** The results are discussed in two conditions for three long vowels /a: /, /i: /, /u: /. The two conditions are highest percent identification (HPI) and the lowest percent identification (LPI) for 10 speakers using MFCC and LPCC. The highest percent identification score for MFCC for long vowels /a: /, /i: /, /u: / was 80%, 80%, 100%, respectively and the lowest percent identification score was 40%, 80%, 30% respectively when the training samples were randomized The highest percent identification score for LPCC for long vowels /a: /, /i: /, /u: / was 50%, 70%, 70% and the lowest percent identification score for randomized training samples was 10%, 30%, 20%. The results indicated that the percent correct identification was higher for MFCC compared to LPCC. Tables 15 to 26 show the diagonal matrix of Euclidian distance for 3 vowels for two conditions (highest and lowest for both MFCC and LPCC). Correct

and false identifications are indicated by red and green colours in bold. Table 27 shows the overall percent correct identification for female disguise condition.

	1	2	3	4	5	6	7	8	9	10
1	5.172	5.831	7.124	5.255	7.581	6.792	**4.119**	6.774	4.566	6.69
2	5.042	**2.92**	7.944	7.542	5.923	5.721	7.4	8.307	6.796	8.575
3	9.679	10.328	**5.554**	8.908	10.106	9.607	8.648	9.343	9.452	8.812
4	10.763	11.274	10.159	7.506	10.677	11.151	7.445	7.532	9.043	**7.324**
5	8.381	9.096	9.252	8.181	**6.599**	8.568	7.911	8.647	9.842	8.442
6	6.463	6.221	9.046	9.423	4.588	**3.957**	10.431	9.72	9.669	9.917
7	8.831	8.95	7.603	4.845	9.42	9.479	**3.452**	5.166	5.141	4.37
8	9.026	9.033	8.776	6.164	9.156	8.601	6.002	**5.063**	6.574	5.157
9	6.674	5.863	7	4.744	8.057	7.059	4.55	5.997	**3.385**	5.462
10	8.455	8.552	7.539	5.13	8.221	7.836	5.26	4.617	5.385	**3.027**

Table 15: Diagonal matrix for female disguise condition (MFCC) for HPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	6.932	8.142	6.065	3.931	7.886	6.769	**3.125**	6.089	5.111	5
2	5.345	4.756	7.271	6.262	5.524	5.463	5.343	5.049	5.349	6.85
3	12.281	12.862	9.221	11.783	11.848	11.465	9.81	10.908	11.426	9.393
4	13.306	14.902	12.049	10.082	13.001	13.125	9.077	10.655	12.047	**8.283**
5	10.275	12.215	11.096	9.541	9.902	10.471	**8.52**	10.067	11.146	9.041
6	7.538	7.549	9.479	10.035	**4.778**	5.212	9.496	7.637	8.914	9.55
7	10.361	11.629	8.991	6.851	10.666	10.45	5.226	7.864	8.341	5.229
8	11.692	12.858	9.834	8.001	11.4	10.608	6.909	8.321	9.535	**5.94**
9	7.469	7.808	6.056	4.021	7.881	7.307	**2.112**	4.44	4.065	2.667
10	10.273	11.459	8.876	7.786	9.733	9.112	6.243	7.426	8.485	4.958

Table 16: Diagonal matrix for female disguise condition (MFCC) for LPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	6.872	11.459	10.647	12.997	9.917	11.38	9.126	12.163	**6.454**	12.509
2	8.271	**4.611**	17.222	19.843	5.76	8.966	16.299	17.453	13.743	20.203
3	18.807	23.089	10.264	12.482	21.933	20.807	11.646	12.464	12.577	**8.765**
4	24.559	27.583	19.061	18.584	26.777	25.766	18.748	20.202	19.996	**17.772**
5	11.415	12.719	19.79	18.528	**11.129**	15.13	17.665	17.939	15.79	20.958
6	11.515	**5.515**	19.77	22.865	7.993	8.251	19.235	20.578	16.59	23.043
7	17.419	21.786	10.856	10.258	20.332	19.634	8.334	13.457	10.586	**7.692**
8	16.961	19.987	13.733	14.877	19.05	20.404	14.587	**12.268**	14.53	13.267
9	11.977	14.23	7.624	12.804	14.054	12.432	7.517	11.391	**6.972**	9.093
10	16.958	21.728	10.915	9.646	20.068	20.274	9.536	11.724	11.219	**7.092**

Table 17: Diagonal matrix for female disguise condition (LPCC) for HPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	8.566	11.501	10.482	8.414	10.51	12.963	**6.022**	8.661	6.508	7.733
2	8.86	7.368	17.018	15.246	6.414	8.585	13.772	**5.303**	11.91	17.012
3	23.229	25.271	16.696	15.449	25.003	25.772	15.328	23.418	17.834	**12.085**
4	33.416	36.018	28.978	26.524	35.667	37.361	26.615	34.195	29.704	**24.412**
5	18.745	19.716	22.084	20.711	17.962	18.302	18.012	**15.691**	18.317	19.313
6	11.39	**7.509**	18.578	17.406	7.54	9.204	16.744	8.825	14.343	20.239
7	21.051	23.686	17.328	13.769	23.16	24.143	13.51	22.483	16.647	**11.408**
8	25.263	26.649	16.701	18.076	26.784	27.745	18.099	24.149	19.698	**14.587**
9	12.819	14.892	8.185	5.661	14.703	16.084	5.367	13.843	7.355	**3.585**
10	21.33	23.693	15.614	14.401	23.302	24.768	13.695	21.8	16.505	**11.248**

Table 18: Diagonal matrix for female disguise condition (LPCC) for LPI /a:/.

MFCC	1	2	3	4	5	6	7	8	9	10
1	15.398	9.614	9.384	**5.523**	8.472	11.274	6.499	7.5	10.117	9.183
2	15.901	**3.422**	9.336	8.668	7.231	9.006	7.498	7.469	12.355	6.504
3	14.259	6.281	5.986	5.082	6.84	8.247	5.47	5.071	8.569	**4.634**
4	14.938	7.754	8.686	**3.445**	7.475	9.398	4.677	5.073	9.148	6.739
5	11.591	7.821	7.576	8.835	**3.18**	6.712	7.24	10.623	10.5	10.636
6	7.31	10.209	6.113	9.854	6.242	**3.441**	7.618	11.858	7.262	11.18
7	12.521	7.206	6.56	5.363	6.31	7.103	**1.898**	7.652	8.151	7.274
8	15.554	7.962	8.332	4.987	8.539	10.056	6.574	**3.977**	9.325	6.105
9	9.159	13.149	6.804	10.098	11.044	8.123	9.311	12.014	**4.374**	11.308
10	16.097	8.319	9.081	8.446	10.495	10.2	8.689	7.054	10.472	**4.373**

Table 19: Diagonal matrix for female disguise condition (MFCC) for HPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	17.786	9.161	10.955	**5.057**	10.464	12.818	5.759	6.321	11.785	8.019
2	17.835	**3.601**	10.559	8.123	7.561	10.171	7.828	7.372	12.7	6.766
3	15.328	5.816	6.313	4.45	8.056	8.829	5.188	**3.934**	8.934	4.078
4	16.485	7.568	9.462	**3.255**	8.978	10.599	4.05	3.727	9.739	6.118
5	12.246	8.368	7.869	7.461	**3.421**	5.771	5.7	7.155	8.798	9.531
6	8.71	10.864	7.37	9.079	7.071	**3.548**	7.107	8.506	5.532	10.754
7	13.152	7.819	6.904	4.686	7.619	7.705	**2.798**	5.393	7.729	6.869
8	17.897	8.103	9.273	4.81	11.468	12.333	6.757	**4.776**	10.587	5.514
9	10.823	14.205	7.461	9.96	13.643	10.112	9.641	10.442	**5.315**	10.989
10	17.37	8.584	9.018	7.794	11.977	11.59	8.839	7.649	10.743	**4.502**

Table 20: Diagonal matrix for female disguise condition (MFCC) for LPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	30.197	34.103	31.344	24.957	34.858	39.475	21.858	**21.567**	28.801	26.05
2	34.187	**11.576**	27.212	25.58	23.569	28.698	29.636	29.127	37.676	42.075
3	35.104	19.739	28.364	**16.876**	29.02	36.381	19.831	17.337	36.304	34.829
4	29.816	26.161	26.25	**13.223**	27.501	32.827	14.125	18.302	31.582	36.944
5	20.628	21.506	17.881	21.497	**8.96**	9.378	23.508	31.833	28.305	48.109
6	18.555	24.408	16.588	23.103	11.157	**8.435**	23.906	32.63	26.367	48.715
7	29.067	29.786	28.396	17.199	33.922	39.19	13.101	**7.878**	26.384	23.469
8	31.048	30.376	29.391	20.831	33.026	38.168	19.283	**16.179**	28.619	27.864
9	21.929	36.004	23.605	33.251	29.577	29.154	32.009	33.738	**19.975**	37.025
10	34.779	35.617	35.404	33.069	41.749	45.947	31.24	23.905	31.426	**14.967**

Table 21: Diagonal matrix for female disguise condition (LPCC) for HPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	29.311	40.571	28.404	24.548	40.583	44.079	24.769	29.5	32.906	**22.423**
2	32.409	**8.052**	22.967	28.42	27.37	32.286	23.182	24.877	26.743	25.299
3	31.875	20.161	18.654	18.882	29.745	35.9	**15.58**	16.421	26.444	25.075
4	26.593	29.311	17.898	13.895	28.717	33.668	14.733	**12.172**	25.205	29.158
5	15.615	22.8	16.398	25.32	**7.525**	10.914	20.22	17.771	14.959	36.093
6	14.241	25.898	17.308	26.551	**9.028**	9.862	21.986	18.961	14.525	38.328
7	25.203	32.466	18.539	**10.442**	34.245	38.574	12.821	16.962	25.891	18.388
8	30.749	36.696	24.899	**16.759**	40.06	43.875	20.099	23.574	31.275	19.767
9	**24.557**	41.468	29.675	36.244	35.638	35.066	33.63	35.106	26.531	34.6
10	33.764	38.625	31.413	31.663	43.464	46.429	30.307	34.736	34.292	**16.779**

Table 22: Diagonal matrix for female disguise condition (LPCC) for LPI of /i:/.

MFCC	1	2	3	4	5	6	7	8	9	10
1	**5.09**	7.58	6.646	7.07	5.387	7.227	9.119	8.142	9.614	9.307
2	5.251	**3.484**	8.182	7.774	4.655	5.849	8.068	6.805	9.861	7.74
3	7.357	7.93	**4.37**	5.494	5.677	6.749	7.333	6.947	7.036	7.517
4	8.174	7.811	3.869	**3.504**	6.779	6.511	5.846	5.443	5.789	6.499
5	5.896	5.138	5.774	5.766	**3.202**	5.483	6.805	5.767	8.224	6.853
6	7.115	5.608	6.447	5.751	6.233	**3.167**	5.467	4.592	5.799	6.044
7	8.085	6.482	6.666	5.525	7.177	5.86	**4.057**	4.836	6.342	6.385
8	7.968	6.01	5.712	4.791	6.647	4.362	4.88	**3.323**	5.595	5.448
9	9.213	6.92	6.012	5.381	7.821	5.153	4.697	4.525	**4.51**	5.382
10	9.538	7.15	6.478	5.965	7.767	6.253	6.872	6.229	6.671	**4.003**

Table 23: Diagonal matrix for female disguise condition (MFCC) for HPI of /u:/.

	1	2	3	4	5	6	7	8	9	10
1	**4.589**	8.863	8.266	6.593	5.165	11.794	8.252	7.919	10.64	11.761
2	5.625	6.516	10.266	7.155	**4.378**	13.455	10.085	8.103	12.876	13.253
3	8.625	12.017	**8.061**	8.538	8.685	12.625	8.788	10.23	10.629	11.029
4	8.977	11.367	**4.754**	6.678	9.241	12.332	5.181	8.805	10.539	7.857
5	6.906	7.701	8.744	6.336	**5.918**	13.379	8.606	8.21	12.468	11.8
6	**4.777**	6.606	9.316	6.634	5.16	10.089	8.867	5.646	9.23	11.685
7	16.295	15.298	19.446	16.936	16.343	20.558	18.697	**14.851**	20.137	19.865
8	6.325	7.386	7.581	**4.916**	5.544	10.773	7.082	6.386	9.5	10.665
9	9.543	8.147	11.102	9.477	9.558	12.2	10.685	**6.615**	11.303	11.123
10	10.574	14.277	**8.755**	10.491	11.506	12.066	9.39	10.953	9.465	10.25

Table 24: Diagonal matrix for female disguise condition (MFCC) for LPI of /u:/.

	1	2	3	4	5	6	7	8	9	10
1	**5.017**	7.331	7.231	6.046	5.363	6.836	9.015	7.17	12.399	10.116
2	4.943	4.045	9.961	8.53	**3.559**	6.439	9.5	6.034	14.54	12.194
3	9.281	10.871	**6.531**	7.898	9.006	10.139	11.06	9.212	11.78	10.256
4	9.13	10.191	6.583	**5.336**	8.666	10.066	9.026	7.79	12.325	9.678
5	6.086	5.959	8.791	7.5	**4.385**	7.545	9.599	6.293	14.324	11.784
6	8.899	10.318	12.411	11.71	10.301	**7.798**	11.555	9.382	12.593	13.483
7	14.936	14.775	16.972	15.391	15.582	15.002	**12.611**	14.358	17.705	15.019
8	7.259	7.338	9.263	8.164	6.866	7.563	8.247	**5.786**	12.28	11.205
9	9.31	9.108	10.297	9.947	10.346	8.573	8.062	**7.626**	10.259	9.678
10	13.124	14.781	**10.436**	11.39	13.906	12.956	12.573	12.56	11.004	10.556

Table 25: Diagonal matrix for female disguise condition (LPCC) for HPI of /u:/.

	1	2	3	4	5	6	7	8	9	10
1	4.054	8.329	11.195	6.654	**3.815**	11.512	9.137	8.366	10.942	15.964
2	5.889	5.712	12.864	8.703	**3.271**	12.758	11.391	8.261	13.19	17.219
3	7.964	11.362	9.632	7.78	**7.575**	12.104	9.591	10.888	10.843	15.533
4	8.207	10.996	5.887	**5.496**	9.729	11.535	5.554	10.018	10.605	11.746
5	6.751	7.019	10.817	7.572	5.514	12.475	9.947	8.768	12.712	15.84
6	**5.164**	6.033	12.01	7.886	5.638	10.049	9.268	5.166	9.587	14.855
7	16.618	15.552	21.637	17.944	17.296	20.591	18.923	**14.819**	20.333	20.374
8	6.065	6.815	9.566	6.196	**6.051**	10.246	8.081	6.928	9.671	14.389
9	9.872	8.093	13.101	10.604	10.943	11.79	10.773	**6.544**	11.547	12.264
10	9.835	13.832	10.134	**8.673**	10.917	12.22	8.976	11.487	9.595	13.324

Table 26: Diagonal matrix for female disguise condition (LPCC) for LPI of /u:/.

		/a:/	**/i:/**	**/u:/**
MFCC	Highest PCI	80%	80%	100%
	Lowest PCI	40%	80%	30%
LPCC	Highest PCI	50%	70%	70%
	Lowest PCI	60%	30%	10%

Table 27: Percent correct identification for female disguise condition.

3) **MFCC and LPCC of male speakers in child disguise condition:** The results are discussed in two conditions for three long vowels /a: /, /i: /, /u: /. The two conditions are highest percent identification (HPI) and the lowest percent identification (LPI) for 10 speakers on MFCC and LPCC. The HPI on MFCC for long vowels /a: /, /i: /, /u: / was 90%, 100%, 80% and the LPI was 30%, 90%, 40%, respectively for randomized training samples. The HPI on LPCC for long vowels /a: /, /i: /, /u: / was 60%, 80%, 60% and the LPI was 60%, 30%, 10%, respectively for randomized training samples. The results indicated that the percent correct identification was higher on MFCC compared to LPCC. Tables 28 to 39 show the diagonal matrix of Euclidian distance for 3 vowels for two conditions (highest and lowest on MFCC and LPCC). Correct and false identifications are indicated by red and green colours in bold. Table 40 shows the overall percent correct identification for female disguise condition.

	1	2	3	4	5	6	7	8	9	10
1	**4.131**	6.916	7.002	6.58	6.644	6.057	5.463	6.585	8.338	6.447
2	6.615	**5.891**	8.371	8.05	8.735	7.606	7.182	6.396	7.991	8.324
3	5.433	6.719	**3.038**	4.093	10.389	6.063	5.6	6.841	5.604	4.467
4	9.61	11.649	9.674	**8.037**	11.208	9.794	8.736	9.452	10.404	8.391
5	8.852	11.485	13.645	11.713	**4.714**	9.872	10.118	10.736	13.929	11.87
6	6.334	7.812	7.898	6.943	6.925	**4.244**	8.223	6.558	9.257	8.764
7	5.675	7.012	6.669	6.248	9.368	8.191	**4.094**	7.381	6.909	4.723
8	9.265	7.693	7.718	7.759	12.468	8.676	9.024	6.67	**6.164**	9.059
9	5.92	4.893	6.662	6.023	9.402	7.482	5.236	5.317	**4.526**	6.673
10	7.927	9.784	8.024	6.47	10.232	7.892	6.664	7.847	8.899	**5.574**

Table 28: Diagonal matrix for child disguise condition (MFCC) for HPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	5.964	8.154	8.492	9.399	7.125	6.364	6.55	6.455	9.637	**5.777**
2	6.072	6.168	8.12	8.11	7.918	6.898	5.721	**4.786**	7.431	6.62
3	5.404	7.244	3.961	5.405	9.782	5.445	4.444	5.818	5.798	**3.377**
4	10.753	13.882	11.842	12.148	11.865	10.673	10.396	10.88	12.831	**8.931**
5	10.04	12.508	14.934	14.589	**6.416**	10.727	11.158	10.771	14.859	10.699
6	7.586	8.482	8.743	8.788	6.567	**4.885**	8.52	5.554	9.637	7.893
7	6.148	8.682	8.08	8.87	9.603	8.08	4.659	7.458	8.671	**4.03**
8	10.016	10.386	8.873	7.869	12.761	9.469	8.652	7.478	**7.184**	9.598
9	5.848	6.361	7.542	6.863	9.372	7.603	4.063	5.195	5.694	5.839
10	9.713	11.674	9.447	9.298	12.029	9.097	**7.93**	9.093	9.819	**6.753**

Table 29: Diagonal matrix for child disguise condition (MFCC) for LPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	13.866	14.946	21.842	19.226	16.232	17.83	18.985	**11.292**	16.08	22.224
2	11.561	**7.704**	21.726	17.773	8.004	11.507	18.856	7.793	11.111	21.174
3	14.04	17.06	**5.231**	7.13	24.917	19.169	10.868	14.857	13.958	9.937
4	25.998	31.626	28.142	26.282	35.161	35.184	**21.174**	29.754	29.651	24.858
5	17.323	12.813	28.366	23.876	**4.109**	14.498	24.426	13.338	16.164	26.627
6	16.795	12.621	20.716	18.142	13.503	**8.693**	21.432	11.836	12.136	21.024
7	17.76	24.445	18.869	17.179	29.292	28.279	**10.635**	22.031	21.972	15.685
8	**13.28**	15.257	18.289	16.26	20.762	20.213	16.251	13.901	16.595	20.242
9	**8.998**	11.259	18.35	15.293	14.664	15.852	14.152	10.018	12.404	17.88
10	25.444	28.817	18.848	19.659	33.767	30.402	18.373	25.843	25.434	**17.615**

Table 30: Diagonal matrix for child disguise condition (LPCC) for HPI of /a:/.

	1	2	3	4	5	6	7	8	9	10
1	**11.305**	15.014	19.897	19.293	17.149	14.909	20.143	12.795	16.565	23.867
2	**5.614**	7.145	19.643	17.932	8.672	11.111	21.199	6.863	11.925	23.547
3	17.877	17.866	**4.583**	9.017	24.766	12.768	13.914	16.27	12.044	13.37
4	29.717	30.845	29.022	25.129	34.583	30.948	**19.864**	29.594	27.988	23.475
5	10.744	10.755	27.007	24.048	**4.051**	16.892	26.572	11.494	17.337	28.703
6	12.016	10.684	20.108	19.785	11.298	**9.322**	25.732	10.217	12.238	24.791
7	22.593	24.106	20.082	15.305	29.053	23.234	**9.523**	22.635	19.326	16.15
8	**15.422**	16.978	16.285	16.762	22.234	16.201	16.752	15.532	16.11	20.77
9	**6.617**	8.905	17.976	14.789	13.025	12.489	16.609	8.393	10.998	21.241
10	29.025	29.149	21.051	20.888	34.135	25.755	19.977	27.41	23.965	**14.802**

Table 31: Diagonal matrix for child disguise condition (LPCC) for LPI of/a:/.

	1	2	3	4	5	6	7	8	9	10
1	**4.074**	8.159	7.75	5.992	6.841	8.321	6.241	6.877	11.423	8.237
2	8.737	**3.457**	9.474	9.97	5.272	7.812	5.641	8.687	15.031	12.016
3	6.892	6.705	**6.244**	7.975	6.269	6.955	6.953	6.514	11.295	9.548
4	5.446	10.146	8.151	**3.431**	9.564	9.597	7.659	7.227	9.906	7.411
5	7.725	4.518	8.915	9.687	**3.311**	6.382	6.585	6.888	14.748	12.65
6	9.042	7.462	6.767	9.961	6.946	**3.444**	7.625	6.49	10.956	10.808
7	5.369	4.834	7.383	6.755	5.593	6.99	**2.792**	6.94	11.872	8.049
8	6.284	7.691	7.767	7.504	6.349	5.918	7.466	**3.739**	11.412	10.813
9	10.825	14.136	7.861	10.106	13.592	10.681	12.236	10.134	**4.328**	8.651
10	6.619	10.62	9.037	6.453	10.379	11.205	7.529	10.016	11.099	**5.613**

Table 32: Diagonal matrix for child disguise condition (MFCC) for HPI of /i:/.

MFCC	1	2	3	4	5	6	7	8	9	10
1	**3.668**	6.663	7.662	6.669	6.111	9.037	5.787	6.302	10.702	6.07
2	7.381	**3.601**	9.188	10.428	5.565	10.276	5.229	10.554	15.126	10.552
3	7.184	6.054	7.126	8.95	**5.81**	9.28	7.199	8.272	12.28	9.926
4	6.079	9.165	7.455	**3.426**	9.275	10.595	7.838	6.346	8.75	5.743
5	7.467	4.426	9.508	10.812	**3.275**	8.502	7.302	7.841	14.743	11.64
6	9.986	7.177	7.16	10.346	7.047	**2.826**	7.898	8.815	11.389	11.919
7	3.802	4.851	6.658	6.277	5.997	9.048	**2.451**	7.641	10.978	5.645
8	8.042	7.84	8.629	8.947	6.681	7.038	8.686	**3.781**	11.249	10.599
9	13.58	14.081	7.952	9.772	14.165	10.262	12.695	11.978	**4.313**	12.13
10	7.948	11.827	9.226	5.904	12.117	13.673	9.014	10.912	9.739	**3.422**

Table 33: Diagonal matrix for child disguise condition (MFCC) for LPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	**11.018**	19.233	17.333	12.392	17.93	19.551	18.239	14.17	18.876	17.398
2	16.525	**3.303**	11.465	15.913	3.832	5.124	3.698	10.215	22.767	24.456
3	14.058	5.135	9.135	12.78	**4.06**	5.276	4.274	8.005	20.09	21.612
4	15.08	25.61	20.883	**12.916**	24.612	25.207	23.71	17.432	18.492	17.805
5	16.182	3.737	11.616	15.862	**3.132**	5.384	5.001	10.538	23.238	24.337
6	19.248	6.062	11.375	18.078	7.351	**4.348**	7.05	11.497	23.653	25.744
7	14.394	4.251	9.712	13.409	4.425	4.784	**3.152**	7.262	20.638	22.702
8	16.184	4.199	9.913	14.99	5.429	**3.48**	4.272	8.189	21.389	23.351
9	22.901	29.878	21.751	21.626	29.446	28.573	28.3	24.304	**17.049**	21.044
10	20.579	31.449	27.183	20.459	30.122	31.404	29.917	25.744	21.938	**17.627**

Table 34: Diagonal matrix for child disguise condition (LPCC) for HPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	19.602	13.499	10.341	12.251	12.043	16.705	12.96	**9.103**	25.828	21.848
2	30.874	**2.686**	6.315	23.097	4.861	7.013	4.055	8.456	34.523	21.782
3	28.155	4.344	3.668	20.02	**2.037**	8.179	3.448	6.791	31.608	19.935
4	18.977	21.92	17.784	**14.062**	19.895	23.474	19.897	16.052	22.564	22.878
5	30.622	**3.662**	5.831	22.681	4.719	8.456	5.572	8.575	34.673	22.636
6	32.083	5.82	7.073	23.633	7.193	**4.529**	6.122	8.896	34.288	19.279
7	27.645	4.461	**2.81**	19.463	4.069	7.322	3.21	5.128	31.244	19.638
8	29.834	4.802	5.05	21.462	5.715	5.503	**3.843**	6.398	32.43	19.111
9	21.453	24.852	21.421	**16.772**	23.543	24.48	22.904	20.53	19.795	18.542
10	**14.638**	37.904	33.996	17.678	35.231	40.091	36.035	32.744	15.064	30.864

Table 35: Diagonal matrix for child disguise condition (LPCC) for LPI of /i:/.

	1	2	3	4	5	6	7	8	9	10
1	6.382	8.022	8.304	7.319	9.361	8.164	**5.563**	6.624	11.094	6.073
2	6.464	**3.4**	9.894	8.489	4.359	7.662	7.671	7.447	13.016	8.194
3	5.26	9.112	**4.008**	5.272	9.08	6.732	6.583	7.284	7.533	6.988
4	**4.4**	6.944	5.897	4.423	6.628	5.298	6.19	5.748	9.004	6.856
5	6.396	5.182	9.139	7.934	**2.791**	6.894	8.786	6.986	12.527	9.418
6	5.436	8.473	5.341	4.757	8.898	**4.335**	6.262	5.679	6.648	6.424
7	5.489	7.216	6.544	5.996	8.88	6.793	**4.04**	6.98	9.104	4.326
8	5.362	7.19	7.886	6.981	7.107	6.191	7.517	**5.308**	10.968	7.946
9	8.998	13.38	5.917	7.328	13.868	8.641	9.008	9.848	**3.861**	8.905
10	4.738	7.123	5.665	5.111	8.04	6.635	3.777	6.63	9.176	**3.66**

Table 36: Diagonal matrix for child disguise condition (MFCC) for HPI of /u:/.

MFCC	1	2	3	4	5	6	7	8	9	10
1	8.103	6.078	6.809	6.958	7.941	7.528	**3.826**	6.353	10.029	6.295
2	6.629	**4.749**	8.397	7.118	5.239	6.295	7.314	8.145	11.619	6.751
3	**4.637**	10.004	5.856	6.3	8.541	5.705	8.137	7.529	6.333	5.616
4	4.557	6.439	5.301	4.118	4.223	**3.58**	7.015	5.878	8.713	4.856
5	7.142	4.86	8.268	7.279	**2.906**	5.4	8.338	7.834	11.746	7.158
6	7.747	11.127	5.79	5.68	9.682	6.661	7.313	5.67	**4.78**	6.344
7	5.135	7.754	4.921	5.38	7.896	6.442	4.932	6.658	7.819	**4.286**
8	4.665	5.773	6.231	5.238	4.798	**3.098**	6.957	5.658	8.838	5.493
9	9.341	15.469	9.416	10.036	14.117	10.592	11.817	10.967	**5.802**	10.034
10	4.888	8.258	4.523	4.501	8.522	6.749	3.58	6.225	7.15	**2.858**

Table 37: Diagonal matrix for child disguise condition (MFCC) for LPI of /u:/.

	1	2	3	4	5	6	7	8	9	10
1	**6.995**	10.878	11.364	9.747	9.826	12.052	9.325	8.1	14.647	13.306
2	5.79	**4.386**	7.729	6.192	5.813	5.7	5.804	7.339	16.572	12.05
3	13.003	14.859	**8.221**	10.935	15.19	10.782	11.611	12.606	14.775	9.138
4	8.55	8.239	6.454	**5.938**	8.373	5.96	5.959	7.517	15.763	10.117
5	11.065	**4.434**	12.364	10.963	5.436	9.806	10.742	11.072	22.073	17.34
6	10.577	13.197	**7.403**	9.603	13.656	8.539	9.169	9.309	11.713	8.316
7	14.038	15.596	**9.405**	11.527	15.899	10.315	11.3	12.155	14.613	10.621
8	**4.733**	7.467	9.704	7.183	7.256	8.839	6.098	6.357	15.482	12.808
9	21.232	25.746	18.188	20.833	26.895	19.992	20.95	22.266	**12.314**	14.921
10	9.795	13.731	7.74	8.31	12.998	10.74	8.596	8.369	13.19	**7.091**

Table 38: Diagonal matrix for child disguise condition (LPCC) for HPI of /u:/.

	1	2	3	4	5	6	7	8	9	10
1	8.086	9.412	11.333	10.256	8.83	11.427	**7.671**	7.878	14.986	13.398
2	5.534	4.768	8.689	8.458	5.217	**4.277**	7.823	7.5	17.453	14.282
3	12.657	15.637	**7.985**	10.48	14.792	14.004	12.223	12.868	15.598	10.344
4	7.896	8.782	**6.61**	6.803	8.254	7.038	8.424	7.877	16.026	12.051
5	11.316	**5.639**	12.882	13.041	6.869	6.701	12.845	11.533	22.349	19.536
6	9.921	13.228	**6.805**	8.117	13.016	11.387	9.431	9.354	11.628	9.225
7	13.249	15.983	**8.841**	9.914	15.652	13.343	12.59	12.242	14.643	11.606
8	**4.809**	5.942	10.275	8.588	6.061	7.369	6.343	6.135	15.948	14.099
9	20.495	26.007	18.216	19.562	25.774	23.492	20.72	21.964	15.553	**14.446**
10	9.392	13.589	7.104	**6.694**	12.225	13.117	7.371	8.493	13.204	6.99

Table 39: Diagonal matrix for child disguise condition (LPCC) for LPI of /u: /.

		/a:/	/i:/	/u:/
MFCC	Highest PCI	90%	100%	80%
	Lowest PCI	30%	90%	40%
LPCC	Highest PCI	60%	80%	60%
	Lowest PCI	60%	30%	10%

Table 40: Percent correct identification for child disguise condition.

4) **Comparison of MFCC and LPCC in normal and female disguise condition. (Male voice Vs Female disguise):** Percent correct identification for vowels /a:/, /i:/, and /u:/ was 33.33 %, 33.33%, and 66.67%, respectively for MFCC and for LPCC it was 33.33%, 33.33%, and 0% Tables 41 and 42 show percent correct identification on MFCC and LPCC when the number of subjects was 3.

	Normal Vs. Female disguise								
	/a:/			/i:/			/u:/		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
S1	9.96	**8.49**	7.77	9.34	9.00	**8.23**	11.96	13.71	**11.76**
S2	10.39	9.45	**7.26**	13.60	**9.83**	13.47	**7.48**	9.36	12.64
S3	12.00	9.88	**7.58**	11.90	**10.38**	12.45	**11.10**	11.87	13.71

Table 41: Diagonal Matrix of Normal's Vs Female Disguise condition for three vowels on MFCC.

	Normal Vs. Female disguise								
	/a:/			/i:/			/u:/		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
S1	0.10	0.07	**0.06**	0.09	0.08	**0.07**	0.12	0.14	**0.08**
S2	0.10	0.10	**0.06**	0.14	**0.12**	0.15	**0.08**	0.10	0.13
S3	0.11	0.07	**0.05**	0.12	**0.10**	0.14	**0.11**	0.11	0.13

Table 42: Diagonal Matrix of Normal's Vs Female Disguise condition for three vowels on LPCC.

5) **Comparison of MFCC and LPCC in normal and child disguise condition. (Male voice Vs Child disguise):** Percent correct identification was 33.33%, 66.67%, and 66.67% for vowels /a:/, /i:/, and /u:/, respectively on MFCC and it was 0%, 66.67%, and 33.33% in LPCC. Tables 43 and 44 show percent

correct identification on MFCC and LPCC when the number of subjects was 3.

	Normal Vs. Child disguise								
	/a:/			**/i:/**			**/u:/**		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
S1	17.77	12.87	**10.05**	16.92	17.32	16.91	**17.53**	20.14	19.49
S2	17.64	16.09	**13.27**	**14.68**	19.22	18.07	14.57	**14.39**	**20.95**
S3	20.49	15.31	**13.23**	17.84	18.13	**14.70**	18.47	**17.34**	19.44

Table 43: Diagonal Matrix of Normal's Vs Child Disguise condition for three vowels on MFCC.

	Normal Vs. Child disguise								
	/a:/			**/i:/**			**/u:/**		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
S1	0.14	0.07	**0.07**	**0.13**	0.17	0.16	0.14	0.14	**0.14**
S2	0.15	0.11	**0.08**	0.19	0.15	**0.14**	**0.11**	0.13	0.19
S3	0.14	0.09	**0.08**	0.32	0.17	**0.15**	0.14	**0.14**	0.14

Table 44: Diagonal Matrix of Normal's Vs Child Disguise condition for three vowels on LPCC.

To summarise, benchmarking was good when normal condition was compared with normal, child disguise with child disguise, and male disguise with male disguise. However, benchmarking was very poor when normal condition was compared with child disguise or female disguise.

Chapter V

Discussion

The results of this study by using electronic vocal disguises showed several points of interest. ***First of all, the percent correct identification was above chance level for electronic vocal disguise on MFCC and LPCC.*** The results of the study were not in consonance with the earlier studies by several authors (Glenn & Kleiner, 1968, Reich; Moll and Curtis 1976; Reich; Duke 1979), where human generated disguises had very poor identification. However, with the present study the electronic vocal disguise was used.

Second, MFCC had higher percent correct identification (HPI) under all conditions compared to LPCC. In this study three conditions were used normal, female electronic vocal disguise and child electronic vocal disguise. Highest percent correct identification was 100% for /i:/, /u:/, and /i:/ for normal, female disguise and child disguise conditions, respectively. It appears that high vowels may be better for speaker identification.

Thirdly, percent correct identification was below chance level when the normal speaking condition was compared with disguise condition. This indicated that if one used disguise voice during a phone call and normal voice during direct recording. Speaker identification is below chance level using MFCC and LPCC. Benchmarking for Telugu vowels was very poor under electronic disguise conditions.

Conditions	**MFCC**			**LPCC**		
	/a:/	/i:/	/u:/	/a:/	/i:/	/u:/
Normal Vs Normal	90	100	60	80	80	60
Female Vs Female	80	80	100	50	70	70
Child Vs Child	90	100	80	60	80	60
Normal Vs Female	33.33	33.33	66.67	33.33	33.33	0
Normal Vs Child	33.33	66.67	66.67	0	66.67	33.33

Table 45: Percent correct identification scores.

The present study has contributed to the field of speaker identification. ***The results of the present study indicate that the extraction of Mel frequency cepstral coefficient and linear predictive cepstral coeffients is useful in speaker identification for long vowels /a: /, /i: / and /u: / in text-independent condition in males under electronic vocal disguise.*** However, comparison of normal and disguise voices may not be possible owing to very poor benchmark.

The present study was restricted to Telugu, male subjects, specific network and the results cannot be generalized to other languages, words and other network connections. The results of the study are quite interesting and future research is warranted on other networks and handsets, other Indian languages, other available disguise conditions.

Chapter VI

Summary and Conclusions

Forensic Speaker Identification is seeking an expert opinion in the legal process as to whether two or more speech samples are of the same person. According to Rose (1992) Speaker recognition can be speaker identification and speaker verification. Speaker identification is deciding if a speaker belongs to a group of known speaker population. Speaker verification is verifying the identity claim of the speaker. Speaker recognition methods can be divided into *text-independent* and *text-dependent* methods. In a text-independent system, speaker models capture characteristics of somebody's speech which show up *irrespective of what one is saying*. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her *speaking one or more specific phrases* (Rabiner, 1993). Every technology of speaker recognition, identification, and verification, whether text-independent or text-dependent, has its own advantages and disadvantages and may require different treatments and techniques. The choice of which technology to use is application-specific. At the highest level, all speaker recognition systems contain two main modules - *feature extraction* and *feature matching.*

Individual`s identity verification is an essential requirement for controlling access to protected resources. Personal identity is usually claimed by presenting a unique personal possession such as a key, a badge, or a password. However, these can be lost or stolen. Further a simple identity claim is not sufficient if the potential for loss is great and the penalty for false identification is severe. Hence verification of that claimed identity is necessary. This can be attempted by examining an individual's ***biometric features***, such as finger prints, hand geometry, or retinal pattern, or by examining certain features derived from the individual's unique activity such as ***speech*** or hand writing. In each case, the features are compared with previously stored features for the person whose identity is being claimed. If this comparison is favourable based on decision criterion, then the claimed identity is verified. In the present era of widely used telephone, mobile phone, radio, and tape recorder communication, the only information available to investigators may consist of a single voice recording, generally made during a telephone/mobile phone conversation. Among these methods, identity verification based on a person's voice has special advantages for practical deployment. Speech is our most natural means of communication and therefore user acceptance of the system would be very high. Apart from access, speaker identification is also used in forensic cases. Therefore, there is a pressing need on the part of police and magistrates for establishment of legal proof of identity from measurements of the voice. In view of this, the considerable interest in obtaining reliable techniques for speaker identification and in using these as the basis for such proof is easily understood.

Speaker Identification may be requested for a number of different criminal offences, such as making genuine or hoax emergency service calls to the police, ambulance or fire brigade, making threatening or harassing telephone calls, blackmail or extortion demands, taking part in criminal conspiracies such as those involving the importation, trafficking or manufacture of illegal drugs, or conspiring to traffic in people, arms, currency, and cultural artifacts. Speaker Identification may also be required in civil cases or for the media. These cases include calls to radio stations, local or other government authorities, insurance companies, or recorded conversations, rallies or meetings.

In speaker identification, the speech sample in question and control may suffer from the problems of noisy and poor quality recordings, vocal disguise, non contemporary, different text, different language and also electronic scrambling such as Voice synthesizers, Text to Speech converter and the almost limitless potential applications of speech processing in modern communication systems and networking such as Voice Over Internet Protocol (VOIP).

There are three methods of speaker identification (Hecker, 1971) - (a) speaker identification by listening (subjective method), (b) speaker identification by visual examination (subjective method), and (c) speaker identification by machine (objective method). In the first method, the expert hears the voices and decides whether two voices belong to the same person or not. In the second method spectrograms of two speech samples are visually matched to identify a speaker. The third method can be semiautomatic or automatic.

In automatic speaker identification, formant frequencies, fundamental frequency, F0 (Fundamental frequency) contour, Linear Prediction coefficients*,* Cepstral Coefficients, and Mel Frequency Cepstral coefficients were used in the past. Of these, Cepstral Coefficients and the Mel Frequency Cepstral Coefficients have been found to be more effective in speaker identification compared to all other features. Hence, the present study will be focusing on usefulness of linear cepstral coefficients and Mel frequency cepstral coefficients (MFCC) on speaker recognition. The aim of the study was to establish ***Benchmark for speaker identification under electronic vocal disguise using Mel frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients. (LPCC)*** The objectives of the study were to provide benchmarks for (a) Mel-frequency cepstral coefficients for Telugu* vowels in electronic vocal disguise condition, and (b) linear prediction cepstral coefficients for Telugu vowels in electronic vocal disguise condition.

Ten Telugu speaking normal male subjects in the age range of 20-35 years participated in this study. Commonly occurring forensically related Telugu meaningful words with long vowel /a:/, /i:/, and /u:/ embedded in five sentences formed the material. Subjects were informed about the study and their written consent was obtained. The written material was provided to the subjects and they are familiarized with the sentences. In the study 2 mobiles phones were used - Micromax X250 and Nokia 7210 supernova. Micromax X250 has a special option called "**MAGIC VOICE**" in which several disguise conditions exist. In the present study only 2 electronic disguises - children voice and man voice - were considered. Female subjects were instructed to speak the sentences in normal, children voice, and man voice. Subjects had the mobile MicromaxX250 and the experimenter had the NOKIA 7210 Supernova mobile. Subjects were instructed to switch on to the mode MAGIC VOICE in the mobile and to make a call to other mobile model Nokia7210 supernova and to speak the sentences thrice in normal and 2 disguise conditions one after another by taking a 30 sec interval between calls. The speech samples were recorded and stored in the microchip of the Nokia7210 supernova mobile.

The recorded samples were transferred on to the computer memory. All the samples were converted from .amr file extension to .wav file and from stereo to mono using you-tube downloader, Adobe Audition and Praat software (Boersma & Weenink 2009). Converted samples were stored in separate folders for each subject and for each disguise condition in the system.

SSL Pro.V4 Software (Voice and Speech Systems, Bangalore, India) and SSL WORK BENCH (Voice and Speech Systems, Bangalore, India) were used for analysis. The files were opened in Wavekep of SSL Pro V4. Second repetition of each recording was considered as test set and the third repetition was considered as training set. The second repetition of a word of the one subject was displayed as first file and the third repetition of the same word of same/different subject/ disguise was considered as the second file. The cursor was positioned at the zero cross of steady state of the vowel in both the files. Cepstrum and Mel frequency cepstrum/ linear prediction of cepstral coefficients were extracted at the point of the cursor. Using the option 'Show Euclidian distance', Euclidian distance between linear Cepstral coefficients and between 13 MFCCs of the first and the second files was extracted and noted down. The same procedure was continued for 5 zero crossings of steady state of the vowel/s.

The unknown subject (second repetition) was labeled as US1, US2, US3, US4, US5, US6, US7, US8, US9, and US10, and the known subjects (third repetition) were labeled as KS1, KS2, KS3, KS4, KS5, KS6, KS7, KS8, KS9, and KS10. The Euclidian distance within and between subjects were noted down and the subjects having the least Euclidian distance was considered as same. If the distance between the unknown and corresponding known speaker was less than any other known speaker the identification was deemed to have been correct. If the distance between the unknown and the corresponding known speaker was more than any other known speaker then the speaker was deemed to be falsely identified. The percent correct identification was calculated using the following formula:

$$\text{Percent correct identification} = \frac{\text{Number of correct identification}}{\text{Number of total identifications}} \times 100$$

The results of this study by using electronic vocal disguises showed several points of interest. ***First of all, the percent correct identification was above chance level for electronic vocal disguise on MFCC and LPCC.*** The results of the study were not in consonance with the earlier studies by several authors (Glenn & Kleiner, 1968, Reich; Moll and Curtis 1976; Reich; Duke 1979), where human generated disguises had very poor identification. However, with the present study the electronic vocal disguise was used.

Second, MFCC had higher percent correct identification (HPI) under all conditions compared to LPCC. In this study three conditions were used normal, female electronic vocal disguise and child electronic vocal disguise. Highest percent correct identification was 100 for /i:/, /u:/, and /i:/ for normal, female disguise and child disguise conditions, respectively. It appears that high vowels may be better for speaker identification.

Thirdly, percent correct identification was below chance level when the normal speaking condition was compared with disguise condition. This indicated that if one used disguise voice during a phone call and normal voice during direct recording. Speaker identification is below chance level using MFCC and LPCC. Benchmarking for Telugu vowels was very poor under electronic disguise conditions.

Conditions	**MFCC**			**LPCC**		
	/a:/	/i:/	/u:/	/a:/	/i:/	/u:/
Normal Vs Normal	90	100	60	80	80	60
Female Vs Female	80	80	100	50	70	70
Child Vs Child	90	100	80	60	80	60
Normal Vs Female	33.33	33.33	66.67	33.33	33.33	0
Normal Vs Child	33.33	66.67	66.67	0	66.67	33.33

Table 45: Percent correct identification scores.

The present study has contributed to the field of speaker identification. ***The results of the present study indicate that the extraction of Mel frequency cepstral coefficient and linear predictive cepstral coeffients is useful in speaker identification for long vowels /a: /, /i: / and /u: / in text-independent condition in males under electronic vocal disguise.*** However, comparison of normal and disguise voices may not be possible owing to very poor benchmark.

The present study was restricted to Telugu, male subjects, specific network and the results cannot be generalized to other languages, words and other network connections. The results of the study are quite interesting and future research is warranted on other networks and handsets, other Indian languages, other available disguise conditions.

References

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

Atal, B. S. (1974), Effectiveness of Linear prediction characteristics of the speech wave for Automatic Speaker Identification and Verification, *The Journal of the Acoustical Society of America*, *Vol.* 55, 1304-1312.

Atal, B.S. (1976), 'Automatic recognition of speakers from their voices'. *Proc.IEEE 64/4: 460-75.*

Bernasconi, C. (1990). On instantaneous and transitional spectral information for text dependent speaker verification. *Speech Communication, 9,* 129-139.

Bolt, R. H., Cooper, F.S., David, E.C., Denes, P.B., Picket, J.M. & Stevens, K.N. (1970). Speaker identification by speech spectrograms. *The Journal of the Acoustical Society of America,* 47, 597-613.

Bolt, R. H., Cooper, F. S., David, E. C. Denes, P. B., Picket, J. M. & Stevens, K. N. (1973),Speaker identification by speech spectrograms: some further observations. *The Journal of the Acoustical Society of America,* 54, 531-534.

Bricker, P. S & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America, 40,* 1441-1450.

Chandrika. (2010). the influence of handsets and cellular networks on the performance of a speaker verification system. *Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysore.*

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. In F. Nolan, 1983, (Ed), *The Phonetic Bases of Speaker recognition. Cambridge: Cambridge University Press*.

Coleman, R. O. (1973). Speaker Identification in the absence of inter- subject differences in glottal source characteristics. *The Journal of the Acoustical Society of America,* 53, 1741- 1743.

Doddington.G (1971), A method of speaker verification, *The Journal of the Acoustical Society of America*, *Vol*-49, p-139(A).

Doddington, G.R., Hyrick, B. and Beek, B. (1974). "Some results on speaker identification using amplitude spectra". *Journal of the Acoustical Society of America,* 55, 463(A).

Furui, S. (1978). Effects of long-term spectral variability on speaker recognition. *The Journal of the Acoustical Society of America,* 64,183.

Furui, S.(1994) 'An overview of speaker recognition technology', *Proc.ESCA workshop on Automatic Speaker Recognition Identification Verification:* 1-8.

Furui, S. (1981), Cepstral Analysis Technique for Automatic Speaker Verification, *IEEE Transactions on Acoustics, Speech and signal Processing,* Vol-29, 254-272.

Glenn. J.W & N. Kleiner (1968), Speaker identification based on nasal phonation, *The Journal of Acoustical Society of America.,* 43, 368-372.

Hasan, R., Jamil, M., Rabbani, G & Rahman, S. (2004). Speaker Identification using Mel frequency cepstral coefficients. *3rd International Conference on Electrical and Computer Engineering.*

Hazen, B. (1973). Effects of differing phonetic contexts on spectrographic speaker identification. *The Journal of the Acoustical Society of America,* 54: 650-660.

Hecker, M.H.L., Stevens, K.N., Bismarck, G. & William, C.E. (1968). "Manifestation of task-induced stress in the acoustic signal". *The Journal of the Acoustical Society of America, 49, 1842-1848.*

Hecker, M.H.L, (1971), Speaker Recognition: basic considerations and methodology, *The Journal of Acoustical Society of America.,* 49.

Hollien, H., & Majewski, W. (1977). Speaker identification by long term spectra under normal and distorted speech condition. *The Journal of Acoustical Society of America.,* 62: 975-980.

Hollien, H. (1974). The peculiar case of 'voiceprint'. *The Journal of the Acoustical Society of America,* 56, 210-213.

Hollien, H. (1990), the acoustics of Crime*, The New Science of Forensic Phonetics*, *Plenum, Nueva York.*

Hollien, H. (2002). *Forensic Voice Identification. San Diego, CA: Academic Press.*

Holmes, J., & Holmes, W. (2001) Speech synthesis and recognition *(2nd edition) New York: Taylor & Francis.*

Imperl, B., Kacic, Z & Horvat, B. (1997). A study of harmonic features for the speaker recognition. *Speech Communication*, 22, 385-402.

Ingemann, F. (1968). Identification of the speaker's sex from voiceless fricatives. *The* Journal *of the Acoustical Society of America,* 44, 1142- 1144.

Jakkhar, S.S. (2009), Bench mark for speaker Identification using Cepstrum, *Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore.*

Johnson, C., Holien .H & Hicks. J, (1984). Speaker identification utilizing selected temporal speech features, *The Journal of Phonetics, Vol. 12, 319-326.*

Kersta, L. G. (1962). Voice identification, *Nature,* 196, 1253-1257.

Kinnunen, T. (2003). Spectral features for automatic text-independent speaker recognition. *Unpublished thesis university of Joensuu, department of computer science. Finland.*

Koenig, B. E. (1986). Spectrographic voice identification: A forensic survey, (letter to the editor). *The Journal of the Acoustical Society of America, 79,* 2088-2090.

Kunzel, H.J. (1994) 'Current approaches to forensic speaker identification,' *Proc.ESCA Workshop on Automatic speaker Recognition identification Verification*: 135-141.

Lakshmi, P., & Savithri. S.R. (2009), Bench mark for speaker Identification using Vector F1 & F2, *Proceedings of the international symposium, Frontiers of Research on Speech & Music, FRSM-2009* , 38 - 41.

Lass, N.J. (Ed.) (1976) *Contemporary Issues in Experimental Phonetics, London: academic Press.*

Luck, J. E. (1969). Automatics speaker verification using cepstral measurements. *The Journal of the Acoustical Society of America,* 46, 1026-1032.

Mao, D., Cao, H., Murat, H., & Tong, Q. (2006). Speaker identification based on Mel frequency cepstrum coefficient and complexity measure. *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi,* 23 (4), 882- 886.

Markel, J. D., & Davis, S. B. (1979), Text independent Speaker Recognition from a Large Linguistically Unconstrained Time spaced Data Base, *IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-27,* 74-82.

McDermott, M. C., Owen, T. & McDermott, F. M. (1996).Voice identification: the aural spectrographic method. In P. Rose, 2002, (ed.), Forensic Speaker Identification. Taylor and Francis, London.

McGehee, F. (1937). The reliability of identification of human voices, *The Journal of General Psychology*, 17, 249- 271.

Medha. (2010). Benchmarking for speaker Identification by Cepstrum measurement using text-independent data. *Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysore.*

Meltzer, D & Lehiste*,* I. (1972). "Vowels and speaker identification in natural and synthetic speech". *The Journal of the Acoustical Society of America, 51, S131 (A).*

Naik, J. (1994). Speaker Verification over the telephone network: database, algorithms and performance, assessment, *Proc. ESCA Workshop Automatic Speaker Recognition Identification Verification,* 31- 38

Nolan, F. (1983). The phonetic Bases of Speaker Recognition, *Cambridge University press,* Cambridge.

Nolan, F. (1997), ‘Speaker recognition and forensic phonetics’, in Hardcastle and Laver (Eds.) (1997): 744-67.

Pamela, S. (2002). Reliability of voice print. *Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysore.*

Plumpe, M., Quatieri, T & Reynolds. D. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Tran. Speech and Audio Proc,* 7, 569-586.

Pollack. I., Pickett, J.M & Sumby W.H (1954). On the identification of speakers by voice. *The Journal of the Acoustical Society of America,* 26, 403-406.

Potter, R. K., Kopp & Green. (1946). Introduction to technical discussions of sound portrayal. *The Journal of the Acoustical Society of America,* 18, 1-3.

Pruzansky. S. (1963). Pattern-matching procedure for automatic talker recognition. *The Journal of the Acoustical Society of America* 35, 354-58.

Quatieri, T.F., Jankowski, C.R., Jr & Reynolds, D.A. (1994) Energy onset times of speaker identification Proc.IEEE*, volume :(1), 160-162.*

Rabiner, L., & Juang, B.H. (1993), Fundamentals of Speech Recognition, *Prentice Hall PTR.*

Ramya. B.M. (2011), Bench mark for speaker identification under electronic vocal disguise using Mel Frequency Cepstral Coefficients. *Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysore.*

Reich & Duke, (1979), Effects of selected vocal disguises upon speaker identification

by listening", *The Journal of the Acoustical Society of America* 66, 1023-1026.

Reich, A.R., Moll, K.L. & Curtis, J.F. (1976). “Effects of selected vocal disguises upon spectrographic speaker identification”. *The Journal of the Acoustical Society of America,* 60, 919-925.

Rose, P. (1990) ‘Thai Phake tones: acoustic aerodynamic and perceptual data on a Tai dialect with contrastive creak’, in R.Seidl (ed.) *Proc. 3^rd^ Australian Intl. Conf. on speech science and Technology: 394-9, Canberra: ASSTA.*

Rose, P. (2002). “Forensic Speaker Identification”. *Taylor and Francis, London*

Sreevidya, N. (2010). Speaker Identification using Cepstrum in Kannada language. *Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysore.*

Stevens, K. N. (1968), Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material, *The Journal of the Acoustical Society of America*, 44: 1596–1607.

Stevens, K. N. (1971), Sources of inter and intra speaker variability in the acoustic properties of speech sounds, *Proceedings 7th International Congress. Phonetic Science. Montreal,* 206-227.

Su. L.S., Li. K.P & Fu, K.S. (1974), Identification of speakers by the use of nasal coarticulation, *The Journal of the Acoustical Society of America* 56, 1876-1882.

Schwartz, M. F. & Rine, H. E. (1968). Identification of the speaker sex from isolated, whispered vowels. *The Journal of Acoustical Society of America,* 44, 1736-1137.

Schwartz, M. F. (1968). Identification of speaker sex from isolated voiceless fricatives, *The Journal of Acoustical Society of America,* 43, 1178- 1179.

Telugu (n.d) *In Wikipedia Online. Retrieved from* http://www.wikipedia.com.

Thompson, C. (1985). Voice Identification: Speaker Identifiability and correction of records regarding sex effects, *Hum. Learn.* 4, 19- 27.

Tiwari, R., Mehra, A., Kumawat, M., Ranjan, R., Pandey, B., Ranjan, S. & Shukla, A. (2010). Expert System for Speaker Identification Using Lip Features with PCA. Intelligent Systems and Applications (ISA), 2010 2nd International Workshop, 1-4.

Tosi, O., Oyer, H. J., Lashbrook, W., Pedrey, C., Nichol, J. & Nash, W. (1972). Experiment on voice identification. *The Journal of the Acoustical Society of America,* 51, 2030-2043.

Wang, L., Ohssuka, S., & Nakagawa, S. (2009). High improvement of speaker identification and speaker verification by combining MFCC & Phase information. *Unpublished study by the department of systems engineering, Schizuoka University, Japan.*

Wolf, J. J. (1972), efficient acoustic parameter for speaker recognition, *The Journal of the Acoustical Society of America* 2044–2056.

Young, M. A., & Campbell, R. A. (1967), Effects of Context on Talker Identification. *The Journal of the Acoustical Society of America*, 42, 1250 - 1260