

**BENCHMARK FOR SPEAKER IDENTIFICATION FOR NASAL  
CONTINUANTS IN URDU IN DIRECT AND MOBILE  
NETWORK RECORDING**

Ayesha Anjum  
**Register Number: 14SLP004**

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, 2016**

## CERTIFICATE

This is to certify that this dissertation entitled “**Benchmark for speaker identification for nasal continuants in Urdu in direct and mobile network recording**” is a bonafide work submitted in part fulfilment for degree of Master of Science (Speech-Language Pathology) of the student Registration Number: 14SLP004. 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, 2016

**Dr. S.R. Savithri**  
**Director**  
All India Institute of Speech and Hearing  
Manasagangothri, Mysore-570006

## CERTIFICATE

This is to certify that this dissertation entitled “**Benchmark for speaker identification for nasal continuants in Urdu in direct and mobile network recording**” has been prepared under my supervision and guidance. It is also been certified that this dissertation has not been submitted earlier to any other University for the award of any other Diploma or Degree.

Mysore  
May, 2016

**Guide**  
**Dr. S.R. Savithri**  
Director  
All India Institute of Speech and Hearing  
Manasagangothri, Mysore-570006

## DECLARATION

This is to certify that this dissertation entitled “**Benchmark for speaker identification for nasal continuants in Urdu in direct and mobile network recording**” is the result of my own study under the guidance of Dr. S.R. Savithri, Director, All India Institute of Speech and Hearing, Mysore, and has not been submitted earlier to any other University for the award of any other Diploma or Degree.

Mysore,  
May, 2016

**Registration No. 14SLP004**

## ACKNOWLEDGEMENTS

Praise be to ALLAH (SWT), his majesty for his uncountable blessings and for bestowing me with good health and patience to complete this dissertation.

Dr. Savithri S.R. Ma'am, you're a true inspiration. Your patience, motivation and immense knowledge is outstanding Ma'am. Thank you for providing me with an enriching learning experience and your insightful comments enabled me to widen my research from various perspectives Ma'am. Your calm, composed nature and quiet strength has greatly helped in the successful completion of this dissertation. You're an unparalleled mentor and I couldn't have imagined completing this dissertation without your guidance Ma'am.

Dr. Jayakumar T. Sir, Thank you for providing me with your valuable inputs and assistance in carrying out my data analysis. You have been a great support Sir. Your patient listening has helped me tremendously in giving shape to this dissertation Sir.

My sincere thanks to Dr. Sreedevi N and Dr. Yeshoda K. for permitting me to use the software in the speech science lab for my data analysis.

I would like to extend my gratitude to Dr. Prema K.S., Dr. Manjula R. and Dr. Shyamala K.C. for providing me with an incredible learning experience during my PG.

I would like to thank all the participants for consenting to this study, without whom this study would not have been possible.

Amit Kumar ji, thank you for being the motivation factor in my life. Your unconditional support and care is irreplaceable.

I would like to express my thanks to all my friends at AIISH for being by my side at all times and for giving me such wonderful memories for life.

I thank my lovely family back home for their unconditional love and support. I feel truly blessed to have such beautiful parents and an amazing family. I love you all very much.

## Table of contents

<b>Chapter No.</b>	<b>Title</b>	<b>Page No.</b>
	Table of contents	vi
	List of tables	vii
	List of figures	viii
1.	Introduction	1- 8
2.	Review of Literature	9-39
3.	Method	40-50
4.	Results	51-59
5.	Discussion	60-66
6.	Summary and Conclusions	67-71
7.	References	72-78

## List of Tables

<b>Sl. No.</b>	<b>Title</b>	<b>Page No.</b>
1.	Diagonal matrix – direct vs. direct recording speaker identification /m/	52
2.	Diagonal matrix – direct vs. direct recording speaker identification /n/	52
3.	Diagonal matrix – direct vs. direct recording speaker identification /n/	52
4.	Diagonal matrix – network vs. network recording speaker identification /m/	53
5.	Diagonal matrix – network vs. network recording speaker identification /n/	54
6.	Diagonal matrix – network vs. network recording speaker identification /n/	54
7.	Diagonal matrix (HPI) of direct vs. network recording for speaker identification /m/	55
8.	Diagonal matrix (HPI) of direct vs. network recording for speaker identification /n/	56
9.	Diagonal matrix (HPI) of direct vs. network recording for speaker identification /n/	56
10.	Diagonal matrix (LPI) of direct vs. network recording for speaker identification /m/	57
11.	Diagonal matrix (LPI) of direct vs. network recording for speaker identification /n/	57
12.	Diagonal matrix (LPI) of direct vs. network recording for speaker identification /n/	58
13.	Percent correct identification for all nasal continuants	58
14.	Summary of the percent correct speaker identification	59

## List of Figures

Sl. No.	Title	Page No.
1.	Illustration of speaker identification	11
2.	Illustration of the types of reference sets	12
3.	Types of errors in speaker identification	13
4.	Illustration of Mel filtering [Taken from Milner, 2003]	35
5.	Segmentation of samples for (a) /m/, (b) /n/ and (c)/n/	42
6.	Illustration of the notepad.	43
7.	Mel frequency filter bank without normalization.	44
8.	Mel frequency filter bank with normalization	44
9.	Notepad of SSL workbench.	44
10.	SSL Workbench window for analysis.	45
11.	Illustration of speaker number being selected for segmentation.	46
12.	Illustration of selecting the session number and occurrence number.	46
13.	Depiction of segmentation window showing one occurrence of /m/ for a speaker.	47
14.	Showing dialogue box asking for confirmation of the highlighted segment in the file.	47
15.	Analysis window of SSL Workbench.	48
16.	Analysis window of SSL Workbench showing diagonal matrix and the final speaker identification score.	49
17.	Percent identification under three conditions	59
18.	Spectrogram of the network recorded nasal /m/	63
19.	Spectrogram of the network recorded nasal /n/	63
20.	Spectrogram of the network recorded nasal /n/	64



## CHAPTER I

### INTRODUCTION

People are identified routinely by their voices in everyday life. People are recognized on a daily basis with their distinctive voices, over a radio, phone line, to name a few. Voice production is facilitated by a structure in the human body called the larynx. The larynx is a highly specialized structure which is responsible for the generation of acoustic signals and it has a vital role in the process of respiration. The structure comprises of a fold-like soft tissue which is referred to as the vocal folds or the vocal cords. The vocal folds form the main vibratory component of the larynx. Sound is generated when the vocal folds are excited by the aerodynamic forces from the lungs and thus setting them into vibration. The air that's passing through the vocal folds causes the vocal folds to vibrate rapidly in a sequence of vibratory cycles which produces the voice. Voice is also interchangeably used with speech which is produced by the modification of air at the level of vocal folds and the articulators.

The process in which this acoustic signal generated by the larynx is perceived and interpreted in the auditory system varies from one person to another. Therefore, the auditory system can be considered to be one of great precision as well as one which is quite deceptive in function (Hollien, 1974)

In our day-to-day life persons who are familiar to us can be easily identified by the quality of their voice, their style of speaking, and so on. A qualitative amount of information can also be inferred from unfamiliar people such as their age, gender, emotional state and language, among others (Jyotsna, 2011).

In order to gain access to high security areas an individual's identity verification is an essential requirement. This requirement is typically met by an exclusive personal possession such as a key, a badge, or a password. However, these can be lost or stolen (Jyotsna, 2011). If such an unanticipated situation befalls and if the penalty for false identification is severe then other verification methods of the claimed identity has to be adopted. Therefore, this can be attempted by examining an individual's biometric features, such as voice prints, finger prints, retinal pattern, or by analyzing certain features derived from the person's unique activity such as speech or hand writing. In any case, the features are compared with previously stored features for the person whose identity is being claimed. If this comparison is favorable, based on decision criterion, then the claimed identity is verified (Prasanna, 2011).

The voice of an individual can be recorded while planning, committing or confessing to a crime. It can be used to directly incriminate the suspect in the act of committing the crime (Rose, 2002).

“Forensic voice identification is a legal process to decide whether two or more recordings of speech are spoken by the same speaker” (Rose, 2002). A voice print is one of the means used to identify a person who has committed a crime and is valid as evidence in a court of law (Saitō & Nakata, 1985). Hence, there are essential practical applications of using a person's voice for identity verification. The most natural means to communicate with people is speech and thus the system's user acceptance would also be very high (Prasanna, 2011).

In the recent years the rate of crime has become greater than before especially after the advent of telecommunication devices such as mobile phone, tablets and portable personal computers. Consecutively the misuse of such devices to create social

menaces has also become more in the form of kidnapping, harassing, bomb threats, ransom demands and many more. In such circumstances the speaker's voice is the only source available for analysis. Perpetrators have a tendency to disguise their voice to avoid detention by concealing their identities. Vocal disguise refers to this deliberate action on the part of a speaker to conceal his/her identity. Out of the many possibilities available to an individual for vocal disguise, falsetto, whisper, change in speaking rate, imitation, pinched nostrils and object in the mouth are popular favorites of perpetrators (Ramya, 2013). Thus, this paved way for the advances in the field of Forensic Speaker Identification.

Speaker recognition is defined as any decision making process that uses the speaker dependent features of the speech signal (Hecker, 1971). Speaker recognition has a number of applications including computer access control (Naik and Doddington, 1987; Higgins, Bahler and Porter, 1991), telephone voice authentication for banking access, intelligent answering machines and law enforcement.

Rose and Reynolds (1990) state that speaker recognition can be in the form of speaker identification or speaker verification. *Speaker identification* is the identification of a particular speaker from a group of unknown speakers. It involves the application of a combination of auditory and acoustic methods which may finally point to the voice on a recording of a telephone conversation or live recording as to belonging to a particular known speaker. On the other hand, *speaker verification* refers to verifying if a particular voice sample of an individual belongs to them as claimed by them. It is also referred to as speaker authentication, talker authentication, voice verification, voice authentication and talker verification.

Speaker recognition can also be classified as *text-dependent and text-independent*. In the latter, voice characteristics are analyzed from the sample recording irrespective of the linguistic content of the recording. In the former, the identification is based on the speaker speaking a particular phrase like a password, pin code etc. (Rabiner and Juang, 1993). However, every technique has to be evaluated for its advantages and disadvantages and then considered. The decision to use the text-dependent or text-independent depends on the application considered for the analysis. Primarily, all modules contain two processes, feature extraction and feature matching.

There are two major problems that are most frequently faced by the forensic analyst. These are the; system distortions and speaker distortions (Rida, 2014). **System distortion** is the result of limited fundamental frequency response like a telephone conversation, noise like wind, fan, clothing friction or automobiles in the background which may obscure the speaker characteristics and make identification a more tedious task, and interruptions. The microphones with limited capability or poor quality tape recorders also can result in the loss of speaker characteristics which may be irrecoverable later. **Speaker distortions** include having cold, under the influence of drugs, alcohol which can change the way a voice sounds in a recording. Some may even try to disguise their voice (Hollien and Rosenberg, 1991).

In speaker identification, the speech sample in question and control may suffer from the problems of noisy and poor quality recordings, vocal disguise, different text, different language and also electronic scrambling such as Voice synthesizers, and Text-to-Speech converters (Ramya, 2013).

Researches in the past have used formant frequencies, fundamental frequencies, F0 contour, Linear Prediction Coefficients (Atal, 1974; Imperl, Kačič & Hovert, 1997),

Cepstral Coefficients (Jakhar, 2009; Medha, 2010; Sreevidya, 2010) and Mel Frequency Cepstral Coefficients (Plumpe, Quatieri & Reynolds, 1999; Hasan, Jamil, Rabbani & Rahman, 2004; Chandrika, 2010; Mehra et. al., 2010; Ramya, 2013; Rida, 2014) to identify speakers. The Cepstral Coefficients and Mel Frequency Coefficients have found to be the most accurate predictors of speaker identification.

Atal (1974) carried out a study for automatic recognition of speakers from their voice by examining several parameters using linear prediction model. *Cepstrum* was found to be the most effective parameter, with an identification accuracy of 70% for speech of 50 ms in duration, which increased to more than 98% for duration of 0.5s. The same speech data was used to find that the verification accuracy was approximately 83% for duration of 50 ms increasing to 95% for duration of 1sec.

Hasan, Jamil, Rabbani and Rahman (2004) used *Mel Frequency Cepstral Coefficients* (MFCC) for feature extraction and vector quantization in security system based on speaker identification. Twenty one speakers consisting of 13 male and 8 female subjects were included in the study. Results revealed 57.14% speaker identification for code book size of 1 and 100% speaker identification for code book size of 16. MFCC technique has been identified as the most efficient method for speaker identification.

Glenn and Kleiner (1968) used vectors from nasal phonation to carry out an automatic speaker identification study. Nasal phonation was chosen in the study because nasal sounds had a relatively fixed position in the oral tract and the open nasal tract generated steady-state power spectrum. In the experiment the accuracy obtained was 93% for 30 speakers and 97% for 10 speakers. Therefore, the results supported the hypothesis that nasal phonation was a strong clue to speaker identity.

The nasal disguise was the most effective disguise in speaker identification by listening experiment (Reich, Moll, & Curtis, 1976). In contrast, the nasal disguise was the least effective in a previous spectrographic matching experiment (Reich & Duke, 1979). Similarly, the power spectra of nasal consonants (Glenn & Kleiner, 1968) and coarticulated nasal spectra seem to provide strong cues for the machine matching of speakers.

The studies mentioned above strongly provide evidence to support the extraction of MFCCs using nasal continuants over other parameters for experiments in speaker identification. Further, review of most of the studies on (Reich & Duke, 1979; Reich, Moll, & Curtis, 1976; Rida, 2014; Nithya, 2015) effective disguise for speaker identification state nasal disguise and slow rate of speech are the *least effective disguises*. Therefore, nasal continuants would be the best speech sounds to investigate speaker identification under disguise.

India is a multilingual country and the phoneme system of languages differs from one another. *Urdu* is a standardised register of the Hindustani language. It is also one of the 22 official languages recognized in the Constitution of India. It is the official language spoken in six states of India. This language is historically viewed to be associated with the Muslim community. Urdu is also spoken in many parts of the world including, Afghanistan, Bahrain, Bangladesh, Mauritius, Nepal and the United Arab Emirates to name a few.

Urdu is closely related to Hindi. A lot of Urdu vocabulary is derived from Persian and Arabic, while Hindi comprises of vocabulary from Sanskrit. Linguists consider Standard forms of both Urdu and Hindi to be different formal registers derived from the Khari Boli dialect, which is also known as Hindustani. Urdu and Hindi have a few

significant differences at the informal spoken level but both are mutually intelligible to its speakers. Hindi and Urdu differ in the pronunciation of vowels and consonants. Nasalization and aspiration are present on certain vowels and consonants in Urdu. The effect of co-articulation in Urdu is also unique from that that of Hindi. The lexicon in Urdu is quite distinct from Hindi (Delacy, 1998).

There are limited studies in the field of Forensic Speaker Identification to train experts on analysis. In order to provide adequate training to experts in this field to make them efficiently identify voices, it is important to have such studies. Scientific testimony impresses any court of law in whichever country that might be. However for any result to be called scientific, it has to be measured, quantified and reproducible if and when the need arises. Therefore, a method to carry out these analyses becomes imperative.

Further, the phonemes differ from one language to another. Shah (2002) states that there are 3 nasal consonants of Urdu are /m/ [bilabial, voiced, plosive], /n/ [alveolar, voiced, plosive], and /ŋ/ [velar, voiced, plosive]. These nasal continuants may be similar to those in Hindi. They also state that /n/ followed by the bilabial stops, gets labialized; /n/ followed by the dental stops, becomes dental; /n/ followed by the alveolar affricates, [dʒ, tʃ], becomes alveolar; /n/ followed by the retroflex stops, becomes retroflex, /n/ followed by the velar stops, becomes velar. /n/ do not change place for following continuants; /n/ becomes the bilabial nasal /m/, when it gets labialized; /n/ becomes velar nasal /ŋ/, when it is followed by /g/. The nasal continuants of Urdu may be different from that of its closest language Hindi. In this context, the present study aimed at establishing benchmark for speaker identification in Urdu nasal continuants using Mel Frequency Cepstral Coefficients (MFCC).

Specifically, the objectives of the study were to provide benchmark for speaker identification for Urdu nasal continuants using, and compare benchmarks in direct and network recording conditions.



## **CHAPTER II**

### **REVIEW OF LITERATURE**

The review is dealt under the following headings:

- (1) Speaker identification
- (2) Methods of speaker identification

#### **(1) Speaker identification**

The investigations into identifying a person based on his or her voice has been debated. Identifying a voice using forensic- quality samples is generally a challenging task for automatic, semiautomatic, and human based methods. As the speech samples that are available for analysis would have been recorded originally in different situations, they may be contaminated by noise or may not contain relevant speech data to provide conclusive evidence. Therefore, all these variables make the speaker recognition process a daunting task.

Speaker recognition has a history dating back to the World War II. During the World War II the first attempt towards speaker recognition was conducted in relation to the assassination of Adolf Hitler on July 21, 1944 at Wolf's Lair. There was no convincing evidence as to whether Hitler had escaped or was killed. At that time, a number of speeches given by Hitler were recorded and stored. A group of phoneticians and engineers were appointed to compare the recent and the old recordings of his speech. Detailed analysis was carried by the investigators which finally revealed that Hitler was still alive (Hollien, 2002).

In a case involving the kidnap of an 11-year-old German girl in 1987, forensic analysis yielded significant findings that the kidnapper's voice and the suspect's

voice had significant similarities (Kunzel, 1987).

In yet another similar case in the United States of America, which involved a telephone bomb threat, the suspect had been acquitted yielding to the evidence from the forensic analysis of the offender's and the suspect's voice (Hollien and Rosenberg, 1991).

An air freight cargo handler, Paul Prinzivalli in Los Angeles, faced trial in the court of law for having threatened his employer, Pam Am. The offender's voice sample was compared to Paul Prinzivalli's voice by conducting forensic-phonetic analysis of the samples. The analysis proved to be a breakthrough in the case wherein it revealed that the dialect of the offender belonged to a typical New England speaker's accent whereas the suspect had an accent from New York. Thus, the evidences established Prinzivalli innocence in the court of law (Labov and Harris, 1994).

In 1999, the crime squad captured illegal drug traffickers in Australia, by carrying out forensic voice analysis of 15 telephonic conversations that was exchanged by the criminals (Lam, 1999).

Thus, the field of forensic voice analysis has been of tremendous use in dealing with law offenders not only in the past but also till date.

The voice identification technique was first adopted by Michigan State Police in 1966 and in the American court in the mid 1960's. This technique was adopted across majority of the states in America (McDermott and Owen, 1996).

Hecker (1971) broadly categorized speaker recognition into two specific tasks - Speaker identification, and Speaker verification.

Speaker identification is defined as “to identify an unknown voice as one or none of a set of known voices” (Naik, 1994). The task in speaker identification 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 speakers (Nolan, 1983).

In Figure 1 below, there is a schematic representation of simple speaker identification. The speaker identification experiment is represented with a reference set of 50 known speaker samples. In the Figure 2, the unknown sample on the left side is compared with the known sample 1 (A) on the right side, with the known sample 2 (B), and so on. The question mark represents the question whether, “the unknown speaker sample matches with the known speaker sample?” If it matches any one of the known speaker sample, say sample 2, then, the result shows that the sample has been identified as speaker B.

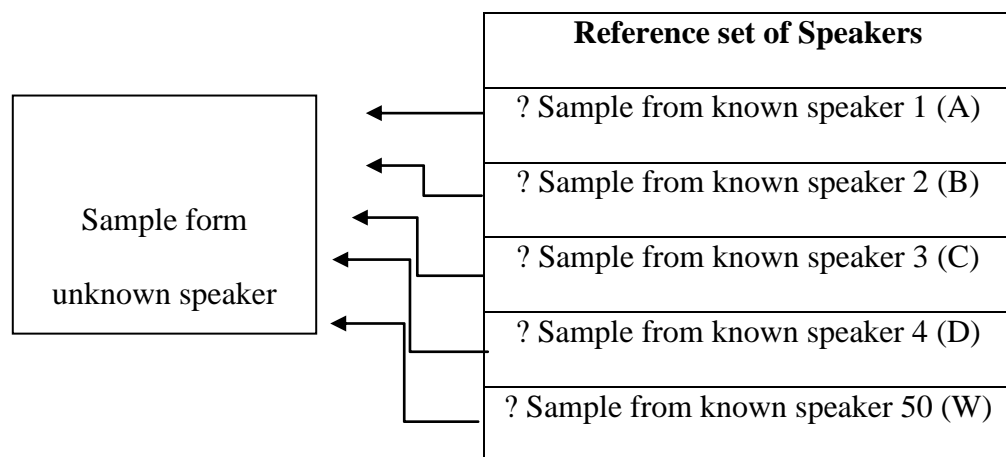


Figure 1: Illustration of speaker identification

In speaker identification, there are possibilities of two types of reference sets (figure 2) of the known speakers.

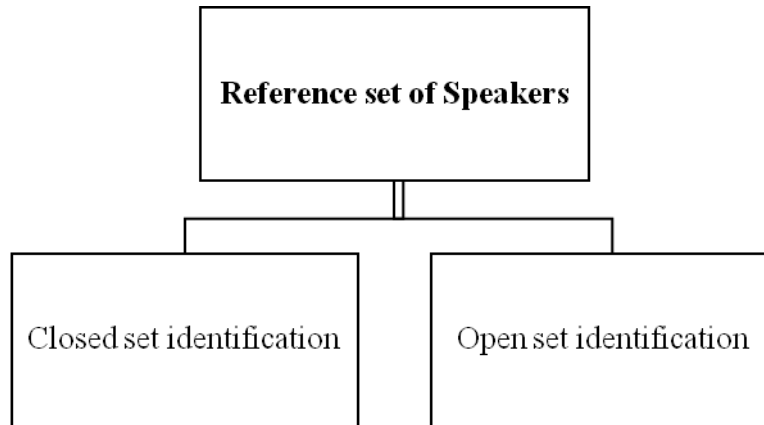


Figure 2: Illustration of the types of reference sets

A closed reference set is when it is known that the unknown speaker is one among the known speaker. An open reference set is when it is not known whether the known speaker is among the known speaker. Closed set speaker identification is an easier process than open set identification, as the possibilities of occurrence of error identification is less in closed set identification. Thus, the closed set identification task includes estimating the distance between the samples of the unknown speaker and each of the known reference speakers, and identifying the known speaker using the sample that is separated by the least distance from the unknown speaker.

The pair of sample separated by the smallest distance is assumed to be from the same speaker (Nolan, 1983). In speaker identification there is no threshold for establishment, since it automatically selects the unknown speaker form the samples given by selecting the one with the least distance from the test sample.

In speaker identification, there are possibilities of only two responses by the

examiner, either the unknown speaker is among the known speaker or it is not. In the open set speaker identification task three types of errors can occur. Figure 3 is a schematic representation of the classification of the three types of errors.

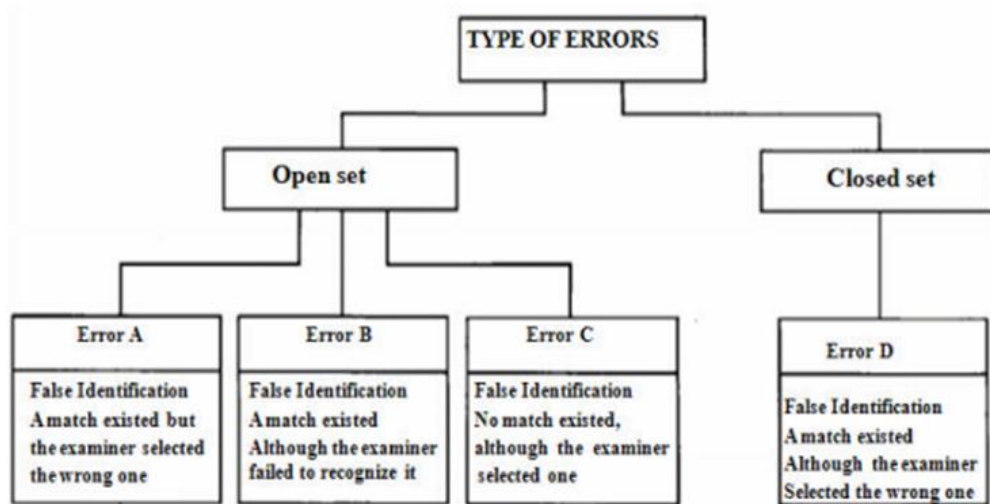


Figure 3: Types of errors in speaker identification

The examiner faces multiple difficulties while carrying out the speaker identification task. The following are some of them:

**Uniqueness:** The identification may involve an open set of trials. In such a task, the known speaker must be detected from a large population of ‘possibilities’.

**Distortion:** Identification of a speaker in the presence of noise causes two main types of distortions - System distortions and Speaker distortions. The system distortions is where the signal may get degraded in terms of frequency response, introduction of noise component and any other sort of frequency or harmonic distortions in the sample. Speaker distortions may arise from sources such as fear or anxiety or stressful conditions that occur when the perpetrator is speaking during the commission of the crime. Factors like drugs or alcohol ingestion and even temporary health states such as a cold can affect the quality of the speech signal.

## **(2) Speaker identification methods**

Speaker identification can be accomplished by listening (subjective method), by visual examination of spectrograms (objective method), and by machine (objective method).

Forensic speaker recognition can be a daunting task for any investigator due to the number of challenges. There are a number as methods that are available for an investigator to choose from to analyse the data including automatic, semi-automatic and human based methods. The data may be contaminated as the quality of recording may vary depending on the amount of background noise, the disguise adopted by the speaker and the influence of general health and drugs on the speaker's voice among a few to mention. Apart from the speaker's variability the system itself has certain amount of distortions that may contribute to the quality of the signal captured. Hence, an investigator has to be cautious while selecting the method for analysis keeping all these variables in cognizance.

### **Speaker identification by listening (subjective method)**

One of the first studies in the field of aural speaker identification was done by McGehee (1937). In this study listeners had to carry out a task of selecting a single target voice from a set of five male voices after delays that ranged from 1 day to 5 months. The percentage of correct identification declined from 83% after day 1 to 80.8% after 1 week, 68.5% after 2 weeks, 57% after 1 month, and to 13% after 5 months. Therefore, this study highlights the inconsistency seen in speaker identification when a subjective method is in use. On the similar lines, Bricker and Pruzansky (1966) conducted a study which also pointed towards the

variability of the results when this method is adopted.

Pollack, Pickett and Sumbly (1954) conducted experiments with speech sample duration which suggested that identification accuracy improves with increasing duration up to about 1200ms; for longer periods, accuracy did not appear to be related to duration, but rather to phonemic repertoire. The authors also found that when they degraded the speech signal by increasing the number of speakers or by substituting whispered speech for normally spoken speech, listeners needed longer samples to identify known talkers.

Bolt et al, (1970) carried out a study to examine speaker authentication and speaker identification. The author used two different methods of presentation of speech material: (1) speech samples were presented aurally through headphones, and (2) speech samples were presented visually as conventional intensity-frequency-time patterns, or spectrograms. Two types of experiments were carried out: (1) 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 (2) 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 the speakers based on utterances of single words or phrases is more accurate than identification from the spectrograms and average error rate obtained by listening is 6% than visual 21% for the closed set identification. For the open visual tests, appreciable numbers of false acceptances (incorrect authentication) were made. The results of the study suggest that for practical situations it would be best if procedures with minimal risk for error are adopted in future.

Schwartz and Rine (1968) conducted a study and determined that listeners could identify the gender of the speakers from the isolated production of /s/ and /ʃ/, but could not from /f/ and /θ/ production. Spectrographic analysis of /s/ and /ʃ/ stimuli revealed that the female spectra tended generally to be higher and parallel in frequency compared to that in the males.

Ingemann (1968) conducted a similar experiment and he supported the findings of Schwartz and Rine (1968) that listeners are able to identify the gender of the speaker from perceiving voiceless fricatives in isolation and that gender was identified better on the fricative /h/.

Schwartz and Ingemann (1968) presented isolated voiceless fricatives as auditory stimuli and they determined that listeners could accurately identify speaker gender from these stimuli, particularly from /h/, /s/, and /ʃ/. The authors inferred that the laryngeal fundamental was not available to the listeners because of the voiceless nature of the consonants. This indicated that accurate speaker gender identification can be done from the vocal tract resonance information alone.

Hollien, Majewski and Doherty (1982) reported high levels of correct identification of known speakers under normal conditions, stressful speaking conditions to be 98% and 97% respectively, but lowered accuracy of 79% for disguise conditions. The authors also found that listeners could identify a particular unfamiliar voice at only about 40% accuracy, and when stress and disguise were added to the paradigm, the accuracy fell to 31% and 21%, respectively. They noted that the values were even lower when they studied



listeners who were unfamiliar with both the speakers and the language spoken.

Hollien and Schwartz (2000, 2001) carried out a study in speaker identification by aural perceptual method using both contemporary and non-contemporary speech samples. Results revealed a score of 76-89% for non-contemporary for 4 weeks to six years period whereas 33% score for 20 years.

Rosenberg (1973) found great variability among the listener's ability to make correct identifications.

There are a number of drawbacks of speaker identification through the aural-perceptual method. This method is purely subjective and is vulnerable to high chances of error identification. In order to yield more accurate results for speaker identification a number of different methods have been adopted henceforth.

### **Speaker identification by visual examination of spectrograms (objective method)**

An instrument called the Sonograph was developed by Bell Telephone laboratory scientists Potter, Kopp and Grey in 1947 while studying speech signals related to communication services which was used during World War II to identify persons for intelligence purposes Lawrence Kersta a Bell System Engineer worked with this voice spectrograph (Sonograph) and observed that "voice spectrograms" renamed by Kersta (1962) as "voiceprints" could provide valuable means for speaker identification. He contended that each voice has its own unique quality and character arising out of individual variations in the vocal mechanisms. According to Kersta (1962) voice print is simply a graphic display of the unique characteristics of the voice. As a result the sound spectrograph has

attracted great interest among criminal investigators. (Saferstein, 2013).

Kersta (1962) examined the “voiceprint” using spectrograms taken from five clue words spoken in isolation using 12 talkers and closed test identification. The examiner trained high school girls 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 subjects 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”.

Though, similar results are not obtained by other researches. The correct identification scores reported by Kersta were exceptionally 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 and Pruzansky(1966), (b) 89% for short words taken from context, reported by Pruzansky (1963), (c) 84%-92%, for short words spoken in isolation, reported by Pollack, Pickett, and Sumbly(1954).

Young and Campbell (1967) studied using three words spoken by five speakers and 10 examiners with spectrogram and reported correct identification rate for words in different context as 37.3%, and word in isolation as 78.4%. The results were interpreted to indicate that different contexts decrease the identification ability of observers because: (a) the shorter stimulus durations of words in context decreases the amount of acoustic information available for matching,

and (b) the different spectrographic portrayals introduced by different phonetic contexts outweighs any intra-talker consistency.

Bolt et al. (1970) 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 reported to be 6% and for visual was 21%. They investigated and observed that 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 concluded that for visual identification, longer utterances increase the probability of correct identification.

Hecker (1971) reported that speaker recognition by visual comparison of spectrograms is used in criminology, but the validity of this method is questionable.

Tosi, Oyer, Lashbrook, Pedrey, Nicol, and Nash (1972) carried out a two-year experiment on voice identification through visual inspection of spectrograms 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 25,000 males speaking American English, all students at Michigan State University. A total of 34,996 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. 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 (1962) 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.

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 (1975) 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.

Reich (1975) 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 speaker identification through spectrograms 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 et. al. (1972).

Reich and 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). 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 and Duke, 1979). In contrast, the nasal disguise was the least effective in a previous spectrographic matching experiment (Reich, 1975). Similarly, the power spectra of nasal consonants (Glenn and Kleiner, 1968) and coarticulated nasal spectra seem to provide strong cues for the machine matching of speakers.

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. The stimuli consisted of twenty-nine bisyllabic meaning Hindi words with 16 plosives, five nasals, four affricates and four fricatives in the word-medial position. 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.

With all these technical uncertainties, forensic applications should be approached with great caution. Along with aural perceptual, spectrographic methods of speaker identifications, objective methods are also recommended in forensic speaker identifications cases.

### **Speaker identification by machine (objective method)**

In the years that followed, voice processing technologies became widespread among examiners to discriminate speaker's voices. One of the simplest methods used initially was to generate and examine amplitude and frequency, time matrices of the speech samples. Another approach was to extract speaker dependent parameters from the speech signals and analyse them by the speech identification software. The other objective method that gained popularity was the, Semi-automatic method and the Automatic method.

The first and primary method of speaker identification by machines the use of

*long term average* of acoustic features such as spectrum representations or pitch.

The second method is to model the speaker-dependent acoustic features within the individual sounds that comprise the utterance. In this method, the acoustic features from sounds in a test utterance is compared with the speaker-dependent acoustic features for similar sounds in a test utterance, the method measures speaker differences rather than textual differences. This method can be accomplished using explicit or implicit segmentation of speech into phonetic classes prior to training or recognition.

The third method of speaker recognition is the use of discriminative neural networks (NN). Discriminative NN's are trained to model the decision function which best discriminates speakers within a known set. Several different networks such as multilayer perceptrons as in the study by Rudasi and Zahorian (1991), and time-delay NN's by Bennani and Gallinari (1991), and radial basis functions by Oglesby and Mason (1990), have recently been applied to various speaker recognition tasks. Generally NN's require a smaller number of parameters than independent speaker models and have produced good speaker recognition performance, comparable to that of vector quantization (VQ) systems.

Automatic speaker verification was carried out 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 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. The data presented, showed that, if the reference samples were collected over a period of many days, then verification is possible after two months, whereas, if the reference data is collected at one sitting, then verification is highly inaccurate one hour later itself. Four authorized speakers and 30 impostors were examined, with error rates obtained from 6% to 13%. The author also noted that, when the impostors attempted to mimic the authorized speaker could deceive the system. It was observed that, the greatest accuracy would be obtained if the final decision was based on a series of two or three repetitions of the test phrase. This means that, the accuracy increases as the information available to the decision mechanism increases.

Wolf (1972) describes an investigation of an efficient approach to selecting parameters, which are motivated by known relations between the voice signal and vocal-tract shapes and gestures. In a mechanical speaker recognition experiment, it is desirable to use acoustic parameters that are closely related to voice characteristics that distinguish speakers. Significant parameters or features of selected segments were used. The investigators located speech events manually within the utterance after feeding it into a simulated speaker recognition system. The 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 2% were obtained.

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/, 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.

Atal (1972) examined the temporal variations of pitch in speech as a speaker identifying characteristics. The pitch data was obtained from 10 speakers consisting of 60 utterances, of six repetitions of the same sentence. 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 percentage of correct identifications was reported to be 97% and suggested that temporal variations of pitch could be used effectively for automatic speaker recognition.

Atal (1974) examined several different parameters using linear prediction model to determine their effectiveness for automatic recognition of speakers from their voices. He determined twelve predictor coefficients approximately once every 50 msec 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 was obtained from 10 speakers consisting of 60

utterances, of six repetitions of the same sentence. He reported that the cepstrum was found to be the most effective parameter, providing an identification accuracy of 70% for speech 50 msec 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 msec, increasing to 98% for a duration of 1sec.

The above studies propose several significant opinions. It may be concluded that n- dimensional Euclidean distance among long-term average speech spectra can be utilized successfully for speaker identification. This method has a number of merits such as, (a) It is easy and simple to carry out the procedure in the laboratory; (b) it eliminates problems of time-alignment; (c) the data generated for the identification does not depend on the power level of the speech sample used; (d) the process is objective and hence, human chance of human error is less; (e) the distortions created by the limited pass band and stress have only minimal effects on the sensitivity of the LTS vector.

Doddington (1971) developed the speaker verification system using of 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 Long Term Average Spectrum (LTAS) discriminative function relative to large populations, different languages, and speaker system distortions. 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. 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 revealed high levels of correct speaker identification for normal speech, slightly reduced scores for speech during stress and markedly reduced correct identifications for disguised

speech. In conclusion, 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 LTAS vector as a speaker identification cue.

Most current speaker recognition systems Eatock and Mason (1994), and Miyajima et al, (2001), used Mel frequency cepstral coefficients (MFCC) as the speaker discriminating features. MFCCs are typically obtained using a non-uniform filter bank which emphasizes the low frequency region of the speech spectrum. Conversely, Sambur (1975) and Orman (2000) opined that the middle and higher frequency regions of the speech spectrum carry more speaker-specific information. A study done by Kumar and Rao (2004), described a method to obtain cepstral coefficients on different warped frequency scales. This method was applied to experimentally investigate the relative importance of specific spectral regions in speaker recognition from vowel sounds. Better performance of Ozgur warping of frequency around 3 to 5 kHz was observed. It seems that for speaker recognition there can be better warping than commonly used Mel scale warping. However, this result is valid for the individual phonemes in question, and may not hold across other phonemes. So other phonemes have to be studied and also with more speakers.

Furui (1994) described the operation of the system which was based on a set of functions of time obtained from acoustic analysis of a fixed, sentence-long utterance. Cepstrum coefficients were extracted by means of LPC analysis on a frame-by-frame basis throughout an utterance. Contours of cepstral coefficients were described by time functions. The author concluded that, the verification

error rate of one percent or less can be obtained even if the reference and test utterances were subjected to different transmission conditions. Nevertheless, this study did not address the issue if the transmission system is over mobile phones.

Reynolds (1995) did a study on text independent speaker identification using GMM. The individual Gaussian components of a GMM are shown to represent some general speaker-dependant spectral shapes that are effective for modelling speaker identity. The focus of their work was on applications which require high identification rates using short utterances from unconstrained conversational speech and robustness to degradations produced by transmission over a telephone channel. The Gaussian mixture speaker model attained 96.8% identification accuracy using five seconds of clean speech utterances and 80.8% accuracy using 15 seconds of telephone speech utterances with a 49 speaker population and is shown to outperform other speaker modelling techniques on an identical 16 speakers telephone speech task.

Glenn and Kleiner (1968), describe a method of speaker identification based on the physiology of the vocal apparatus, independent of the spoken message. This experiment was based on the spectrum of nasal sounds for speaker identification in different environments in test and reference data. Power spectra produced during nasal phonation were transformed and statistically matched. Initially, the population of 30 speakers was divided into three subclasses, each containing 10 speakers. Subclass 1 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.

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. 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 and 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.



Meltzer and Lehiste (1972) investigated the relative quality of synthetic speech. They selected three speakers one man, one women and one child. They recorded a set of 10 monophthong English vowels by each speaker. Ten vowels were synthesized on a Glace-Holmes synthesizer 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.

Several studies (Jakhar, 2009; Medha, 2010; & Sreevidya, 2010) carried out to find out benchmark for speaker identification using cepstrum as a feature.

Jakhar (2009) carried out study in Hindi language in order to develop benchmark for text dependent speaker identification using cepstrum of three long vowels both live and telephone recording conditions. The results show that 88.33%, 81.67% and 78.33% for five speakers, 81.67%, 68.33%, 68.33% for 10 speakers, 60%, 50% 43.33% for 20 speakers live vs. live, mobile vs. mobile and live vs. mobile conditions respectively. This indicates that the scores increased with decrease in number of known speakers and identification score is more in similar recording condition. Among three long vowels /a: / yielded better results compared others in live recording and vowel /i: / in mobile recording.

Medha (2010) study reports that benchmark was established for text independent

speaker identification using cepstrum including both male and female participants in direct recording. Results of this study states that benchmarking for female speakers was below chance level whereas for male speakers it was 80% for the vowels /a:/ and /i:/.

Sreevidya (2010) attempted to set the benchmark in Kannada language by text independent speaker identification method using cepstrum in both direct and mobile recording conditions. The results of the study quotes vowel /u:/ with highest score (70 and 80%) in direct speech and reading and for vowel /i:/ with the highest score as (70 and 67%). Also quotes that for both the direct vs. mobile recordings, for all vowels and for groups of speakers the results were below chance level.

Therefore, 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 LTAS vectors. However, no parameter was seen to be 100% efficient across conditions and disguise.

Psychophysical studies of the frequency resolving power of the human ear has motivated several investigators to model the non-linear sensitivity of the human ear to different frequencies. The frequency response of the basilar membrane in the human ear has a very selective response pattern. This selective response pattern simulates as a bank of band pass filters equally spaced in the Bark scale. Figure 4 shows the linear spacing between 100 Hz to 1 kHz and the logarithmic spacing above 1 kHz. It has been observed that in the high frequencies, the  $F0$  must change more than a human listener can hear a difference between two tones. *Mel* is a unit of perceived fundamental frequency. It was originally

determined by listening tests, and several analytic models have been proposed for approximating the Mel-scale. The relative amplitudes of different frequencies determine the overall *spectral shape*. Studies of the human hearing mechanism revealed that in the early phases of the human peripheral auditory system, the input stimulus is split into several frequency bands within which two frequencies are not distinguishable. These frequency bands are referred to as *critical bands*. The ear averages the energies of the frequencies within each critical band and thus forms a compressed representation of the original stimulus. This observation has given incentive for designing perceptually motivated filter banks as front-ends for speech and speaker recognition systems.

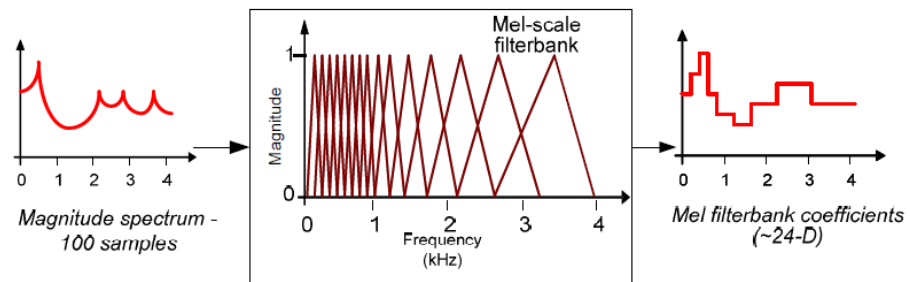


Figure 4: Illustration of Mel filtering [Taken from Milner, 2003]

Kinnunen (2003) stated that the *Mel-frequency Cepstral Coefficients* (MFCC) are the most evident example of a feature set that is extensively used in speaker recognition. While using MFCC feature extractor, one makes an assumption that the human hearing mechanism is the optimal speaker recognizer. The authors carried out a study, to put forth the general guidelines about the analysis parameters. They carried out investigations on two speech corpora using vector quantization (VQ) speaker modelling. The corpora consisted of a 100 speaker subset of the American English TIMIT corpus, and a Finnish corpus of 110

speakers. The results indicated 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. The system has been implemented in Matlab 6.1 on windows XP platform. Results showed 57.14% speaker identification for code book size of 1, 100% speaker identification for code book size of 16. Mao, Cao, Murat & Tong (2006) used *linear predictive coding (LPC) parameter and Mel Frequency Cepstrum Coefficient* (MFCC) for speaker identification. 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%.

Wang, Ohtsuka, & Nakagawa (2009) used a method that integrated the phase information with MFCC on a speaker identification task. The speech database consisted of normal, fast and slow speaking modes. 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 was done with mobile handsets over a cellular

network as against digital recording. The average MFCC vector over the entire segment was extracted using MATLAB coding. Results revealed that the overall performance of speaker verification system using MFCCs was about 80% for the data base considered. The overall performance of speaker recognition was about 90% to 95% for vowel /i/. 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 MFCCs for speaker identification and the results indicated that the percent correct identification was above chance level for electronic vocal disguise for females. Interestingly vowel /u: / had higher percent identification (96.66%) than vowels /a: / 93.33 %, and /i: / 93.33%.

Rida (2014) investigated speaker identification for nasal continuants using MFCC in 10 Hindi speaking participants in the age range of 20 to 40 years. Results indicated 90 to 100% speaker recognition in Live vs. Live recording and 50% to 90% Network vs. network recording.

Nithya (2015) reported benchmark for speaker identification using Tamil nasal continuants in live recording and mobile network recording conditions. Twenty participants were considered in the study and ten sentences with three nasal continuants in Tamil were selected for the stimuli during the task. Results of the study showed that the percentage of correct identification in live recording condition was 97.6%, 85.6% and 76.5% for the nasals /m/, /n/ and / ŋ /, respectively. In mobile network conditions the scores were 83.5%, 65.8% and 68.3%.

Chandrika (2015) established benchmark for speaker identification for nasal continuants in Kannada using Mel Frequency Cepstral Coefficients in Kannada. The study included 30 male participants, 10 each in the age range of  $20 \leq 30$  years,  $30 \leq 40$  years, and  $40 \leq 50$  years. The results indicated that the percent correct speaker identification for /m/, /n/ and /ɳ/ were 82, 89, 93 in the age range of  $20 \leq 30$  years, 66, 82, 88 in the age range of  $30 \leq 40$  years, and 86, 78, 93 in the age range of  $40 \leq 50$  years, respectively for direct recordings. In network recording it was 96, 90, 84 in the age range of  $20 \leq 30$  years, 86, 91, and 84 in the age range of  $30 \leq 40$  years and 90, 88, 88 in the age range of  $40 \leq 50$  years using MFCC. Percent SPID was highest for nasal continuant /n./ i.e. 93 for age range  $20 \leq 30$  years and  $40 \leq 50$  years whereas 88 for  $30 \leq 40$  years age range; in case of network recording samples, the highest score for speaker identification is 96 for  $20 \leq 30$  years age group and 90 for  $40 \leq 50$  years age group for nasal continuant /m/; 91 for  $30 \leq 40$  years age group for the nasal continuant /n/. Percent correct identification was increased for network recorded samples. The results indicated that nasal continuant /n./ has the highest percent of correct speaker identification score in case of direct recording and /m/ and /n/ had the highest score in case of network recorded samples.

The studies mentioned above strongly provide evidence to support the extraction of MFCCs using nasal continuants over other parameters for speaker identification. Further, review of most of the studies (Reich & Duke, 1979; Reich, Moll, & Curtis, 1976; Rida, 2014) on effective disguise for speaker identification state nasal disguise and slow rate of speech are the *least effective disguises*. Therefore, nasal continuants would be the best speech sounds to

investigate speaker identification under disguise.

Nevertheless, till date there are limited studies on nasal continuants as strong phonemes for speaker identification. Scientific testimony impresses any court of law in whichever country that might be. However, for any result to be called scientific, it has to be measured, quantified and reproducible if and when the need arises. Therefore, a method to carry out these analyses becomes a must. Thus in this context, the present study aimed at establishing benchmark for speaker identification using nasal continuants in Urdu using Mel-frequency cepstral coefficients (MFCC). Specifically, the objectives of the study were to provide benchmark for speaker identification in Urdu nasal continuants using MFCC, and compare benchmarks in direct and network recording conditions.

## CHAPTER III

### METHOD

**Participants:** Ten male participants in the age range of 20 to 40 years with Urdu as their native language for oral communication were included in the study. The inclusion criteria of the speakers was,

- a.) No history of speech, language and hearing problems,
- b.) Normal oral structure,
- c.) No other associated psychological or neurological problem and
- d.) Reasonably free from cold and other respiratory illness and oral restructuring at the time of recording as assessed by the experimenter by means of history and oral structure examination.

**Stimulus:** Commonly occurring forensically related Urdu meaningful words with nasal continuants – bilabial /m/, alveolar /n/ and velar /ŋ/ were selected. The nasal continuants were embedded in 3-4 word sentences in word - initial, - medial and - final positions to maintain the naturalness of speech. The words that were selected to make up the sentences were derived based on the colloquial/ informal Urdu spoken in Chennai, Tamil Nadu. In total /m/, /n/, and /ŋ/ occurred 7, 8, and 4 times in the sentences which were as follows:

- 1) /Mand̪ɔ̃ nɔ la:kʰ hona /
- 2) /ladka dʊn'ga/
- 3) /boʊltʊn ʃʊn/



- 4) /pulɪf kʊ nəkʊ bətə/
- 5) /pãntʃ bədʒ ku a:/
- 6) /bɑʃən fən kəru n'gə/
- 7) /Məndʒ kə:m fi:ə/
- 8) /Mere kənə fi:ə/
- 9) /Mə:r du n'gə/
- 10) /duka:n Mən Məlu n'gə/

**Procedure:** The speech samples of the participants were recorded individually. The recordings were carried out in Chennai, Tamil Nadu. Informed written consent was obtained from each of the participant. The speech stimuli consisting of the sentences were written on a card. The subjects were seated comfortably, and were given the stimuli prior to the recordings to familiarize themselves to utter the sentences. Each card with one stimulus was presented to the participant visually. They were instructed to utter the sentences thrice at an interval of 1 minute. They were instructed to speak under two conditions, directly into the recorder (direct) and through another mobile into the recording mobile phone (network). The participants read out the sentences which were recorded simultaneously in the recorder and the network using an Olympus LS-10S PCM recorder (Olympus America Inc.) at a sampling frequency of 96 kHz and 24 bits rate resolution. The recorder was held at an approximate distance of 10 cm from the mouth of the participant. The network used for making the calls was Vodafone and the receiving network was Vodafone on a Lenovo mobile phone. The speech communicated at the receiving end were recorded and saved in the SD

card of the mobile phone. Later the .amr format files were converted to .wav files using Media.io an online audio converter website, so that analysis could be carried out in an effective manner on the computer.

**Speech Segmentation:** The .wav converted speech sample wave was opened with the Praat software (Boersma and Weenink, 2016) and the words with nasal continuants at word - initial, - medial and - final positions were identified and segmented base on visual inspection of spectrogram. A portion of the nasal phonation (min 30 ms), in each occurrence, for one session and speaker was segmented and saved as .wav file for each speaker for all the nasal continuants. Figure 5 illustrates segmentation.

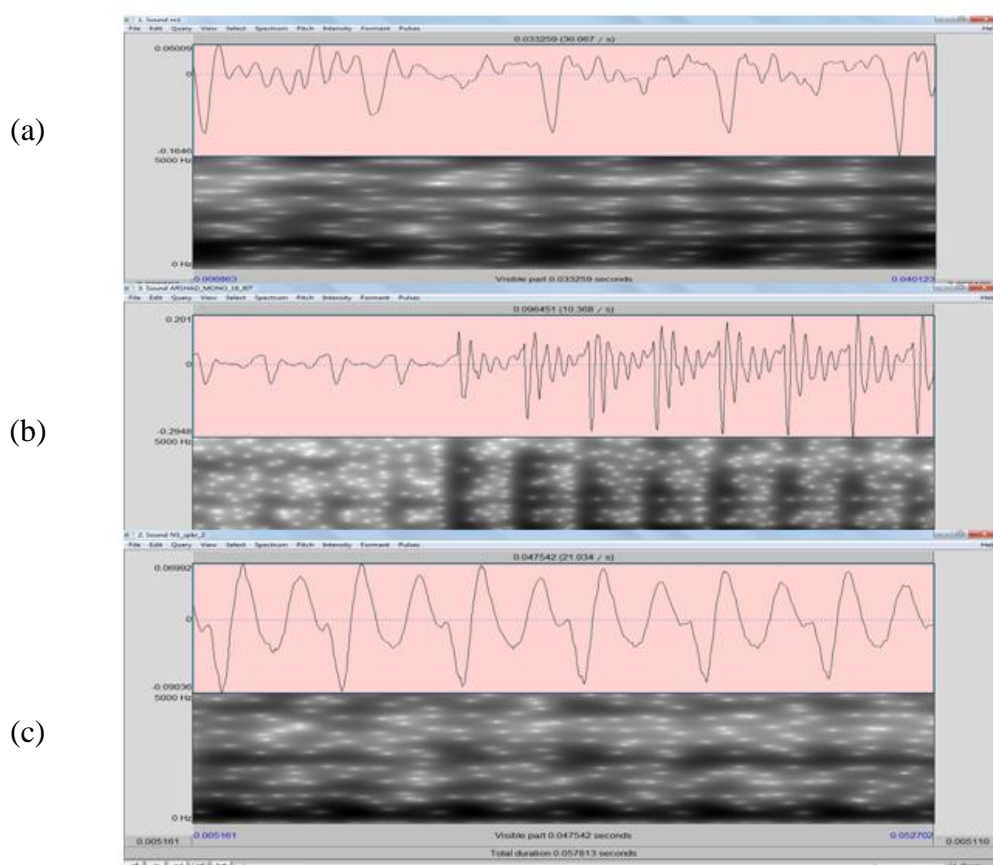
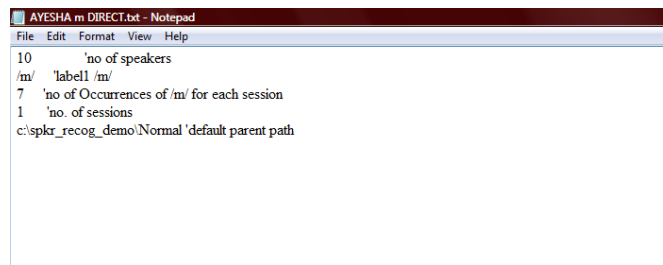


Figure 5: Segmentation of samples for (a) /m/, (b) /n/ and (c) /n/

A total of 19 nasal continuants occurred in these 10 sentences. Thus, the total number of samples for each speaker was 114 ( $19 * 3 * 2$ ), and the total number of samples for 10 speakers were 1140.

**Procedure:** SSL Work Bench (Voice and Speech Systems, Bangalore, India) was used for analyses. The nasal continuants were segmented. Initially the files were specified using a notepad and .dbs file that is extension of the notepad file were created. Figure 6 illustrates the notepad.



```
AVESHA m DIRECT.txt - Notepad
File Edit Format View Help
10 'no of speakers
/m/ 'labell /m/
7 'no of Occurrences of /m/ for each session
1 'no. of sessions
c:\spkr_recog_demo\Normal 'default parent path
```

Figure 6: Illustration of the notepad.

The segmented material was analyzed to extract 13 MFCCs (In the SSL Workbench, the sampling frequency is 8 kHz and therefore the analysis can be done up to 4 kHz, within 4 kHz only 13 Mel-frequency cepstral co-efficients (MFCC) can be computed efficiently).The formula for linear frequency to Mel frequency transformation used was constant times  $\log(1+f/700)$ . The frequency response of Mel filter bank for un-normalized and normalized conditions is shown in figures 7 and 8, respectively.

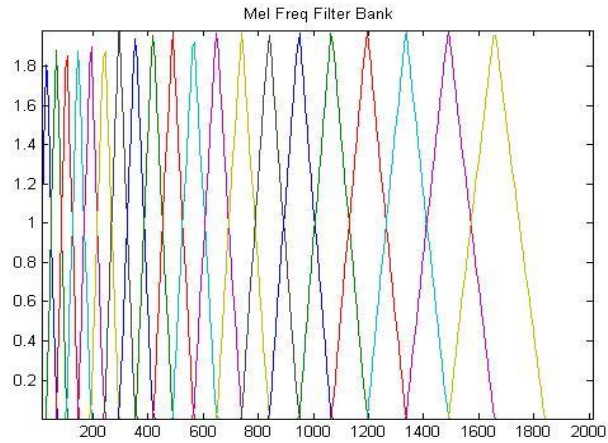


Figure 7: Mel frequency filter bank without normalization.

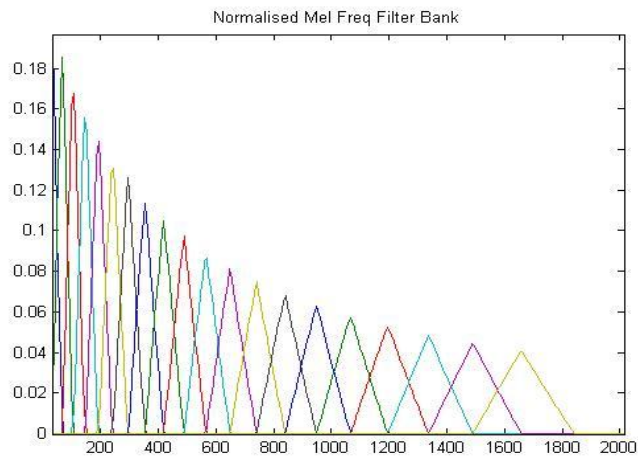


Figure 8: Mel frequency filter bank with normalization.

The notepad file was opened in SSL Workbench as in figure 9.

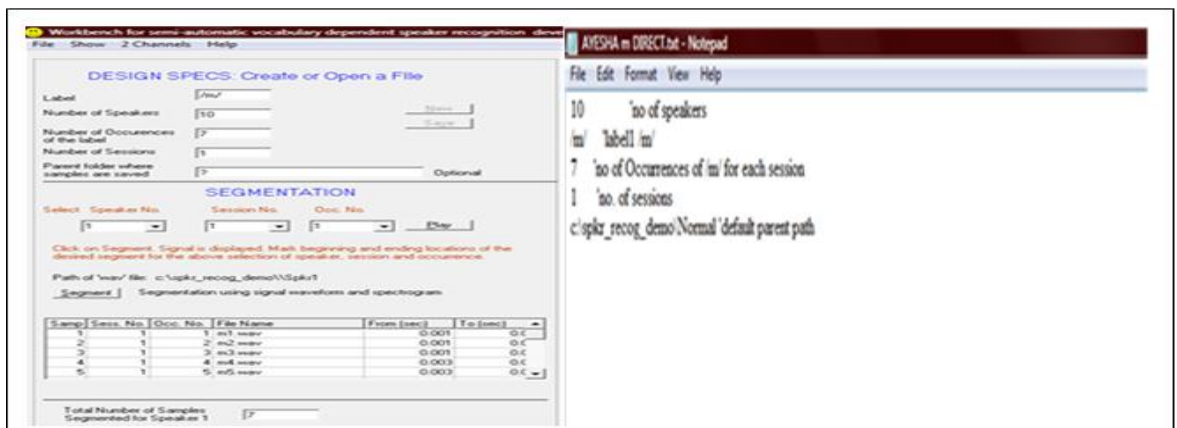


Figure 9: Notepad of SSL workbench.

The 'number of occurrence' was specified according to the occurrence of nasal continuant being studied. The 'number of sessions' was specified as 1 for the results, as the participants will utter each sentence thrice, but only one set of utterance was considered for the analysis. The parent file name was also specified in the notepad file. This is the file where the recordings were saved and was the database for the software search. The notepad file was opened in SSL Workbench. When this is opened, the 'label', 'number of occurrence', and 'number of sessions' will appear on the window as they are already fed in to the software. The experimenter selected the recording to be analyzed and marked the segment according to the session number and occurrence number. This was done by clicking on the 'segment' button which opened the location specified in the parent file path of notepad file. Following this, the experimenter chose the file from the folder. Figure 10 shows the workbench window for analyses.

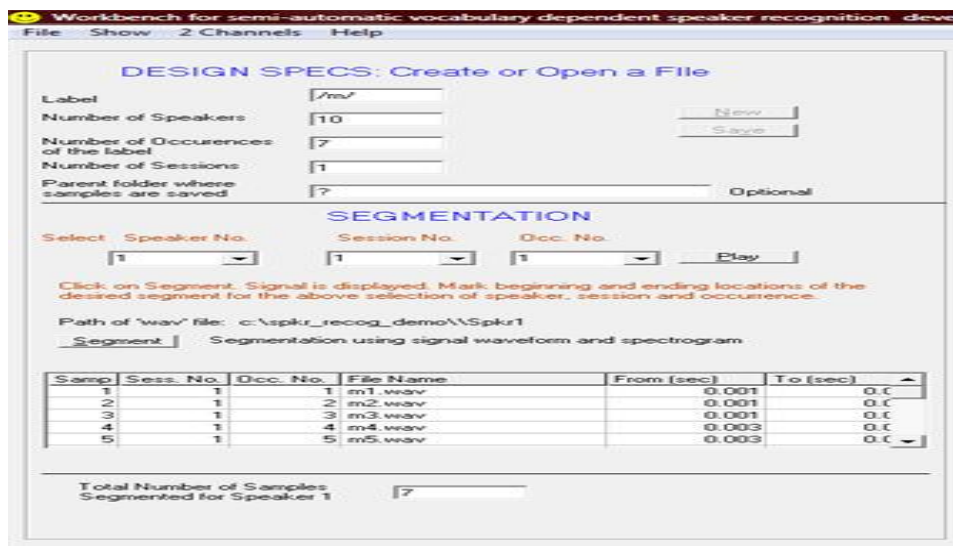


Figure 10: SSL Workbench window for analysis.

Following this, samples for analyses were segmented. To do this, the speaker number, session number and occurrence number were specified because averaging and

comparison takes place between the same samples at different sessions. Figure 11 illustrates the speaker number being selected for segmentation.

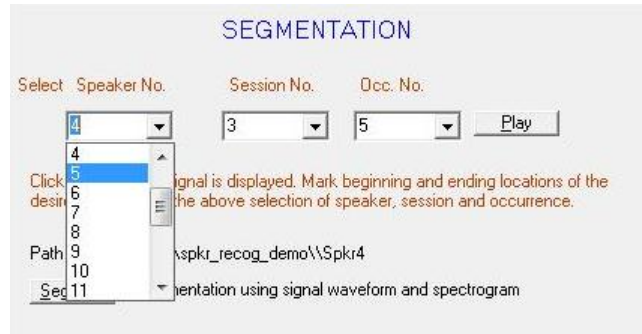


Figure 11: Illustration of speaker number being selected for segmentation.

The speaker number was selected from the options given which was already fed into the system according to the number specified for that result in the notepad file. In the same manner the session number and occurrence number were selected. Figure 12 illustrates selecting the session number and occurrence number.

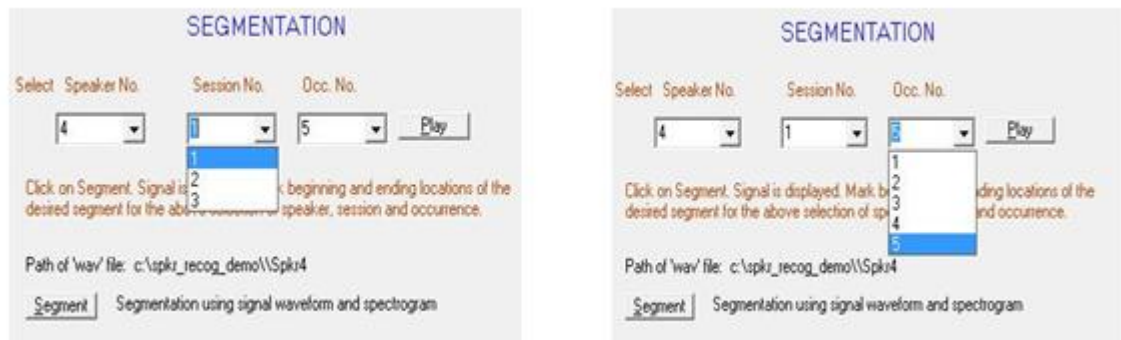


Figure 12: Illustration of selecting the session number and occurrence number.

Once these selections were made, 'segment' button was clicked on to open the dialogue box for selecting the file from the parent path specified. Following this the window will open for segmentation. Figure 13 illustrates segmentation window showing one occurrence of /m/ for a speaker.



Figure 13: Depiction of segmentation window showing one occurrence of /m/ for a speaker.

The segment of the file required was selected, and the option of 'assign highlighted' were selected from the 'Edit' menu. After this, confirmation was done. Figure 14 shows the dialogue box seeking for confirmation of the highlighted segment in the file.

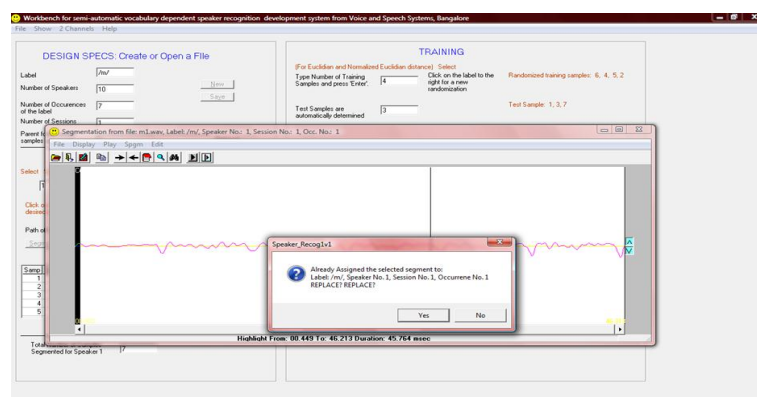


Figure 14: Showing dialogue box asking for confirmation of the highlighted segment in the file.

After all files were segmented for all the speakers, 'save segmentation' option was selected from the 'File' menu and the highlighted segment was saved onto the .dbs file created as the extension of the notepad file. Following segmentation, training was done in another window. In this window, 13 MFCC was selected and the sample for identification was tested. Figure 15 shows the analysis window of SSL Workbench.

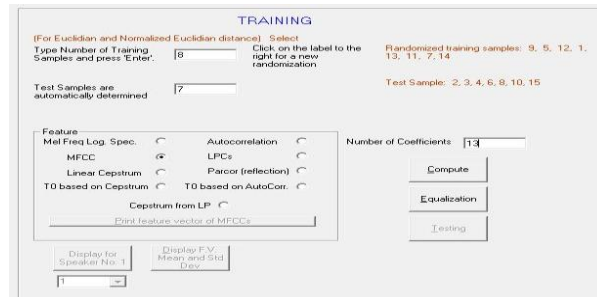


Figure 15: Analysis window of SSL Workbench.

Training sample numbers was specified and the rest were automatically selected as test samples by the system. Once this was done, 'compute' was clicked on. On clicking this option the system will check all the samples and compare them grossly and give a qualitative analysis of each speaker. Following this, the 'testing' button was clicked on. This will open a window in which 'compute score for identification' was clicked on. This gave the diagonal matrix in the lower half of the window (figure 16) and a final percentage for correct speaker identification.



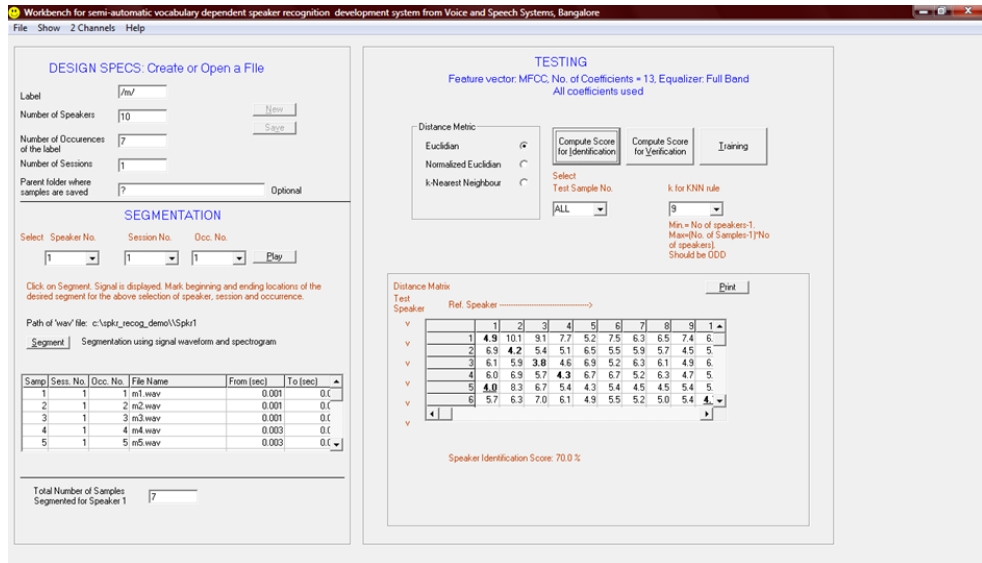


Figure 16: Analysis window of SSL Workbench showing diagonal matrix and the final speaker identification score.

This data was stored and the same procedure was repeated. Direct and network recordings were repeated 5 times. Repetitions were done by randomizing the training samples and the speaker identification thresholds were noted for the highest score and the lowest score.

Euclidian Distance for the mobile and network derived MFCC were extracted. The Euclidean distance between point's p and q is the length of the line segment connecting them  $(\overline{pq})$ . In Cartesian coordinates, if  $p = (p_1, p_2, \dots, p_n)$  and  $q = (q_1, q_2, \dots, q_n)$  are two points in Euclidean n-space, then the distance from p to q, or from q to p is given by:

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

The Euclidian distance between 13 MFCCs was extracted and noted within and between participants was. Participants having the least Euclidian distance were

considered to be the same speakers. If the distance between the unknown and corresponding known speaker is less, the identification were considered as correct. If the distance between the unknown and the corresponding known speaker is more, then the speaker is considered to be falsely identified as another speaker. The percent correct identification was calculated using the following formula:

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

In this study, all the speech samples were contemporary, as all the recordings of participants were carried out in same session. Closed set speaker identification tasks were performed, in which the experimenter was aware that the ‘unknown speaker’ is one among the ‘known’ speakers. Also, text-independent mode was adopted since the unknown and known speaker’s samples used for analyses were of different context.

## CHAPTER IV

### RESULTS

Results of the study will be discussed under the following headings:

- 1) Comparison of MFCC of the speakers for the three nasal continuants /m/, /n/ and /n:/ - direct recordings vs. direct recording
- 2) Comparison of MFCC of the speakers for the three nasal continuants /m/, /n/ and /n:/ - network recordings vs. network recording
- 3) Comparison of MFCC of the speakers – direct recordings vs. network recordings – but each considered as a different speaker for the three nasal continuants /m/, /n/ and /n:/

- 1) **Comparison of MFCC of the speakers for the three nasal continuants /m/, /n/ and /n:/– direct recordings vs. direct recording**

Results indicated correct percent identification score for /m/, /n/ and /n./ as 70, 80 and 100, respectively. The reference average is taken along the row and the test sample is taken along the column. The Euclidian distance of the data was averaged by the software separately for the test sample and the reference sample of the same speaker. These were then compared against all the speakers. The one with the least displacement from the reference was identified as the test speaker. The items in bold in the table below indicates the correct identification of the speaker as belonging to the same speaker as the reference. The tables 1 to 3 represent the Euclidian distance as given by the workbench software. Sp refers to speaker in all the tables below.

Sp	1	2	3	4	5	6	7	8	9	10
1	<b>4.924</b>	10.123	9.055	7.711	5.240	7.447	6.233	6.476	7.366	6.090
2	6.895	<b>4.170</b>	5.480	5.109	6.497	5.519	5.951	5.676	4.456	5.152
3	6.120	5.857	<b>3.819</b>	4.616	6.917	5.189	6.280	6.069	4.921	6.574
4	5.969	6.865	5.752	<b>4.262</b>	6.663	6.666	5.153	6.262	4.746	5.621
5	<b>4.008</b>	8.278	6.737	5.378	4.263	5.437	4.517	4.525	5.373	5.330
6	5.685	6.302	6.990	6.138	4.941	5.493	5.266	5.046	5.364	<b>4.702</b>
7	4.411	7.193	5.909	4.853	4.662	5.371	<b>3.880</b>	4.378	4.295	4.729
8	5.244	7.490	6.719	5.844	4.818	5.153	4.562	<b>3.346</b>	4.701	4.390
9	5.371	5.666	5.214	4.444	5.985	6.159	3.284	4.146	<b>3.186</b>	3.490
10	4.250	5.651	4.870	4.378	4.623	4.571	3.347	<b>2.747</b>	2.855	3.170

Table 1: Diagonal matrix – direct vs. direct recording speaker identification /m/

Sp	1	2	3	4	5	6	7	8	9	10
1	<b>3.589</b>	7.272	5.582	5.551	4.716	6.046	4.722	5.741	4.150	4.178
2	7.332	<b>4.213</b>	4.800	5.030	6.754	5.831	6.582	7.270	6.507	5.800
3	5.609	5.151	2.949	<b>2.834</b>	4.764	5.587	4.586	5.829	4.369	4.373
4	6.136	5.439	4.375	<b>3.797</b>	5.659	6.168	5.024	6.785	4.359	5.130
5	6.358	9.986	7.749	7.254	<b>5.065</b>	6.989	6.166	5.360	6.443	6.019
6	6.285	6.681	5.802	5.746	4.876	<b>3.013</b>	6.018	5.165	6.376	4.947
7	5.522	8.051	6.159	5.734	<b>4.002</b>	6.177	4.249	4.309	4.605	4.709
8	5.393	7.812	5.978	5.426	3.699	5.503	4.796	<b>2.636</b>	4.890	4.009
9	4.343	7.297	5.578	5.066	3.850	6.177	3.899	4.124	<b>3.564</b>	4.157
10	4.607	7.200	5.219	5.652	4.162	5.021	4.963	4.091	5.005	<b>3.190</b>

Table 2: Diagonal matrix – direct vs. direct recording speaker identification /n/

Sp	1	2	3	4	5	6	7	8	9	10
1	<b>2.700</b>	9.419	7.839	5.980	4.684	5.425	4.255	5.624	4.748	5.501
2	9.436	<b>2.903</b>	5.665	6.209	8.582	6.282	7.158	8.376	7.199	7.745
3	8.932	4.893	<b>2.767</b>	4.969	6.792	5.969	6.645	6.423	6.238	6.484
4	7.727	4.656	3.726	<b>2.412</b>	5.603	4.873	4.998	5.073	4.632	5.872
5	6.174	7.538	6.412	4.746	<b>2.537</b>	3.654	4.026	3.012	3.825	3.561
6	7.186	6.150	5.065	4.779	5.102	<b>2.476</b>	5.029	4.754	4.489	4.752
7	5.178	7.649	7.425	5.425	4.402	4.444	<b>2.847</b>	4.332	3.845	3.398
8	7.583	7.495	6.425	5.442	4.217	4.523	5.241	<b>2.239</b>	5.169	3.533
9	6.636	6.828	6.386	3.345	3.057	4.119	3.239	3.320	<b>2.238</b>	4.370
10	6.435	7.621	6.383	5.694	3.839	4.717	4.037	3.439	4.065	<b>2.177</b>

Table 3: Diagonal matrix – direct vs. direct recording speaker identification /n/

2) **Comparison of MFCC of the speakers – network recordings vs. network recording for the three nasal continuants /m/, /n/ and /n/**

Results indicated correct percent identification score for /m/, /n/ and /n/ as 60, 70 and 60, respectively. The reference average is taken along the row and the test sample is taken along the column. The Euclidian distance of the data was averaged by the software separately for the test sample and the reference sample of the same speaker. These were then compared against all the speakers. The one with the least displacement from the reference was identified as the test speaker. The items in bold in the table below indicates the correct identification of the speaker sample as belonging to the same speaker as the reference sample. In the tables below some of the items have been identified as different speakers which have been indicated as bold. The tables 4 to 6 represent the Euclidian distance as given by the workbench software. Sp refers to speaker in all the tables below.

Sp	1	2	3	4	5	6	7	8	9	10
1	<b>6.165</b>	7.512	6.196	6.309	7.089	6.758	7.564	8.221	8.002	8.266
2	8.239	<b>5.369</b>	8.920	7.022	9.845	9.162	6.171	9.618	6.431	7.281
3	7.132	7.695	7.861	7.132	8.385	8.209	<b>6.819</b>	8.545	7.411	8.076
4	7.576	7.482	<b>6.578</b>	7.008	8.641	8.019	8.236	9.994	8.720	9.313
5	8.129	8.939	8.719	7.525	7.971	7.875	7.564	<b>7.363</b>	7.455	7.622
6	5.737	9.092	4.990	5.553	5.419	<b>4.607</b>	8.158	6.654	8.416	8.196
7	7.439	4.739	9.555	6.655	9.264	9.069	<b>4.444</b>	7.775	4.515	5.425
8	7.678	5.733	9.706	6.988	8.913	9.107	5.215	7.126	<b>4.917</b>	5.595
9	8.039	6.455	8.996	7.237	8.956	8.661	6.596	8.042	<b>6.297</b>	6.607
10	7.705	6.037	10.717	7.553	9.311	9.097	5.584	7.087	5.167	<b>4.914</b>

Table 4: Diagonal matrix – network vs. network recording speaker identification /m/

Sp	1	2	3	4	5	6	7	8	9	10
1	<b>3.80</b>	7.13	7.722	5.18	7.681	7.858	6.351	6.486	5.256	4.790
2	5.543	<b>1.952</b>	11.467	4.329	10.185	12.575	3.144	4.555	4.243	5.897
3	4.142	5.221	8.266	<b>3.931</b>	8.237	9.382	4.455	5.383	4.035	4.603
4	3.434	4.590	7.780	<b>2.414</b>	7.283	8.944	3.946	3.779	3.127	4.218
5	7.374	8.753	6.072	7.424	<b>4.320</b>	6.659	8.436	7.696	7.230	5.522
6	5.363	8.748	<b>4.689</b>	6.081	5.432	5.029	7.683	7.455	6.204	5.082
7	8.413	7.133	10.297	7.700	9.970	10.806	<b>6.956</b>	7.884	7.213	7.761
8	5.838	5.495	8.286	4.486	6.808	8.542	5.268	<b>3.417</b>	4.108	4.378
9	5.477	4.037	9.536	4.384	7.899	10.408	4.314	<b>3.500</b>	3.912	5.165
10	4.771	4.112	8.648	4.389	7.113	9.503	4.479	4.963	4.242	<b>3.837</b>

Table 5: Diagonal matrix – network vs. network recording speaker identification /n/

Sp	1	2	3	4	5	6	7	8	9	10
1	<b>4.831</b>	8.267	6.421	6.910	9.305	7.646	8.124	8.542	8.521	8.885
2	5.118	3.982	3.950	<b>3.796</b>	5.725	6.486	4.706	4.966	4.618	5.775
3	5.867	6.242	<b>4.562</b>	4.950	9.558	8.741	6.016	8.153	7.270	8.733
4	4.343	4.420	2.654	<b>2.378</b>	7.246	6.958	4.162	5.338	4.883	6.704
5	5.323	4.920	5.216	3.790	4.599	5.467	4.509	<b>3.445</b>	3.689	4.191
6	6.708	9.948	7.898	8.424	4.309	<b>3.056</b>	9.717	6.631	8.587	4.534
7	6.805	<b>2.239</b>	5.628	3.186	7.457	8.357	2.242	5.243	3.008	6.744
8	5.953	7.006	6.609	5.347	3.540	4.955	6.175	<b>2.330</b>	4.469	3.564
9	6.026	4.660	5.561	4.106	4.614	5.852	4.396	<b>3.144</b>	3.592	3.765
10	6.002	5.024	5.158	4.502	3.966	5.055	5.038	4.166	4.279	<b>3.147</b>

Table 6: Diagonal matrix – network vs. network recording speaker identification

3) **Comparison of MFCC of the speakers – direct recordings vs. network recordings – but each considered as a different speaker for the three nasal continuants /m/, /n/ and /ŋ/**

The results are discussed two situations. The Highest Percent Identification (HPI) and the Lowest Percent Identification (LPI) for each nasal continuant. The reference average is taken along the row and the test sample is taken along the column. The Euclidian distance of the data was averaged by the software separately for the test sample and the reference sample of the same speaker. These were then compared against all the speakers. The one with the least displacement from the reference was identified as the test speaker. The items in

bold in the tables below indicates the correct identification of the speaker sample as belonging to the same speaker as the reference sample. In the tables below some of the items have been identified as different speakers which have been indicated as bold.

It was found that the HPI for the nasal continuants /m/, /n/ and /n./ was 50, 85 and 85, respectively. The LPI for the nasal continuants /m/, /n/ and /n./ were found to be 45, 70 and 70, respectively. This indicated that /n./ was found to be the best nasal continuant for speaker identification through MFCC. Percent speaker identification was very poor when direct recordings were compared with network recordings. Tables 7 to 13 show the results obtained under these conditions.

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	<b>4.5</b>	9	7.7	6.3	5.9	6.2	5.5	5.7	6.8	5.3	11.2	12.5	10.7	9.9	10.4	12.1	13.8	13.1	12.2	14.2
2	8.1	<b>3.5</b>	5.6	5.4	6.9	6.4	6.7	6.1	4.5	5.8	8.3	9.7	7.0	5.9	7.2	8.1	11	11.1	9.6	12.9
3	7.8	5.5	<b>4.8</b>	6.1	8.4	6.9	7.2	6.8	5.6	6.8	6.6	8.5	5.6	5.1	6.2	6.8	9.8	9.6	8.2	11.8
4	5.9	5.6	4.2	<b>3.5</b>	7.9	6.7	4.5	5.8	4.1	5.2	6.8	7.9	6.6	5.9	5.9	7.4	9	8.4	7.4	9.5
5	<b>4.7</b>	8.7	7	6.1	6.1	5.5	5	5.2	6.5	6.1	9.8	11.3	9.1	8.6	8.8	10.5	12.7	11.6	10.8	13.7
6	6.8	5.2	5.7	5.2	5.7	5.6	6.1	5.7	<b>4.8</b>	5.4	8.6	10.1	7.5	6.4	7.9	8.9	11.5	11.3	9.9	13.2
7	5.6	6.4	6.3	5	5.5	5.9	<b>4.3</b>	4.9	4.5	4.5	9.3	10.3	8.5	7.5	8	9.7	11.7	11.3	10.2	12.5
8	6	7.1	6.6	5.8	5.8	5.3	4.7	<b>3.7</b>	4.8	4.3	10	11.2	8.8	7.9	8.4	10	12.6	11.7	10.7	13.8
9	5.4	5.2	5.1	4.2	6.5	6.1	3.4	4	<b>2.9</b>	3	8.4	9.1	7.7	6.6	6.9	8.7	10.4	10.0	8.8	11.1
10	5	5.3	5	4.1	5.8	5.1	3.3	3.2	<b>3.1</b>	3.2	8.4	9.5	7.6	6.6	7.1	8.8	11	10.3	9.2	11.8
11	11.3	9.7	8	9.6	13.6	11.7	10.8	11.5	10	11.7	<b>6.3</b>	8	6.9	7.6	7.3	7.1	8.7	7.2	7.2	8.6
12	9.7	<b>6.8</b>	8.7	8	9.9	9.6	9.1	9.3	7.6	8.9	9.1	7.3	8.6	6.8	9	9.9	7.9	10.3	8.5	11.1
13	12.1	10.7	8.8	10.7	14.2	11.8	11.6	11.8	10.8	12.5	<b>6.5</b>	7.9	6.9	7.9	7.5	6.9	8.5	6.9	6.9	9.7
14	11.1	8.8	7.5	9.1	13.1	11.1	10.1	10.5	9.13	10.9	5.7	6.6	5.9	6.4	6.03	5.8	7.2	5.6	<b>5.5</b>	8.1
15	14.5	13.4	11.1	13.2	17.3	15.1	13.4	14.1	13.1	14.6	8.2	9.01	9.07	10.4	8.5	7.9	9.2	<b>6.8</b>	7.7	8.5
16	11.1	8.2	6.8	8.8	12.4	10.7	9.6	9.9	8.4	10.1	5.8	8.5	5.4	6.3	5.2	<b>5.05</b>	9.4	7.6	7.2	9.8
17	10.7	9.2	9.5	9.6	11.5	10.6	10.3	10.5	9.4	10.6	8.9	<b>6.5</b>	8.9	8.08	9.2	9.8	6.8	8.9	7.6	10.0
18	12.7	12.6	11.09	11.7	16.1	14.5	11.8	13.1	11.98	13.3	9.33	7.07	10.1	10.2	9.2	9.8	6.6	<b>6.4</b>	6.7	6.5
19	11.6	10.7	10.2	10.4	14.7	13.4	10.7	12.07	10.5	11.9	8.5	4.6	9.2	8.6	8.4	9.4	<b>4.1</b>	5.8	5.2	5.4
20	10.7	8.6	8.8	9.1	13.06	11.7	9.8	10.6	8.9	10.6	7.4	<b>4.7</b>	7.8	6.9	7.6	8.3	5.1	6.09	5.1	6.6

Table 7: Diagonal matrix (HPI) of direct vs. network recording for speaker identification /m/

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	<b>3.0</b>	6.8	5.0	5.2	4.0	5.5	4.2	5.1	3.9	3.7	11.4	11.8	11.1	12.2	13.9	13.5	11.5	13.5	12.9	12.6
2	6.7	<b>2.7</b>	4.6	4.2	8.5	5.4	6.1	6.9	6.4	5.9	11.7	11.9	10.8	11.8	15.4	14.3	12.2	13.8	13.5	13.3
3	5.2	4.9	<b>3.0</b>	3.7	5.1	4.8	4.3	4.8	4.3	3.9	10.1	11.0	9.31	10.5	12.9	11.8	10.2	12.4	11.9	11.5
4	5.3	5.2	3.5	<b>3.5</b>	5.5	5.3	4.4	5.4	4.1	4.6	9.2	10.2	8.7	9.8	12.2	11.2	9.4	11.4	11.0	10.7
5	5.3	8.0	6.2	6.3	<b>3.2</b>	5.4	4.3	3.9	4.5	4.8	12.6	13.5	12.0	13.1	14.4	13.7	12.3	14.3	13.7	13.7
6	6.1	6.5	6.2	6.0	6.5	<b>3.3</b>	5.6	5.1	6.3	5.1	14.2	15.3	13.5	14.7	17.3	15.8	14.7	16.5	16.0	16.0
7	5.3	7.3	5.5	5.9	<b>3.8</b>	5.8	3.9	4.6	4.0	4.9	11.4	12.0	10.7	11.7	13.2	12.6	10.9	13.0	12.5	12.3
8	5.2	7.4	6.2	6.1	3.9	5.4	4.0	<b>2.2</b>	4.0	4.2	13.5	13.7	12.6	13.6	15.3	14.6	12.9	14.7	14.1	14.4
9	5.1	7.5	6.3	6.2	5.0	6.4	4.8	5.4	<b>4.5</b>	5.1	12.3	12.6	11.8	12.9	14.4	13.9	12.0	13.9	13.4	13.3
10	4.3	6.0	5.1	5.5	5.2	5.3	3.9	4.2	4.2	<b>3.3</b>	13.2	13.2	12.3	13.6	15.7	14.9	12.9	15.1	14.5	14.2
11	12.8	11.7	9.98	10.3	12.0	12.7	12.1	13.0	11.4	12.3	<b>3.88</b>	7.04	4.4	4.34	5.94	7.06	5.34	6.13	5.94	4.61
12	11.8	10.9	9.84	9.86	12.1	13.2	11.7	12.7	10.7	12.3	5.7	<b>2.4</b>	4.2	4.29	7.5	10.3	5.2	4.6	4.0	4.3
13	15	13.7	12.1	12.7	13.4	14.3	13.6	14	13.3	13.7	10.4	12.6	9.9	10.1	10.5	<b>9.5</b>	10.1	11.7	11.8	10.8
14	12.1	10.6	9.19	9.35	11.6	12.2	11.3	12.1	10.5	11.8	4.5	4.81	3	<b>2.9</b>	6.6	8.39	4.5	4.4	4.1	4.1
15	16.4	15.6	13.8	14.4	14.3	16.4	14.8	15.4	14.2	15.1	8.9	11.5	9.1	8.5	<b>5</b>	6.6	7.1	8.8	9.2	7.8
16	15	13.5	11.8	12.3	13.1	14.2	13.3	13.7	12.8	13.4	7.5	11.4	7.8	7.5	6.7	<b>5.1</b>	6.9	8.7	9.3	7.8
17	11.4	10.4	9.32	9.40	11.6	12.5	11.1	12.1	10.3	11.7	6.2	<b>4.5</b>	4.9	4.9	7.9	10	5.5	5.5	5.3	5.4
18	13.7	12.4	11.2	11.1	12.9	14.1	12.8	13.4	11.8	13.4	5.9	5.9	5.1	4.1	5.6	8.3	4.6	<b>3.3</b>	3.6	4.7
19	12.3	11.5	10.1	10.2	11.6	13.0	11.6	12.3	10.7	11.9	6.3	5.5	5.8	5.6	7.1	9	5.6	5.2	<b>5</b>	5.2
20	13.1	12.3	11	11.2	12.6	14.1	12.5	13.5	11.7	12.9	6.4	5.8	6.2	6.1	6.5	8	5.4	6.2	6.3	<b>5.3</b>

Table 8: Diagonal matrix (HPI) of direct vs. network recording for speaker identification /n/

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	<b>2.2</b>	9.2	7.8	6.4	5.5	5.5	4.2	6.7	4.9	6.5	13.4	13.9	13.1	12.8	15.7	15.9	14	15.4	15.1	15.4
2	9.7	<b>3.0</b>	4.7	4.5	7.5	6.5	6.9	7.9	7.4	7.8	11	11.7	9.3	9.5	13.9	14.2	11.5	13.7	13	13.3
3	8.8	5.8	<b>2.6</b>	3.9	6.7	5.5	7.5	6.8	7	7	8.7	10	7.6	8.2	11.8	12.1	10.3	11.7	11.3	11.4
4	6.9	6.4	4.8	<b>1.8</b>	5.1	4.5	5.4	5	3.9	6.4	8.7	9.3	8.1	7.4	11	11.3	9	10.4	10	10.8
5	5.1	8.5	6.4	4.9	<b>2.4</b>	4.1	4.2	3	2.8	4.1	11	12.2	11.2	10.7	12.5	12.8	12	11.9	12.1	12.1
6	6.9	5.7	5.2	4.6	4.1	<b>2.4</b>	4.6	4.9	4.8	4.8	11	13.2	10.8	11.1	13.9	14.2	13	13.8	13.8	13.7
7	4.9	7.7	6.4	4.5	3.8	4.6	<b>2.3</b>	4.3	3	3.9	12.3	12.7	11.6	11.2	13.8	14	12.6	13.4	13.3	13.2
8	6.5	8.2	5.9	4.9	3.3	4.4	5	<b>1.7</b>	4	3.5	12	13.3	11.7	11.4	13.4	13.7	13	12.6	13.1	13.0
9	6.4	6.9	5.9	4	3.9	4.3	4	4.5	<b>2.8</b>	4.7	10.3	10.9	10	9.5	11.9	12.1	10.7	11.5	11.3	11.3
10	5.1	7.9	5.8	5.1	3.6	4.5	3.7	3.7	3.8	<b>2.5</b>	13	13.9	12.4	12.4	14.7	14.9	13.8	14.2	14.3	14.1
11	12	10.7	9	8.5	10	9	11.3	10.1	9.8	11.5	<b>4.3</b>	7.6	6	6.4	7.8	8	7.6	7.9	7.5	8.3
12	14.9	13.3	12	11	13.7	13.1	13.9	13.2	12.5	14.5	6.1	<b>3.4</b>	5.8	4.6	5.5	5.6	3.8	6.3	4.7	5.6
13	12.4	9.3	8.4	7.2	10.6	9.6	10.6	10.1	9.7	11.3	5.4	5.2	<b>3.8</b>	3.8	7.1	7.4	5.4	7.7	6.5	7.2
14	13.7	11.2	10	8.9	11.9	11.1	12.2	11.2	10.8	12.7	4.3	2.8	3.3	<b>1.2</b>	5.1	5.4	2.6	5.1	3.6	5.3
15	14.7	14.6	12.5	11.4	13	13	13.9	12.2	12.1	13.7	6.7	5.5	7.7	6.2	3.5	3.4	5.4	<b>3.4</b>	3.6	3.6
16	14.6	14.4	12.4	11.3	12.9	12.9	13.8	12.1	12	13.7	6.7	5.4	7.6	6	3.5	3.5	5.3	<b>3.5</b>	3.6	3.6
17	13.6	12.4	11.7	9.9	12.7	12.4	12.5	12.3	11	13.6	7.7	3.6	7	4.8	7.1	7.2	<b>2.9</b>	6.8	4.9	6.8
18	15.2	14.2	12.6	11.2	13.3	13	13.9	12.2	12.2	14.2	6.1	5	6.9	4.9	3.4	3.5	4.4	<b>2</b>	2.0	4.2
19	13.5	12.7	11.4	9.8	12.3	12.1	12.4	11.6	10.8	13.1	6.6	3.4	6.3	4.4	5.3	5.4	<b>3</b>	5	3.4	5.1
20	14.4	13.8	12	11	12.7	12.6	13.3	11.8	11.6	13.2	6.3	4.8	6.7	5.6	2.8	2.8	4.9	3.5	3	<b>2.3</b>

Table 9: Diagonal matrix (HPI) of direct vs. network recording for speaker identification /n/ /



Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	6	7.2	6.7	5.7	<b>5.3</b>	5.6	5.4	6.3	6.1	5.6	10.9	7.5	8.1	9.5	11.5	10.4	7.7	11.7	10	10.1
2	10.2	<b>4.2</b>	7	6.1	8	6.2	6.8	7.8	5.1	6.2	10.6	4.9	7.3	8	11.1	9.2	5.6	11.4	8.9	8.7
3	9.2	<b>4.8</b>	5.2	6	7	5.7	6.2	6.9	5.5	5.7	9.6	5.5	6	7.2	9.7	8.2	5.9	10.6	8.8	8.8
4	7.8	5.6	5.4	<b>4.2</b>	6.1	5.9	5.2	6.5	5.2	5.3	9	5.7	6.6	7.5	9.9	8.5	5.9	9.9	8	8.1
5	6.7	8.1	9.4	7.8	<b>5.6</b>	5.7	6.7	6.8	7.7	7.3	14.9	10.7	11.4	13.1	15.1	13.3	11.1	15.8	13.9	14.2
6	8.2	6.2	6.3	6.6	6.5	<b>5.7</b>	6.7	6.6	7	6.4	11.4	8.6	7.7	9.5	11.6	9.7	8.8	12.9	11.3	11.6
7	5.5	5.9	6.4	3.9	4.2	4.3	<b>2.9</b>	4	3.6	3.3	11	7.3	8.2	9.2	11	9.9	7.5	11.3	9.3	9.5
8	6	5.5	6.2	5	4.2	4	3.9	<b>3.2</b>	4.1	3.4	11.5	7.8	8	9.2	11	9.7	8	11.9	10.1	10.5
9	6.8	5	5.4	4.3	5.3	4.9	4	4.9	4	<b>3.8</b>	10	7.1	7.2	8.4	10	8.7	7.4	11	9.2	9.4
10	5.9	6	7.4	5.1	5.2	4.8	4.5	4.8	4.2	<b>4.1</b>	12.5	8.5	9.6	10.8	12.6	11	8.9	13.3	11.1	11.2
11	11.4	7.2	6.3	8.4	9.1	8.2	8.8	9.5	8.7	8.5	8.1	6.6	<b>5.0</b>	6.2	8.7	7.8	6.6	9.5	8.7	8.7
12	13.7	11.1	10.6	10.5	12.4	11.9	11.3	12.5	10.2	10.8	8.9	6.3	8.7	7.3	9.9	10.7	6	6.5	<b>5.6</b>	5.7
13	13.8	9.8	7.7	10.3	11.7	11.4	10.9	11.8	10.9	10.6	6.5	8.6	<b>5.7</b>	5.9	6.6	7	8.2	7.8	8.3	8.5
14	10.8	6.2	6.4	7.4	8.3	7.3	7.8	8.4	7.4	7.4	8.7	6.4	<b>5.9</b>	6.7	9.5	8.4	6.3	9.5	8.3	8.5
15	14.4	11.9	10.9	11.9	12.7	12.5	12.1	12.7	12.1	12	9.7	10.8	9.8	9.1	9.7	10.7	10.5	<b>8.6</b>	9.4	9.5
16	13.3	8.3	7.5	9.4	10.8	10	9.8	10.4	9.3	9.3	7.6	7.9	<b>5.8</b>	5.8	7.3	7.3	7.8	8.5	8.3	8.5
17	14.8	12.5	11	11.4	13.7	13.6	12.4	13.9	11.6	12	7.5	7.1	8.7	6.4	8.7	10	6.4	4.3	<b>3.8</b>	3.9
18	13.2	11	8.6	9.7	12	12.3	10.6	11.7	10.4	10.3	6.6	9.1	7.7	6.6	6.8	7.5	8.7	<b>6.4</b>	6.5	6.9
19	13.6	10.8	9.6	10.2	12.2	12.0	11.1	12.3	10.4	10.7	7.6	7.3	7.8	6.6	8.5	9.4	6.9	<b>5.6</b>	5.7	6.1
20	14.4	13.4	11.3	11.3	14.1	14.4	12.4	14.1	12	12.2	7.3	9.7	10.3	8.6	9	10.4	9.3	<b>5.8</b>	6.0	6.0

Table 10: Diagonal matrix (LPI) of direct vs. network recording for speaker identification /m/

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	<b>3.5</b>	7.2	5.5	5.4	4.6	6	4.6	5.7	4.1	4	11.7	11.8	10.6	12.4	14.4	13.2	11.6	14.1	13.1	13.2
2	7.3	<b>4.2</b>	4.7	5	6.8	5.8	6.5	7.2	6.4	5.3	12.01	12.09	9.9	12.1	15.1	13.1	11.8	14.2	13.7	13.8
3	5.5	5.7	3.5	<b>3.4</b>	4.3	5.6	4.3	5.5	4.1	3.6	9.3	10.3	7.9	10	12.2	10.5	9.9	11.8	11.5	11.2
4	6	5.5	4.4	<b>3.9</b>	5.5	6.1	4.9	6.7	4.2	4.5	9.4	10.1	8.2	10	12.7	10.8	9.6	11.8	11.3	11.2
5	6	10.3	7.9	7.2	<b>4.8</b>	7.6	5.8	5.3	5.9	6.4	12.9	13.6	11.9	13.5	14.6	13.4	13	14.7	14.3	14.1
6	6.3	6.7	5.8	5.6	4.9	<b>3.1</b>	6.1	5	6.4	5.4	13.3	14.4	11.9	13.9	16.6	14.5	14	15.9	15.7	15.6
7	5.4	7.9	6	5.6	<b>3.8</b>	5.9	4.2	4.2	4.6	4.7	12.5	12.9	11	12.8	14.4	12.9	12.4	14.3	14	13.9
8	5.3	7.7	5.9	5.3	3.6	5.5	4.7	<b>2.6</b>	4.8	4.3	12.9	13	11.3	13	15	13.5	12.4	14.4	14	14.4
9	5.4	8.3	6.7	6.2	4.7	6.5	5.2	<b>4.6</b>	5.01	5.08	13.2	13.1	11.7	13.4	15.2	13.9	12.6	14.8	14.1	14.4
10	4.5	7.1	5.1	5.5	4.1	5	4.8	4	4.9	<b>3.3</b>	13.4	13.4	11.7	13	15.9	14.5	13.2	15.5	14.9	15.2
11	12.7	11.8	10.8	10.2	12.1	13.7	11	13.6	10.7	12	<b>3.3</b>	6.5	5.1	4.5	6.1	6.3	6.4	5.4	6.3	4.4
12	12	10.9	10.6	10	12.3	14	10.7	13.5	10.3	12.1	6.2	<b>2.1</b>	5.1	4.4	8	10.05	3	5.4	3.06	5.6
13	15.6	15.2	13.7	13.1	14.2	15.9	13.6	15.2	13.6	14.3	9.2	11.6	9.5	9.4	9.3	<b>8.1</b>	11.3	9.9	11.5	9.6
14	12.1	10.5	9.8	9.2	11.4	13	10.2	12.7	10	11.4	3.9	4.6	3.3	<b>2.5</b>	6	7.1	4.2	3.8	4.5	4.6
15	15.9	16.3	14.5	13.9	14.5	17.2	13.7	15.7	13.7	14.8	8.4	10.9	9.7	8.8	<b>4.6</b>	5.8	10.1	7.5	9.8	7
16	14.9	14.5	12.8	12.2	13.2	15	12.6	14.4	12.6	13.2	7.3	11.3	8.9	8.5	6.4	<b>3.4</b>	10.6	8.1	10.8	7.4
17	13.4	12.7	11.8	11.3	13	14.7	11.8	14.2	11.5	12.9	8.3	7.8	7.9	7.5	8.9	9	<b>7.5</b>	7.8	7.8	7.7
18	13.3	12.2	11.8	10.8	12.8	14.7	11.5	13.8	11	12.9	6.4	5.3	6.3	4.6	6.8	8.1	4.4	<b>3.7</b>	3.8	4.9
19	12.5	12.1	11	10.2	11.6	13.5	10.6	12.61	10.3	11.7	6.2	6.9	6.6	5.6	6.6	7.2	6.3	<b>5</b>	5.9	5.4
20	12.1	11.6	10.5	10.08	11.8	13.9	10.4	13.1	10.1	11.6	5.7	<b>4.6</b>	5.1	4.8	5.7	7.1	4.9	5.3	5.2	4.7

Table 11: Diagonal matrix (LPI) of direct vs. network recording for speaker identification /n/

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	<b>2.3</b>	10	8.4	6.8	5.1	6.2	4.8	7	6	5.9	11.2	14.9	11.1	13.5	14.2	14.1	14.2	15.4	14.4	14.7
2	8.9	<b>2.4</b>	5.3	5.3	7.8	6.2	6.6	8	6.6	7.3	9.6	13.4	7.8	10.6	13.3	13.2	12.7	14.5	13.3	13.6
3	8.2	4.8	<b>2.1</b>	4.1	6.9	5	7.3	6.9	6	6.6	7	10.5	6	8.3	10.7	10.5	10.9	11.8	10.8	10.9
4	6.5	5.4	4.3	<b>1.2</b>	4.8	4.2	5.2	5	3.5	5.7	6.7	10.3	5.8	8	9.6	9.4	9.5	10.7	9.7	10.3
5	5.5	7.9	6	4.8	<b>2.5</b>	3.8	4.1	3.5	2.8	4.3	7.9	12.5	8.8	10.8	11.1	11	12.1	11.8	11.3	11.6
6	6.2	6.3	5.5	5	4.8	<b>2.6</b>	4.5	5.2	4.8	5	8.6	13.8	8.9	11.5	13.2	13.1	13.6	13.9	13.4	13.6
7	4.6	7.7	6.6	4.5	3.4	4.7	<b>2.1</b>	4.3	3.3	3.2	10	13.4	9.2	11.6	12.4	12.3	12.6	13.3	12.4	12.7
8	6.2	8	5.9	5	3	4.2	4.9	<b>2.0</b>	4.2	3.8	9.3	13.2	9.3	11.2	12	11.9	13.1	12.4	12.1	12.5
9	5.4	7.6	6.7	4.3	3.9	4.3	3.5	4.3	<b>2.6</b>	4.6	8.5	12.2	8.6	10.5	11	10.9	11.3	11.8	11.1	11.5
10	5.6	7.9	5.9	5.7	3.1	4.3	4.3	3.3	4.1	<b>2.0</b>	10.4	14.5	10.4	12.7	13.4	13.3	14.2	14.0	13.4	13.5
11	14.1	11.8	10.5	10	12.8	11.1	13.4	12.4	11.3	13.7	4.4	4.6	5.8	<b>4.3</b>	5.3	5.2	6.6	5.7	5.5	5.9
12	13.9	11.5	11.4	9.9	13.3	12.4	13.1	13.2	11.1	13.8	7.9	3.7	6.3	4.7	4.6	4.5	<b>2.3</b>	6.8	4.4	4.9
13	14.2	10.5	10	9.4	13.3	11.3	13.0	12.5	11.1	13.2	6.1	3.4	4.3	<b>2.6</b>	5.3	5.2	5.5	6.5	5.1	5.1
14	13.1	10.2	10	8.4	11.9	10.8	11.9	11.4	9.9	12.5	6.3	3.2	4.5	<b>2.2</b>	3.5	3.3	2.8	5.1	3.5	4.4
15	16.1	14.9	13.5	12.6	14.4	13.7	15.1	13.5	12.9	15.0	8.9	5.6	9.4	7.2	4.1	4.2	7.4	<b>3.5</b>	4.3	4
16	16.4	15.3	13.9	12.9	14.6	14.1	15.4	13.8	13.2	15.3	9.3	5.8	9.8	7.6	4.4	4.5	7.6	<b>3.6</b>	4.6	4.2
17	13.5	11.3	11.3	9.4	12.5	12	12.5	12.3	10.5	13.3	8	4.8	6.3	4.7	4.7	4.6	<b>2.9</b>	6.4	4.6	5.5
18	15.1	13.3	12.3	10.9	13.3	12.8	14.0	12.4	11.7	14.2	8	4.4	7.8	5.2	2.7	2.7	5	<b>1.8</b>	2.6	4
19	14.1	12.4	11.9	10.1	12.9	12.4	13.2	12.4	11.1	13.8	8	4	6.9	4.7	3.3	3.3	<b>2.9</b>	4.5	3.0	4.4
20	15.2	13.6	12.4	11.5	13.2	12.8	14	12.4	11.6	13.8	8.5	5.5	8.5	6.7	3.7	3.7	6.6	3.4	3.3	<b>3</b>

Table 12: Diagonal matrix (LPI) of direct vs. network recording for speaker identification /n/

	/n'/	/n/	/m/
Highest Percent Identification	85	85	50
Lowest Percent Identification	70	70	40

Table 13: Percent correct identification for all nasal continuants

To summarize, the percent correct speaker identification for /m/, /n/ and /n'/ was 70, 80 and 100, respectively when direct recordings were compared with direct recordings using MFCC. The percent correct speaker identification score for /m/, /n/ and /n'/ was 60, 70 and 60, respectively when network recordings were compared with network recordings using MFCC. The Highest Percent Identification (HPI) score on twenty randomizations for /m/, /n/ and /n'/ was 50%, 85% and 85%, respectively when direct recordings were compared with network recordings using MFCC. The Lowest Percent Identification (LPI) score on twenty randomizations for /m/, /n/ and /n'/ was 45, 70

and 70, respectively when direct recordings were compared with network recordings using MFCC. Overall, the results revealed that the nasal continuant /n/ had the best percent correct speaker identification among /m/ and /n/ that were considered in the current study. Table 14 shows the summary of the percent correct speaker identification. Figure 17 shows a graphical representation of the percent correct identification under the three conditions.

CONDITION	Percent correct identification		
	/m/	/n/	/ŋ/
Direct vs. Direct recording	70	80	100
Network vs. Network recording	60	70	60
Direct vs. Network recording - HPI	50	85	85
Direct vs. Network recording - LPI	45	70	70

Table 14: Summary of the percent correct speaker identification.

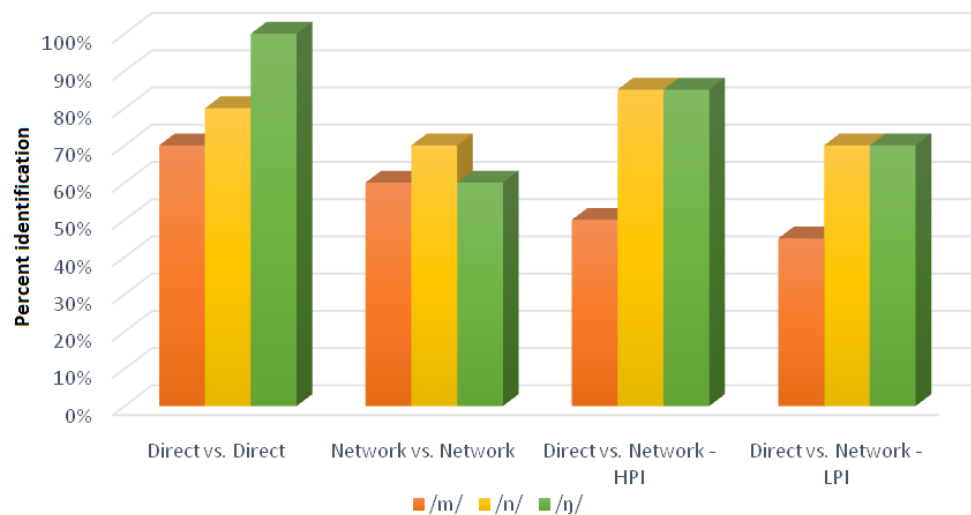


Figure 17: Percent identification under three conditions.

## CHAPTER V

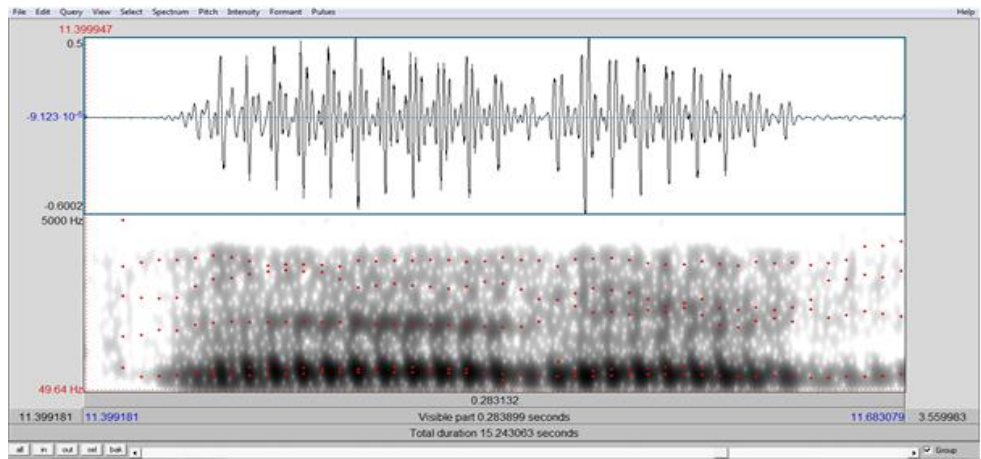
### DISCUSSION

The results of the present study revealed several significant findings. *The percent correct speaker identification score for /m/, /n/ and /ŋ/ was 70, 80 and 100, respectively when direct recordings were compared with direct recordings using MFCC.* These results are in consonance with several studies done in past. Hasan, Jamil, Rabbani, & Rahman (2004) reported speaker identification score of 57.14% for code book size of 1, 100% speaker identification for code book size of 16, using MFCCs for feature extraction and vector quantization based in speaker identification in security system. Mao et al., (2006) reported that 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%. Rajsekhar (2008) reported 75% identification in MFCC using the word “zero”. Wang et al., (2009) reported that the 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. Tiwari (2010) reported improvement in percent correct speaker identification when the number of filters used was increased in MFCC. The author reported 85% efficiency using MFCC with 32 filters in a speaker recognition task. Chandrika (2010) reported that the overall performance of speaker verification system using MFCCs was about 80% for the data base considered. The overall performance of speaker recognition was about 90% to 95% for vowel /i/. Ramya (2011) reported that the percent correct identification was above chance level for electronic vocal disguise for females using MFCC for speaker identification. Remarkably vowel /u: / had higher percent identification (96.66) than

vowels /a: / 93.33, and /i: / 93.33. Patel and Prasad (2013) reported an error rate of 13% for the word “hello” using MFCC. Rida (2014) reported speaker identification scores for nasal continuants in Hindi using MFCC. Scores ranged from 90 to 100% for speaker identification in live vs. live recording and 50% to 90% for network vs. network recording. Nithya (2015) reported benchmark for speaker identification using three Tamil nasal continuants in live recording and mobile network recording conditions. Results of the study showed that the percentage of correct identification in live recording condition was 97.6%, 85.6% and 76.5% and in mobile network conditions the scores were 83.5%, 65.8% and 68.3%. Chandrika (2015) reported benchmark for speaker identification using three Kannada nasal continuants in live recording and mobile network recording conditions. The author had also compared the MFCCs across three age groups of  $20 \leq 30$  years,  $30 \leq 40$  years, and  $40 \leq 50$  years. Results of the study revealed that the nasal continuant /n / had the highest percent of correct speaker identification score in case of direct recording and /m/ and /n/ had the highest score in case of network recorded samples.

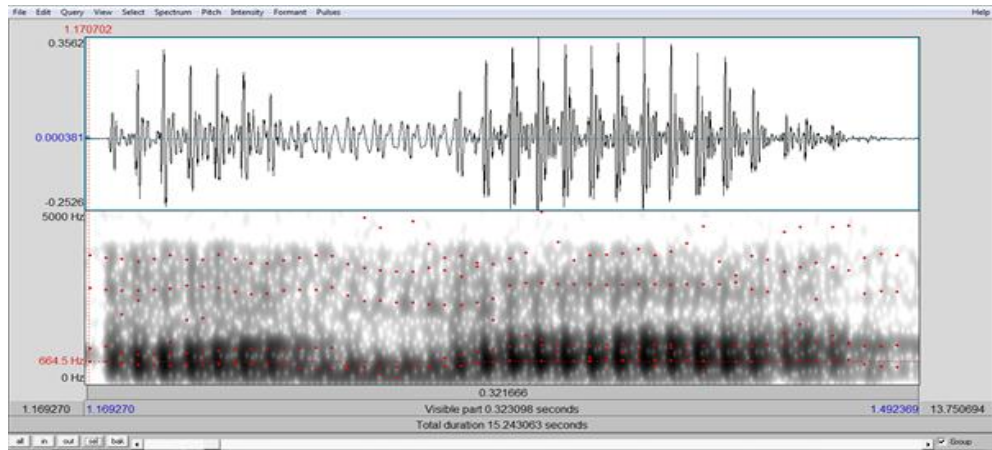
***The percent correct speaker identification score for /m/, /n/ and /n/ was 60, 70 and 60, respectively when network recordings were compared with network recordings using MFCC.*** The percent correct speaker identification scores decreased drastically when network recordings were compared with network recordings. This finding was consistent with several other studies done earlier. Rida (2014) reported drop in the speaker identification scores for the three nasal continuants in Hindi to 50%, 80% and 90% when network recordings were compared with network recordings. The author assumed that the network frequency bandwidth (900/1800 for Vodafone) would mask the characteristic of the nasals that would have helped for speaker identification in the direct versus direct recording condition. Correspondingly, in the present study the

network recordings were carried out through the Vodafone network. Hence, the low scores for speaker identification when network recordings were compared to network recordings can be attributed to the network frequency bandwidth (900/1800) for Vodafone that may have masked the characteristic of the nasals. Nithya (2015) reported significantly lower scores for speaker identification for network recordings. The percentage of speaker identification for the nasal /m/ was 83.5, /n/ was 65.8 and /ŋ/ was 68.3. Barinov, Koval, Ignatov and Stolbov (2010) conducted a study to examine the characteristics of speech transmitted over a mobile network. They stated that the non-linearity of the GSM (Global System for Mobile Communications) channel's frequency response in the 750-2000 Hz range might cause a change in the energy distribution and affect the 2<sup>nd</sup> and the 3<sup>rd</sup> formants (F2 and F3). The authors also reported a fall-off in the channel's frequency response at 3500 Hz which led to the shifting of the 4<sup>th</sup> formant (F4). As the nasal murmur is present below 400 Hz, it would have been lost due to the transmission characteristics of the mobile network. In the present study in direct recording, the first formant of /m/ was 388 Hz, that of /n/ was 371 Hz and for /ŋ/ it was 388 Hz. As the network frequency bandwidth for Vodafone is 900/1800, it would have definitely masked the first formant and further, no information is available beyond 1800 Hz as shown in figures 18 to 20. The poor scores in network recording can be attributed to the limited bandwidth of the network used in the present study.



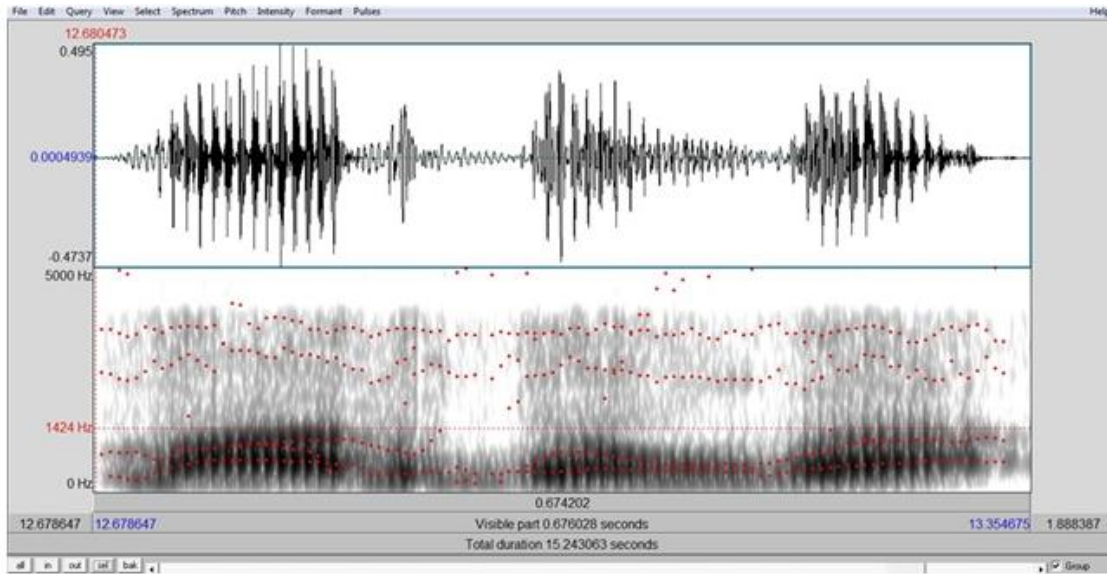
**Me** **re**

Figure 18: Spectrogram of the network recorded nasal /m/



**ho** **n** **a**

Figure 19: Spectrogram of the network recorded nasal /n/



Ma: r du n ga

Figure 20: Spectrogram of the network recorded nasal /n/

The speaker identification scores decreased when the direct recordings were compared with the network recordings with each sample treated as a different speaker. *The Highest Percent Identification (HPI) score on twenty randomizations for /m/, /n/ and /n/ was 50, 85 and 85, respectively when direct recordings were compared with network recordings using MFCC. The Lowest Percent Identification (LPI) score on twenty randomizations for /m/, /n/ and /n/ was 45, 70 and 70, respectively when direct recordings were compared with network recordings using MFCC.* This decreasing trend can be attributed to the reason that as the total number of speakers considered increase the more are the chances for error identification by the software. The scores for speaker identification declined when direct recordings were compared with network recordings due to the interference of the recording network characteristics. The GSM (Global System for Mobile Communications) is the pan-European cellular mobile standard. Speech signals are compressed before transmission by the speech coding algorithms that are a part of the GSM. The



algorithms reduce the number of bits in digital representation, at the same time maintaining acceptable quality. Consequently, this process may modify the speech signal and can have an influence on the speaker recognition performance (Barinov, Koval, Ignatov and Stolbov, 2010). The findings in the present study are consistent with the results of the previous studies. Rida (2014) reported a decrement in the speaker identification score when the live recordings were compared with the network recordings. The author reported speaker identification scores for the three nasal continuants in Hindi as 90%, 90%, and 30%.

***Overall, the results revealed that the velar nasal continuant /n/ had the best percent correct speaker identification in the current study.*** The velar nasal continuant in Urdu has a mid frequency spectra, the bilabial /m/ has a low frequency spectra and the alveolar /n/ has a high frequency spectra. The overall energy in nasal continuants is dampened due to the characteristics of the nasal cavity. The velar nasal continuant /n/ which had the highest percent correct identification may be attributed to its acoustic properties. During the production of the velar nasal /n/, the point at which the closure occurs in the vocal tract may depend largely on the following vowel. The first resonance occurs at approximately around 370 Hz. The first anti-resonance is above 3 kHz. Furthermore, the frequency of the first anti-resonance has little side branching, and the velar nasal /n/ had the lowest number of occurrence in the stimuli of the present study. It occurred overall four times in the 10 sentences used in the study. Rida (2014) reported that, in Hindi, the velar /n/ had the highest percent correct identification.

The results of the present study indicate a high benchmark for speaker identification using the nasal continuants in Urdu using MFCC in direct recording only. However,

while comparing the network recording with network recording, alveolar /n/ only can be considered and if one compares direct recording with network recording alveolar /n/ or velar /n/ can be considered. The benchmark obtained from the present study is as follows:

CONDITION	/m/	/n/	/nʰ/
Direct vs. Direct recording	70%	80%	100%
Network vs. Network recording	60%	70%	60%
Direct vs. Network recording - HPI	50%	85%	85%
Direct vs. Network recording - LPI	45%	70%	70%

The present study was limited to male participants. Future studies on a larger sample size are warranted in other Indian languages for better generalization of the findings.

## CHAPTER VI

### SUMMARY AND CONCLUSIONS

People are identified routinely by their voices in everyday life. People are recognized on a daily basis with their distinctive voices, over a radio, phone line, to name a few. In order to gain access to high security areas an individual's identity verification is an essential requirement. This requirement is typically met by an exclusive personal possession such as a key, a badge, or a password. The voice of an individual can be recorded while planning, committing or confessing to a crime. It can be used to directly incriminate the suspect in the act of committing the crime (Rose, 2002). "Forensic voice identification is a legal process to decide whether two or more recordings of speech are spoken by the same speaker" (Rose, 2002). A voice print is one of the means used to identify a person who has committed a crime and is valid as evidence in a court of law (Saitō & Nakata, 1985).

Researches in the past have used formant frequencies, fundamental frequencies, F0 contour, Linear Prediction Coefficients (Atal, 1974; Imperl, Kačič & Hovert, 1997), Cepstral Coefficients (Jakhar, 2009; Medha, 2010; Sreevidya, 2010) and Mel Frequency Cepstral Coefficients (Plumpe, Quatieri & Reynolds, 1999; Hasan, Jamil, Rabbani & Rahman, 2004; Chandrika, 2010; Mehra et al, 2010; Ramya, 2013; Rida, 2014) to identify speakers. The Cepstral Coefficients and Mel Frequency Coefficients have found to be the most accurate predictors of speaker identification. There are limited studies in the field of Forensic Speaker Identification to train experts on analysis. In order to provide adequate training to experts in this field to make them efficiently identify voices, it is important to establish benchmark in all languages for speaker identification. Thus, the aim of the present study was to establish Benchmark

for speaker identification for nasal continuants in Urdu using Mel-frequency Cepstral Coefficients (MFCC). Specifically, the objectives of the study were to provide benchmark for speaker identification in Urdu nasal continuants using, and to compare benchmarks in direct and network recording conditions.

Ten male participants between the age ranges of 20 to 40 years with Urdu as their native language for oral communication were included in the study. Commonly occurring forensically related Urdu meaningful words with nasal continuants – bilabial /m/, alveolar /n/ and velar /ŋ/ were selected. The nasal continuants were embedded in 3-4 word sentences in word - initial, - medial and - final positions to maintain the naturalness of speech. In total /m/, /n/, and /ŋ/ occurred 7, 8, and 4 times, respectively. Participants were specifically instructed to adopt a casual conversational style while reading out the sentences. Each card with one stimulus each was presented to the participant visually. They were instructed to utter the sentences thrice at an interval of 1 minute. They were instructed to speak under two conditions, directly into the recorder (direct) and through another mobile into the recording mobile phone (network). The participants read out the sentences which were recorded simultaneously on to the digital recorder and the network. Thus, the direct and the network recordings were carried out simultaneously. The recorder used in this study was an Olympus LS-10S PCM recorder (Olympus America Inc.). The network used for making the calls was Vodafone and the receiving network was Vodafone on a Lenovo mobile phone. The speech communicated at the receiving end were recorded and saved in the SD card of the mobile phone. Data analysis was carried out using SSL Workbench (Voice and Speech Systems. Bangalore, India). Euclidian Distance for the mobile and network derived MFCC were extracted. The diagonal matrix and a final percentage for correct speaker identification were obtained. A speaker was

presumed to be identified correctly when the Euclidian distance between the training and the test sample was the least. The percent correct identification was calculated using the following formula:

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

The results of the present study revealed several significant findings. The percent correct speaker identification score for /m/, /n/ and /n:/ was 70, 80 and 100, respectively when direct recordings were compared with direct recordings using MFCC. The percent correct speaker identification score for /m/, /n/ and /n:/ was 60, 70 and 60, respectively when network recordings were compared with network recordings using MFCC. The percent correct speaker identification scores decreased drastically when network recordings were compared with network recordings. The speaker identification scores decreased when the direct recordings were compared with the network recordings with each sample treated as a different speaker. The Highest Percent Identification (HPI) score on twenty randomizations for /m/, /n/ and /n./ was 50, 85 and 85, respectively when direct recordings were compared with network recordings using MFCC. The Lowest Percent Identification (LPI) score on twenty randomizations for /m/, /n/ and /n:/ was 45, 70 and 70, respectively when direct recordings were compared with network recordings using MFCC. This decreasing trend can be attributed to the reason that as the total number of speakers considered increase, the more are the chances for error identification by the software. The scores for speaker identification declined when direct recordings were compared with network recordings due to the interference of the recording network characteristics. The GSM (Global System for Mobile Communications) is the pan-European cellular mobile standard. Speech signals are compressed before

transmission by the speech coding algorithms that are a part of the GSM. The algorithms reduce the number of bits in digital representation, at the same time maintaining acceptable quality. Consequently, this process may modify the speech signal and can have an influence on the speaker recognition performance (Barinov, Koval, Ignatov and Stolbov, 2010).

Overall, the results revealed that the velar nasal continuant /ŋ/ had the best percent correct speaker identification in the current study. The velar nasal continuant in Urdu has a mid frequency spectra, the bilabial /m/ has a low frequency spectra and the alveolar /n/ has a high frequency spectra. The overall energy in nasal continuants is dampened due to the characteristics of the nasal cavity. The velar nasal continuant /ŋ/ which had the highest percent correct identification may be attributed to its acoustic properties. During the production of the velar nasal /ŋ/, the point at which the closure occurs in the vocal tract may depend largely on the following vowel. The first resonance occurs at approximately around 370 Hz. The first anti-resonance is above 3 kHz. Furthermore, the frequency of the first anti-resonance has little side branching, and the velar nasal /ŋ/ had the lowest number of occurrence in the stimuli of the present study. It occurred overall four times in the 10 sentences used in the study.

The results of the present study indicate a high benchmark for speaker identification using the nasal continuants in Urdu using MFCC in direct recording only. However, while comparing the network recording with network recording, alveolar /n/ only can be considered and if one compares direct recording with network recording alveolar /n/ or velar /ŋ/ can be considered. The benchmark obtained from the present study is as follows:

CONDITION	/m/	/n/	/nʹ/
Direct vs. Direct recording	70%	80%	100%
Network vs. Network recording	60%	70%	60%
Direct vs. Network recording - HPI	50%	85%	85%
Direct vs. Network recording - LPI	45%	70%	70%

The present study was limited to male participants. Future studies on a larger sample size are warranted in other Indian languages for better generalization of the findings.

## REFERENCES

- Anantapadmanabha, T. (2015). SSL Workbench. . *Voice and Speech systems* . Bangalore, India.
- Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, 52(6B), 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*, 55(6), 1304-1312.
- Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 64(4), 460-475.
- Barinov, A., Koval, S., Ignatov, P., & Stolbov, M. (2010, June). Channel compensation for forensic speaker identification using inverse processing. In *Audio Engineering Society Conference: 39th International Conference: Audio Forensics: Practices and Challenges*. Audio Engineering Society.
- Bennani, Y., & Gallinari, P. (1991). *A modular connectionist architecture for text-independent talker identification*. Université de Paris-Sud, Centre d'Orsay, Laboratoire de Recherche en Informatique.
- Boersma, P. & Weenik, D. (2016). *Praat: doing phonetics by computer [Computer program]*. Retrieved from <http://www.praat.org/>: <http://www.praat.org/>
- Bolt, R. H., Cooper, F. S., David Jr, E. E., Denes, P. B., Pickett, J. M., & Stevens, K. N. (1970). Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes. *The Journal of the Acoustical Society of America*, 47(2B), 597-612.
- Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, 40(6), 1441-1449.
- Chandrika, S. (2010). The influence of handsets and cellular networks on the performance of a speaker verification system. *Project of Post graduate Diploma in Forensic Speech Sciences and Technology submitted to University of Mysore* . Mysore, Karnataka, India.
- Chandrika. (2015). Benchmark for speaker identification for nasal continuants in Kannada using Mel Frequency Cepstral Coefficients in Kannada. *Project of Post graduate Diploma in forensic Speech Sciences and Technology submitted to University of Mysore*. Mysore, Karnataka, India.
- Delacy, R. (1998). *Hindi & Urdu Phrasebook*. Lonely Planet.
- Doddington, G. R. (1971). A Method or Speaker Verification. *The Journal of the Acoustical Society of America*, 49(1A), 139-139.



- Eatock, J. P., & Mason, J. S. (1994, April). A quantitative assessment of the relative speaker discriminating properties of phonemes. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on* (Vol. 1, pp. I-133). IEEE.
- Furui, S. (1978). Effects of Long-term Spectrum Variability on Speaker Recognition. In *Proc. Joint Meeting of ASA and ASJ*.
- Furui, S. (1994). An overview of speaker recognition technology. *Proc. ESCA Workshop on Automatic Speaker Recognition*, (pp. 1-8).
- Glenn, J. W., & Kleiner, N. (1968). Speaker identification based on nasal phonation. *The Journal of the Acoustical Society of America*, 43(2), 368-372.
- Hasan, M. R., Jamil, M., & Rahman, M. G. R. M. S. (2004). Speaker identification using mel frequency cepstral coefficients. *variations*, 1, 4.
- Hecker, M. H. (1971). *Speaker recognition: An interpretive survey of the literature*. Washington, DC: American Speech and Hearing Association.
- Higgins, A., Bahler, L., & Porter, J. (1991). Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1(2), 89-106.
- Hollien, H. (1974). Peculiar case of "voiceprints". *The Journal of the Acoustical Society of America*, 56(1), 210-213.
- Hollien, H. (1977). Vocal and speech patterns of depressive patients. *Folia Phon*, 29: 279 -291 .
- Hollien, H. F. (2002). *Forensic voice identification*. Academic Press.
- Hollien, H. M. (1982). Aural versus visual identifications resulting from a simulated crime. *Journal of Forensic Sciences*, 28, 208-221.
- Hollien, H., & Rosenberg, A. E. (1991). The Acoustics of Crime: The New Science of Forensic Phonetics. *The Journal of the Acoustical Society of America*, 90(3), 1703-1704.
- Hollien, H., & Schwartz, R. (2000). Aural-perceptual speaker identification: problems with noncontemporary samples. *Forensic linguistics*, 7, 199-211.
- Hollien, H., & Schwartz, R. (2001). Speaker identification utilizing noncontemporary speech. *Journal of Forensic Science*, 46(1), 63-67.
- Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*.
- Imperl, B., Kačič, Z., & Horvat, B. (1997). A study of harmonic features for the speaker recognition. *Speech Communication*, 22(4), 385-402.
- Ingemann, F. (1968). Identification of the speaker's sex from voiceless fricatives. *The Journal of the Acoustical Society of America*, 44(4), 1142-1144.

- Jakhar, S. (2009). Benchmark for speaker identification using Cepstrum. . *Project of Post graduate Diploma in Forensic Speech Sciences and Technology submitted to University of Mysore . Mysore., Karnataka.*
- Jyotsna. (2011). Speaker identification using Cepstral Coefficients and Mel Frequency Cepstral Coefficients in Malayalam nasal Co-articulation. *Project of Post graduate Diploma in Forensic Speech Sciences and Technology submitted to University of Mysore, .*
- Kent, R., & Read, C. (2002). *Acoustic Analysis of Speech*, 2nd edn (San Diego, California: Singular, Thomas Learning).
- Kersta, L. G. (1962). Voiceprint identification. *The Journal of the Acoustical Society of America*, 34(5), 725-725.
- Kinnunen, T. (2003). Spectral features for automatic text-independent speaker recognition. *Licentiate's Thesis, University of Joensuu.–2003.*
- Koenig, B. E. (1986). Spectrographic voice identification: a forensic survey. *The Journal of the Acoustical Society of America*, 79(6), 2088-2090.
- Kumar, P., & Rao, P. (2004). A study of frequency-scale warping for speaker recognition. *Proc. NCC 2004*, 203-207.
- Künzel, H. J. (1987). *Sprechererkennung: Grundzüge forensischer Sprachverarbeitung*. Kriminalistik-Verlag.
- Labov, W., & Harris, W. A. (1994). Addressing social issues through linguistic evidence. *Language and the Law*, 265-305.
- Lam, D. (1999). District Court of New South Wales, 99-11-0711.
- Lei, H., & Gonzalo, E. L. (2009). Importance of nasality measures for speaker recognition data selection and performance prediction. In *INTERSPEECH* (pp. 888-891).
- Luck, J. E. (1969). Automatic speaker verification using cepstral measurements. *The Journal of the Acoustical Society of America*, 46(4B), 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= Journal of biomedical engineering= Shengwu yixue gongchengxue zazhi*, 23(4), 882-886.
- McDermott, M. C., Owen, T., & McDermott, F. M. (1996). Voice Identification: The Aural/Spectrographic Method. *Owl Investigations Web Site*, [http://www.owlinvestigations.com/forensic\\_artic\\_les/aural\\_spectrographic/fulltext.html](http://www.owlinvestigations.com/forensic_artic_les/aural_spectrographic/fulltext.html).
- McGehee, F. (1937). The reliability of the identification of the human voice. *The Journal of General Psychology*, 17(2), 249-271.
- Medha, S. (2010). Benchmark for speaker identification by Cepstrum measurement using text-independent data. *Project of Post graduate Diploma in Forensic Speech Sciences and Technology submitted to University of Mysore, Mysore.*

- Mehra, A., Kumawat, M., Ranjan, R., Pandey, B., Ranjan, S., Shukla, A., & Tiwari, R. (2010, May). Expert system for speaker identification using lip features with PCA. In *Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on* (pp. 1-4). IEEE.
- Meltzer, D. & Lehiste, I. (1972). Vowel and Speaker Identification in Natural and Synthetic Speech. *The Journal of the Acoustical Society of America*, 51: S131 (A).
- Milner, B.P. (2003). *Speech and Language Processing Lecture Notes*, University of East Anglia, UK.
- Miyajima, C., Watanabe, H., Tokuda, K., Kitamura, T., & Katagiri, S. (2001). A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction. *Speech Communication*, 35(3), 203-218.
- Naik, J. (1994, September). Field trial of a speaker verification service for caller identity verification in the telephone network. In *Interactive Voice Technology for Telecommunications Applications, 1994. Second IEEE Workshop on* (pp. 125-128). IEEE.
- Naik, J. M., & Doddington, G. R. (1987, April). Evaluation of a high performance speaker verification system for access Control. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*. (Vol. 12, pp. 2392-2395). IEEE.
- Nithya (2015). Benchmark for speaker identification using Tamil nasal continuants in live recording and mobile network recording. *Project of Post graduate Diploma in Forensic Speech Sciences and Technology submitted to University of Mysore, Mysore*.
- Nolan, F. (1983). The phonetic bases of speaker recognition. Cambridge studies in speech science and communication.
- Oglesby, J. & Mason, J.S. (1990). Optimization of Neural Models for Speaker Identification. In *Proc. of IEEE Intl. Conf. Acoust. Speech and Signal Proc. (ICASSP '90)*.
- Orman, Ö. D. (2000). *Frequency analysis of speaker identification performance* (Doctoral dissertation, Boğaziçi University).
- 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*.
- Patel, K., & Prasad, R. K. (2013). Speech recognition and verification using MFCC & VQ. *International Journal of Emerging Science and Engineering*, 1(7), 33-7.
- Plumpe, M. D., Quatieri, T. F., & Reynolds, D. A. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *Speech and Audio Processing, IEEE Transactions on*, 7(5), 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(3), 403-406.
- Potter, R. K., Kopp, G. A., & Grey, H. G. (1947). Visible Speech. *D. Van Nostrand Co., Inc., N.Y.*
- Prasanna, L. (2011). Benchmark for nasal continuants in Telugu for speaker identification. *Unpublished dissertation of Master of Science in Speech Language Pathology submitted to University of Mysore, Mysore.*
- Pruzansky, S. (1963). Pattern-Matching Procedure for Automatic Talker Recognition. *The Journal of the Acoustical Society of America*, 35(3), 354-358.
- Rabiner, L., & Juang, B. H. (1993). Fundamentals of speech recognition.
- Rajsekhar, A. (2008). Real time speaker recognition using MFCC and VQ. *Unpublished dissertation of Master of Technology in Electronics and Communication Engineering submitted to National Institute of Technology, Rourkela.*
- 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.*
- Ramya, B.M. (2013). Bench mark for speaker identification under electronic vocal disguise using Mel Frequency Cepstral Coefficients. *Unpublished dissertation of Master of Science in Speech Language Pathology submitted to University of Mysore, Mysore.*
- Reich, A. R. (1975). *Certain effects of selected vocal disguises upon spectrographic speaker identification.*
- Reich, A. R. (1981). Detecting the presence of vocal disguise in the male voice. *The Journal of the Acoustical Society of America*, 69(5), 1458-1461.
- Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America*, 66(4), 1023-1028.
- 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(4), 919-925.
- Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech communication*, 17(1), 91-108.
- Rida, Z, A. (2014). Benchmarks for speaker identification using nasal continuants in Hindi in direct mobile and network recording. *Unpublished dissertation of Master of Science in Speech Language Pathology submitted to University of Mysore, Mysore.*
- Rose, P. H. I. L. I. P. (2002). Forensic Speaker Identification.

- Rose, R. C., & Reynolds, D. A. (1990, April). Text independent speaker identification using automatic acoustic segmentation. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on* (pp. 293-296). IEEE.
- Rosenberg, A. E. (1973). Listener performance in speaker verification tasks. *Audio and Electroacoustics, IEEE Transactions on*, 21(3), 221-225.
- Rudasi, L., & Zahorian, S. A. (1991, April). Text-independent talker identification with neural networks. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on* (pp. 389-392). IEEE.
- Saferstein, R. (2013). *Criminalistics*. Pearson Education.
- Saitō, S., & Nakata, K. (1985). *Fundamentals of speech signal processing*. Academic Pr.
- Sambur, M. R. (1975). Selection of acoustic features for speaker identification. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 23(2), 176-182.
- Schwartz, M. & Ingemann. (1968). Identification of speaker sex from isolated, voiceless fricatives. *The Journal of the Acoustical Society of America*, 44(4), 1142-1144.
- Schwartz, M. F., & Rine, H. E. (1968). Identification of speaker sex from isolated, whispered vowels. *The Journal of the Acoustical Society of America*, 44(6), 1736-1737.
- Shah, A. M. (2002). Urdu nasal consonants and their phonological behaviour. *Center of research in Urdu language processing (CRULP)*, 133-140.
- Sreevidya, M. S. (2010). Speaker identification using Cepstrum in Kannada language. *Project of Post Graduate Diploma in Forensic Speech Sciences and Technology submitted to University of Mysore, Mysore*.
- Su, L. S., Li, K. P., & Fu, K. S. (1974). Identification of speakers by use of nasal coarticulation. *The Journal of the Acoustical Society of America*, 56(6), 1876-1883.
- Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, 1(1), 19-22.
- Tosi, O., Oyer, H., Lashbrook, W., Pedrey, C., Nicol, J., & Nash, E. (1972). Experiment on voice identification. *The Journal of the Acoustical Society of America*, 51(6B), 2030-2043.
- Wang, L., Ohtsuka, S., & Nakagawa, S. (2009, April). High improvement of speaker identification and verification by combining MFCC and phase information. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (pp. 4529-4532). IEEE.

- Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, 51(6B), 2044-2056.
- Young, M. A., & Campbell, R. A. (1967). Effects of context on talker identification. *The Journal of the Acoustical Society of America*, 42(6), 1250-1254.