BENCHMARK FOR SPEAKER IDENTIFICATION USING NASAL CONTINUANTS IN KANNADA IN DIRECT MOBILE AND NETWORK RECORDING

Chandrika G

PGDFSST Register No. 14FST002

A Project Work Submitted in Part fulfillment of third term of Post Graduate Diploma in Forensic Speech Science and Technology

University of Mysore, Mysuru.

ALL INDIA INSTITUTE OF SPEECH AND HEARING

MANASAGANGOTHRI, MYSORE-570 006

JULY-2015

CERTIFICATE

This is to certify that this project work entitled **"*Benchmark for Speaker Identification using Nasal continuants in Kannada in Direct mobile and Network Recording*"** is the bonafide work submitted in part fulfillment for third term of the Post Graduate Diploma in Forensic Speech Science and Techonology of the student (Register No. 14FST002). This has been carried out under the guidance of Dr. S. R. Savitrhi, Director of this institute and has not been submitted earlier to any other University for the award of any other Diploma or Degree.

Mysore
July, 2015

Dr. S. R. Savithri
Director
All India Institute of Speech and Hearing
Manasagangothri
Mysore - 570 006

CERTIFICATE

This is to certify that this project work entitled **"*Benchmark for Speaker Identification using Nasal continuants in Kannada in Direct mobile and Network Recording"*** has been prepared under my supervision and guidance. It is also certified that this has not been submitted earlier in other University for the award of any Diploma or Degree.

Mysore

May, 2014

Dr. S. R. Savithri
Director
All India Institute of Speech and Hearing
Manasagangothri, Mysore -570 006

DECLARATION

This is to certify that this project work entitled **"*Benchmark for Speaker Identification using Nasal continuants in Kannada in Direct mobile and Network Recording*"** is the result of my own study under the guidance of Dr. S. R. Savitrhi, 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

July, 2015 PGDFSST Register No.14FST002

Acknowledgements

I am greatly honored to express my respect and gratitude to Dr. S.R Savithri, Director, All India Institute of Speech and Hearing and specially to my guide "Savithri Ma'am" for accepting me and allowing to carry out project under her able guidance. You are the only inspiration that evoked an aspiration to join for this course. Your knowledge, support and guidance in the project has enlightened me. You are the synonym for complete human being and I learnt beyond text which is very important in life. Your simplicity and honesty, totally you are a role model ma'am. Thank you for everything ma'am.

I would like to remember the help provided by my present Director, FSL, Bengaluru, & B. Dayananda former in-charge Director, FSL, Bengaluru, V.G. Nayak, Deputy Director, B.C. Ravindra, Assistant Director, for their constant support and encouragement throughout the course period, I express my sincere thanks to both technical and non technical staff of FSL, Bengaluru.

I extend my special thanks to Dr. Sreedevi , H.O.D department of Speech Language Sciences for her cooperation. I also wish to express my gratitude to my lecturers Dr. Santhosh, Dr Hema, Dr Jayakumar and Rajasudhakar for their unwavering support, encouragement and sharing their experience throughout the course. I also thank all technical and non technical staff of Speech Language Sciences for their constant support till the end.

I grab this opportunity to register my endless thanks to my mother who is very courageous and my father who supported extensively for my studies. Amma my life is your message, Ma you are great. You are always right ma your right guidance in

right time making me alright every time. Love you Ma. Anna thank you very much for your unwavering faith towards me till today.

My very special gratitude to my sister Kalavathi. G , my jeeju Dr.S.T.Ramachandra and my nephew cute pearl Pratham.K.R for their love, affection, caring and hospitality during my journey to AIISH. Thank you Kala for your ironing words and bava for your courage and encouragement in hard situations. I also extend my thanks to my first sister Pallavi.G for her love and caring and Shashidhar.S . My hearty blessings to my nephews Preetham and Dasara (Maana) and thank them for their joyful pranks. Finally I wish to thank my little sis Vidya.G for her affection and support.

I am happily acknowledge my gratitude to my mother-in law Meena Kumari for her true affection and caring and my father in law C.M.Madaiah for conditional love and caring. I wish to express my thanks to my co sister Sunitha and brother in law Pravin Kumar.M . I also extend my special thanks to my second co-sister Shalini and brother in law Arun Kumar.M for their constant support. I also thank my sister in law Chithra.M and brother in law Eeresh for their affection and support. Finally I thank all junior members of my family Inchara, Mokhsith, Dishitha and junior Shalini.

Success is a dream without the support of companion. Mere Jeevan saathi Kiran Kumar.M, whose love and patience always add new color to my life and supported me in every step of my life. You are my best friend. My two daughters, Hithishi and Stuti are the two stars gifted in my life, Love you kids for your cooperation, patience and support. My this whole course is dedicated to my kids. I owe my exclusive blessings and wishes to my kids for ever and ever.

I thank my class mates Arjun, Nithya and Suman for cherishing my college days memories. One of my favorite spot in AIISH is library. My acknowledgement is incomplete if I miss this. I must specially thankful to H.O.D Library and all library staff for their service.

Finally I must thank the creator almighty for everything. I thank one and all who supported me directly or indirectly.

Table of contents

List of Tables

List of Figures

CHAPTER I

INTRODUCTION

"*If the law has made you a witness, remain a man of science, you have no victim to avenge, no guilty or innocent person to convict or save – you must bear testimony within the limits of science"*

Brouardel, P.C.H.

Any science, used for the purpose of the law is a Forensic Science. Forensic science is science used in public, in a court or in the justice system. Forensic expert is a scientist, who applies his scientific knowledge to assist juries, attorneys and judges in understanding science (Steve Cain, 2015).Various physical evidences fingerprints, foot prints, blood, semen, saliva, skin, nail, hair fibers, bone which are originated from human beings are considered as very good evidence in the court of law.

A voice print is one means used to identify a person who has committed crime and is valid as evidence in a court of law (Shuzo Saito & Kazuo Nakata, 1985). Fingerprints are static images that don't change unlike some damage is done to the fingerprint ridge detail but, voiceprint has dynamic qualities such as pitch of voice varies with respect to time (Steve Cain, 2015). Like finger prints, voiceprints also a helpful way of identifying a criminal. "Forensic voice identification is a legal process to decide whether two or more recordings of speech are spoken by the same speaker" (Rose, 2002).

The most natural way to communicate is through speech. People all over the world, irrespective of language, make use of their larynx to produce voice. Voice is an acoustic signal produced by the modification of air at the level of vocal folds. Voice is also interchangeably used with speech which is produced by the modification of air at

the level of vocal folds and the articulators. This acoustic signal travels in the air, and is heard and interpreted in a different manner by each individual as our hearing mechanism differs from person to person. Therefore, the auditory system can be considered to be one of great precision as well as one which is quite deceptive in function (Hollien, 1990)

However, voice is more than just a string of sounds. It is also a media through which we identify other humans known to us like members of our family, friends, popular figures etc. This information is retrieved from the tone of the voice, rate of speaking, style of speaking etc., which is additional information apart from the intended linguistic message. Other characteristics of the individual like age, gender, language, emotional state and so on can also be identified by listening to their voice even if they are unfamiliar to us.

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 Science is well accomplished with various identifying features namely Finger print, Palm print, Gait Pattern, Handwritings, Signatures, Iris, Retina, DNA. Among these Voice is also one of the very useful features in the identification of a person. Handwritings show both inter- and intra-person variability similarly speech patterns also shows more variability with respect to both inter and intra person.

“There has been an increase in the crime rate at a world-wide scale. A tendency to disguise ones voice is a popular method for perpetrators to avoid capture by concealing their identities specially while making threatening phone calls,

kidnapping, extortion or emergency police help calls. The deliberate action of the speaker to conceal or falsify their identity is referred to as ***vocal disguise***. 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. Recent times have seen an exponential increase in the use of mobile phones. It was only a matter of time before these were also used in committing crimes. When a crime is committed through telecommunication, voice is the only evidence available for analysis" (Ramya, 2013). Therefore expert opinion is always being sought to establish whether two or more recordings are from the same speaker. This has brought the field of Forensic Speaker Identification into limelight. Rose (1992) states that speaker recognition can be either speaker identification, or speaker verification. ***Speaker identification*** refers to the identification of a particular speaker from a group of unknown speakers. It requires 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 be ***text- dependent and text- independent***. In the former the same text should be present in the test and training samples; in the latter, voice characteristics are analyzed from the sample recording irrespective of the linguistic content of the recording (Rabiner, 1993). However the choice of the technique is

application-specific. But, both the tasks involve two processes, - feature extraction and feature matching.

The forensic scientist may encounter with another problem namely, system distortions and speaker distortion. ***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, 1990). The correct speaker identification is degraded by background noise, different transmission channels, emotional states etc. If the disguises are more deliberate, then identification becomes more difficult (Ramya, 2013). Therefore it is necessary to study the effect of disguise on speaker identification. Especially if the speaker identification will focus on speech sounds with less association with the oral cavity as the perpetrators focus on changing the characteristics of this cavity to disguise voice. The nasal cavity is a relatively tougher choice when it comes to manipulation (Lei, Lopez-Gonzalo, 2009)

Researchers, in the past, have used formant frequencies, fundamental frequency, F0 contour, Linear Prediction coefficients (Atal, 1974; Imperl, Kacic & Hovert, 1997), Cepstral Coefficients (Jakkhar, 2009; Medha, 2010; Sreevidya, 2010) and Mel Frequency Cepstral coefficients (Plumpe, Quateri & Reynolds, 1999; Hassan, Jamil, Rabbani & Rahman, 2004; Chandrika, 2010; Tiwari et. al., 2010) to identify speaker.

However, the Cepstral Coefficients and the Mel Frequency Cepstral Coefficients have been found to be more effective in speaker identification compared to other features.

Atal (1974) examined various parameters using linear prediction model for their effectiveness for automatic recognition of speakers from their voices. Results revealed ***cepstrum*** 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. Using the same speech data, verification accuracy was approximately 83% for duration of 50 ms increasing to 95% for duration of 1sec.

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

Some experiments were conducted by Reich, Moll, & Curtis (1976) to investigate the effect of vocal disguises upon speaker identification. The results 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). The mean performance level (56.67% correct) on the undisguised task was considerably poorer than the data for similar experimental conditions (approximately 80%) (Tosi, Oyer, Lashbrook, Pedrey, Nichol & Nash, 1972). ***In general, results of this experiment show that nasal and slow rate were the least effective disguises***, while free disguise was the most effective on the spectrographic speaker identification. The exclusion of low confidence decisions produced significantly higher correct percentages. ***It was also found that stimulus words containing nasal phonemes (i.e., me, on, and) were considered quite useful for spectrographic speaker identification.*** Reich et al, (1976) found that the inclusion

of disguised speech samples in the spectrographic matching tasks significantly interfered with speaker identification performance and had a significant effect on the types of errors made by the examiners. The errors of false identification increased, accompanied by a proportional decrease in the errors of false elimination.

Reich, & Duke (1979) describe another experiment involving the effects of selected vocal disguises upon speaker identification by listening. The reduction in speaker identification performance by vocal disguise in naïve listeners was 22.0% (slow rate) to 32.9% (nasal) and in sophisticated listeners it was 11.3% (hoarse) to 20.3% (nasal). In general, ***results show that nasal disguise (naïve and sophisticated listeners) was the most effectiv***e, while slow rate disguise (naïve listeners) and hoarse disguise (sophisticated listeners) were the least effective disguises on the speaker identification by listening. Further, ***nasal disguise was the most effective disguise in speaker identification by listening experiment (Reich et al., 1979). In contrast, the nasal disguise was the least effective in a previous spectrographic matching experiment (Reich et al., 1976). Similarly, the power spectra of nasal consonants (Glenn & Kleiner, 1968) and coarticulated nasal spectra (Su; Li and Fu, 1974) seem to provide strong cues for the machine matching of speakers***. Thus, the nasal phonemes have been identified as being more reliable as a speaker cue because nasal cavity is both speaker specific and fixed so as its volume and shape cannot be changed.

Glenn & Kleiner (1968) hypothesized that each of the speakers produce a unique and identifiable ***power spectrum during nasal phon*ation** in recognition experiments. The result obtained showed 97% of identification accuracy with /n/ nasal sound for the entire experiment.

Kinnunen (2003) indicated that the ***Mel-frequency Cepstral Coefficients*** (MFCC) is the most evident example of a feature set that is extensively used in speaker recognition. In using MFCC feature extractor, one makes an assumption that the human hearing mechanism is the optimal speaker recognizer. 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 ***MFCC***s 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% Net work vs. network recording.

Psychophysical studies of the frequency resolving power of the human ear has motivated modeling the non-linear sensitivity of human ear to different frequencies. MFCC's are based on the known variation of the human ears critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies. In addition, MFCC's are shown to be less susceptible to the variation of the speaker's voice and surrounding environment. Initially, Fast Fourier

Transformation (FFT) of a speech sample is extracted which is converted to Mel frequency. Cepstral coefficients are extracted on Mel frequencies.

It is evident from the review that MFCCs is, perhaps, the best parameter for speaker identification. Also, nasal continuants may be the most suitable, among speech sounds, for speaker identification. However, 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 analysis becomes a must. ***In this context, the present study was planned.*** The aim of the study was to establish ***Benchmark for speaker identification for nasal continuants in Kannada using Mel Frequency Cepstral Coefficients in Kannada*** [***"***Kannada is a widely used language and one amongst the most spoken languages within the world. Those speak Kannada by birth are called as Kannaḍigas and there are roughly forty million people use this language for regular purpose and also the administrative and official language of the Karnataka state" (retrieved from http://en.wikipedia.org/ wiki/Kannada_language)]. In Mysore dialect of spoken Kannada the frequency of occurrence of bilabial /m/ is 2.76%, dental /n/ is 7.59% and retroflex /n̪/ is 0.29% (Sreedevi-2013).

The objectives of the study were two-fold and as follows:

1. to find out the Mel frequency Cepstral Coefficients for Kannada nasal continuants in direct and mobile recording, thus providing benchmark for speaker identification, and

2. to compare the MFCCs across three age groups of 20≤30 years, 30≤40 years, and 40≤50 years.

CHAPTER II

REVIEW OF LITERATURE

When you have eliminated the impossible, whatever remains, however improbable, must be the truth

Sir Arthur Conan Doyle

Speech or speech waveform is unique individual trait. Individual speech waveforms are unique because individual physical dimensions of vocal organs and their physical characteristics are different. Factors causing individuality of a voice are vocal-tract length and vocal cord vibration, frequency and waveform. Characteristics of these factors are observed physically as high or low formant frequencies, wide or narrow bandwidths, high or low average pitch frequency and variation of the slope and curvatures of the spectrum envelope. These are used for speaker recognition as a set of features which are relatively independent of phonemic content of a word or phrase. These can be named as individualities of speech (Shuzo Saito & Kazuo Nakata, 1985).

Speaker Recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech signal. This process makes it possible to use the speaker's voice to verify his/her identity and thereby control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers etc. Speaker recognition can be classified into *verification* and *identification.* ***Speaker identification*** refers to the identification of a particular speaker from a group of unknown speakers.

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.

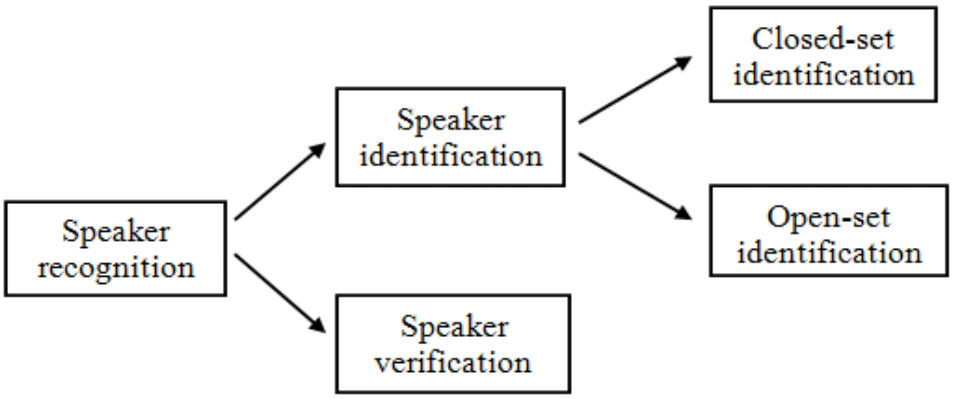


Figure 1: Schematic diagram of a speaker recognition system.

Forensic Speaker Identification involves verifying a speaker from speech recorded under less than ideal conditions typical in forensics. A comparison may have to be made of a disguised voice sample recorded over a telephone channel with the voice sample recorded under laboratory conditions. In forensic applications, it is common to first perform a speaker identification process to create a list of "best matches" and then perform a series of verification processes to arrive at a conclusive match.

Forensic speaker identification can often be classified as a kind of speaker recognition. The task is to compare the sample from the unknown speaker with the known set of samples, and determine whether it was produced by any of the known speaker (Nolan, 1983, Nolan, 1997). The objective of speaker identification is merely not, identification: 'to identify an unknown voice as one or none of a set of known voice' (Naik, 1994). The kind of activity covered by term speaker recognition is conceptually straight forward, and definitions abound. Hecker (1971) defines that

speaker recognition is “any decision-making process that uses the speaker-dependent features of the speech signal,” and Atal (1976) offers the formulation “any decision-making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance.”

Speaker recognition can be ***text- dependent and text- independent***. In the ***text dependent*** the same text should be present in the test and training samples. In the ***text-independent***, voice characteristics are analyzed from the sample recording irrespective of the linguistic content of the recording (Rabiner, 1993). However the choice of the technique is application-specific. But, both the tasks involve two processes, - feature extraction and feature matching.

Speech samples are mainly of two types **–Contemporary and Non-contemporary.** Speech samples recorded at same session in case of **Contemporary** and at different intervals of time (i.e. after hours, days or may be years) in **Non-contemporary** (Tosi,Oyer, Lashbrook, Pendrey, Nicol and Nash, 1972)

In speaker identification the decision is very simple either it is correctly identified or not. Whereas in speaker verification there were four kinds of decisions. (Thevenaz & Hugli, 1995)It may be correct acceptance, false acceptance, correct rejection and false rejection.

Most often encountered problem in forensic speaker identification is distortions. They are of two types, system distortion and channel distortion. It becomes very difficult to identify a speaker by his/her voice, especially when they are talking in an environment which distorts or masks their utterances (channel distortions) or when they are excited or stressed (speech distortions). The distortions are broadly classified

into two types, System distortion and Speaker distortion. ***System distortion*** includes several kinds of signal degradation. One is reduced frequency response, i.e., the signal pass band can be limited when some one talks over a telephone line or mobile phone, poor quality tape recorders are used to 'store' the utterances and / or microphones of limited capability are employed. In these cases, the important information about the talker is lost and these elements are not usually retrievable. Such limited signal pass band can reduce the number of helpful speaker specific acoustic factors. Second, noise can create a particularly debilitating type of system distortion as it tends to make the talker`s voice and, therefore, can obscure elements needed for identification. Examples of noise included those created by wind, motors, fans, automobile movement and clothing friction. The noise itself may be intermittent or steady state saw tooth or thermal and so on. Third, any kind of frequency or harmonic distortion can also make the task of identification more difficult. Examples include intermittent short circuits, variable frequency response, and harmonic distortion and so on. ***Speaker distortion*** is related to the speaker. The speaker themselves can be the source of many types of distortions. Fear, anxiety or stress like emotion can occur when the perpetrator is speaking during the commission of crime. They often will degrade identification as the speech shifts triggered by these emotions can markedly changed one or more the parameters within the speech signal. The effects of ingested drugs or alcohol; and even a temporary health state such as a cold can affect the speech. The suspect may sometimes attempt to disguise their voice (Holien, 1990). All those affect the speaker identification process horrendously. If the disguises are more deliberate, then the identification becomes more difficult (Ramya, 2013).

Speaker identification method is classified into three general categories as per the interpretive survey made by Hecker in 1971.

(1) Aural perceptual or by listening (subjective method)

(2) Spectrographic or Visual examination of spectrograms (subjective method)

(3) Semi automatic or Automatic using machines (objective method)

Of these approaches, the third method (semi automatic and automatic) appears to be the most promising for the future, primarily because (1) specific parameters within the speech signal can be selected and analyzed serially or simultaneously, (2) the selected vectors may be used in various combinations, and (3) subjective analysis by human is eliminated.

1. Aural Perceptual method

The speaker is identified here by merely listening to the voice. Hecker (1971) reported that speaker recognition by listening appears to be the most accurate and reliable method at that time. It is evident that the identification accuracy reduces as a function of time. In a very important historical Charles Lindberg's child kidnap case McGhee carried out research to assess the ability to identify the kidnapper. Speaker read the passage standing behind the screen, fifteen listeners heard the voice. Second time they heard the voice where there were five unknown foil speakers. In this listener had to write the number of the speaker they thought they had heard originally. The same procedure was repeated after 1, 2 and 3 days, 1,2, and 3 weeks, 1,3,5 months. The speaker identification score for 1 day-83%, 2 weeks-68%, 3 months-35% and 5 months-13%.

The study made by Bricker and Pruzansky (1966) pertaining to speaker identification by aural perceptual method at various intervals of time. Ten male talkers but not familiar and sixteen familiar listeners were selected for the study. The result of the present study shows improvements in the identification score. They have got 98% score for day one and 56% score for second day.

Thompson (1985) experimented in the speaker identification using voice line up. He used male voices in a six-voice line up. In this study the listeners has to rate each voice as to whether it was the voice they had heard 1 week previously. They have to decide is the voice heard previously was not in the line up or that they were not sure whether it was in the line up. The experimental result shows that correct identifications score was 62.1%, incorrect identifications score was 22.1%, and 15.8% “not in line up" or "not sure if in line up" response.

Further Yarmey and Mattys (1992) attempted to study the speaker identification scores in the telephonic speech samples. Results showed that for 1 week there was no significant reduction in speaker identification.

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

Speaker identification by aural perceptual method is worth using in ideal situations like if the speakers are known, of unique voice quality, samples can be heard several times, large speech sample, speech should be under natural speaking

characteristics, extensive training, speech may naturally good or bad, the accents and dialects are also more advantageous.

The major drawbacks in aural perceptual method are if the speaker is unknown, numberof talkers are greater, degraded speech samples due to channel and system distortions, disguised speech samples, talkers are from same family and non contemporary speech samples.

Aural perceptual method of Speaker identification is purely a subjective method. The results obtained using only one method is insufficient in drawing final conclusions in criminal justice system. Hence, more accurate methods having different approach is required in the forensic speaker identification to achieve 100% score .

2. Spectrographic Method

Bell Telephone laboratory scientists Potter, Grey and Kopp developed an instrument called Sonograph (Sono means sound and graph means visual display) in 1941 while studying speech signals related to communication services which was used during World War II to identify persons for intelligence purposes the voice broad cast by German Millitary communications. Lawrence Kersta a Bell System Engineer worked with this voice spectrograph (Sonograph) and observed that “voice spectrograms” renamed by Kersta 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 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. (Richard Safferstien, 2001).

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

However, similar results are not obtained by other researches. The correct identification scores reported by Kersta are outstandingly high, 99%-100%, for short words spoken either in isolation or in context, as compared to (a) 81%-87%, for short words spoken in isolation, reported by Bricker 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 Sumby (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 is 37.3%, and word in isolation is 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.

Stevens et al. (1968) compared aural with the visual examination of spectrograms using a set of eight talkers and a series of identification tests was carried out. The average error rate for listening is 6% and for visual is 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 also 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 coming into use in criminology, but the validity of this method is still in question.

Enormous complete study (Tosi et al., 1972) were published in which attempts were made to more closely imitate law enforcement conditions, but only spectral comparisons were made. A two-year experiment on voice identification through visual inspection of spectrograms was performed with the twofold goal of checking Kersta's (1962) claims in this matter and testing models including variables related to forensic tasks. The 250 speakers used in this experiment were randomly selected from a homogeneous population of 25000 males speaking general American English, all students at Michigan State University. A total of 34996 experimental trials of identification were performed by 29 trained examiners. Each trial involved 10 to 40 known voices, in various conditions: With

closed and open trials, contemporary and non-contemporary spectrograms, nine or six clue words spoken in isolation, in a fixed context and in a random context, etc.

The examiners were forced to reach a positive decision (identification or elimination) in each instance, taking an average time of 15 minutes. Their decisions were based solely on inspection of spectrograms; listening to the identification by voices was excluded from this experiment. The examiners graded their self-confidence in their judgments on a 4-point scale (1 and 2, uncertain; 3 and 4, certain). Results of this experiment confirmed Kersta's experimental data, which involved only closed trials of contemporary spectrograms and clue words spoken in isolation. Experimental trials of this study, correlated with forensic models (open trials, fixed and random contexts, non-contemporary spectrograms), yielded an error of approximately 6% false identifications and approximately 13% false eliminations.

The examiners judged approximately 60% of their wrong answers and 20% of their right answers as "uncertain." This suggests that if the examiners had been able to express no opinion when in doubt, only 74% of the total number of tasks would have had a positive answer, with approximately 2% errors of false identification and 5% errors of false elimination. Main differences of conditions that could exist between models and real cases are as follows:

(1) Population of known voices: In forensic cases, the catalog of known voices could theoretically include millions of samples. In the present practical situations that police must handle. In these cases the catalog of known voices is open, true, but limited to a few suspected persons. Therefore, it seems reasonable to disregard

size of the population of known voices as a differential characteristic that could hamper extrapolation of results from the present experiment to real cases.

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

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

(4) Availability of clues: In the experimental models of this study, only spectrograms of nine or six clue words were available to the examiners for visual inspection. Rather, a professional examiner is entitled to request as many samples as he deems necessary to reach a positive conclusion. In real forensic cases the

professional examiner must necessarily listen first to the unknown and known voices while processing the spectrograms for visual comparison. A combination of methods of voice recognition by listening and by visual enhances the accuracy of voice identifications.

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

Hazen (1973) investigated how well the file card system of voiceprint identification reported by Kersta fulfils its purpose of minimising the effects of contextually caused spectral variations and how well it serves as either an identification or population reduction tool. He reported that for reduced population, error rates were higher for closed tests (12.86% and 57.14%) than for open tests (11.91% and 52.38%), but were almost five times as great for the different context condition (57.14% and 52.38%) than for the same context condition (12.86% and 11.91%). Hollien (1974) comments on spectrographic speaker identification, it now appears that the controversy about "voiceprints" is doing the judicial system and the relevant scientific community a considerable disservice. Final perspective of the letter is to urge responsible investigators interested in the problem to focus their research activities on the development of methods. That will provide efficient and objective ways to identify individuals from their speech, especially in the forensic situation. All these may be possible under undisguised voice. However, with vocal disguise the situation may be different. Reich et al. (1976) reported that the examiners were able to match

speakers with a moderate degree of accuracy (56.67%) when there was no attempt to vocally disguise either utterance. In spectrographic speaker identification nasal and slow rate were the least effective disguises, while free disguise was the most effective. Most of the speaker identifications are conducted in laboratory condition. The results may differ in actual conditions.

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

Following procedures were used in voice identification comparisons made by FBI examiners. (1) Only original recordings of voice samples were accepted for examination. (2) Recordings were played back on appropriate professional tape recorders and recorded on a professional full-track tape recorder at 7½ ips. (3) Spectrograms were produced on Sound Spectrograms, model 700, using linear expand frequency range (0 - 4000 Hz), wideband filter (300 Hz) and bar display mode. All spectrograms for each separate comparison were prepared on the same spectrogram. (4) When necessary, enhanced tape copies were also prepared from the original recordings. (5) Similarly pronounced words were compared between two voice samples. Normally, 20 or more different words were needed for a meaningful comparison. Less than 20 words usually resulted in a less conclusive opinion, such as possibly instead of probably. (6) Examiners made a spectral

pattern comparison between the two voice samples by comparing beginning mean and end formant frequency, formant shaping, pitch, timing, etc., of each individual word. (7) Aural examination was made of each voice sample to determine if pattern similarities or dissimilarities noted were the product of pronunciation differences, voice disguise, obvious drug or alcohol use, altered psychological state, electronic manipulation, etc. (8) Aural comparison was then made by repeatedly playing two voice samples simultaneously on separate tape recorders, and using high quality headphones. (9) Examiner then had to resolve any differences found between the aural and spectral results, usually by repeating all or some of the comparison steps. (10) If the examiner found the samples to be very similar (identification) or very dissimilar (elimination), an independent evaluation was always conducted by at least one, but usually two other examiners to confirm the results.

If differences of opinions occurred between the examiners, they were then resolved through additional comparisons and discussions by all the examiners involved. No or low confidence decisions were usually not reviewed by another examiner. Most of the no or low confidence decisions were due to poor recording quality and/or an insufficient number of comparable words. Decisions were also affected by high pitched voices (female) and some forms of voice disguise.

Pamela (2002) investigated the reliability of voiceprints by extracting acoustic parameters in the speech samples. Six normal Hindi speaking male subjects in the age range of 20-25 years participated in the study. Twenty-nine bisyllabic meaning Hindi words with 16 plosives, five nasals, four affricates and four fricatives in the word-medial position formed the material. Subject read the words

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

Some experiments were conducted by Reich et al (1976), to find out the effect of vocal disguises upon speaker identification. Reich (1976) described an experiment involving the effects of selected vocal disguises upon spectrographic speaker identification. The results of this experiment suggest that certain vocal disguises markedly interfere with spectrographic speaker identification. The reduction in speaker identification performance ranged from 14.17% (slow rate) to 35.00%

(free disguise). These experimental data obviously contradict Kersta's (1962) claim that spectrographic speaker identification is essentially unaffected by attempts at disguising one's voice. The mean performance level (56.67% correct) on the undisguised task was considerably poorer than the data for similar experimental conditions (approximately 80%) Tosi et al (1972).

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

The nasal disguise, for example, was the most effective disguise in speaker identification by listening experiment (Reich et al., 1979). In contrast, the nasal disguise was the least effective in a previous spectrographic matching experiment (Reich et al., 1976). Similarly, the power spectra of nasal consonants (Glenn and Kleiner, 1968) and coarticulated nasal spectra seem to provide strong cues for the machine matching of speakers. It is somewhat surprising then that the listeners in the present study were unable to successfully utilize these seemingly speaker dependent cues. The free (i.e., extemporaneous) disguise proved to be very

effective in both the spectrographic matching experiment (Reich et al., 1976) and the present listening experiment.

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

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

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.

3. Semi automatic or Automatic using machines (Objective Method)

The first and earliest method of is Speaker identification by machine to use *long term average* of acoustic features such as spectrum representations or pitch. In

some of the early studies by Furui (1972) and Markel and Davis (1979) the idea was to average out the other factors influencing the acoustic features such as the phonetic variations, leaving only the speaker dependant component. In this method the averaging process discards much speaker-dependant information and can require long (>20s) speech utterances to derive stable long-term speech statistics. This has been used successfully for several difficult text-independent speaker identification tasks by Gish (1985).

The second method is to model the speaker-dependent acoustic features within the *individual* sounds that comprise the utterance. By comparing acoustic features from sounds in a test utterance 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. In studies by Matsui & Furui (1991), and Rao et. al (1992) explicit segmentation was performed using a HMM based continuous speech recognizer as a front-end segment for text-independent speaker recognition systems. It was found in both the studies that the front-end speech recognizer provides little or no improvement in speaker recognition performance compared to the absence of front-end segmentation. Moreover, this imposes a significant increase in computational complexity on both training and recognition. Implicit segmentation by Soong et al., (1985), Helms (1981) and Higgins et al., (1993) on the other hand, relies on some form of unsupervised clustering to provide implicit segmentation of the acoustic features during both training and recognition. While this technique has demonstrated good performance on restricted vocabulary

(digits) tasks, it is limited in its ability to model the possible variability encountered in an unconstrained speech task.

The third method to 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 (1991), 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. The major drawback of many of the NN techniques is that the complete network must be retrained when a new speaker is added to the system.

Most current speaker recognition systems Eatock and Mason (1994), and Miyajima (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. However, Sambur (1975) and Orman (2000) have suggested that middle and higher frequency regions of the speech spectrum carry more speaker-specific information. A study done by Kumar and Rao (2004), a general method to obtain cepstral coefficients on different warped frequency scales was proposed. 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 has been 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.

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 modeling speaker identity. The focus of the 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 modeling techniques on an identical 16 speakers telephone speech task.

Furui (1981) describes 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 are extracted by means of LPC analysis on a frame-by-frame basis throughout an utterance. The frequency response distortions introduced by transmission systems are removed. Contours of cepstral coefficients are described by time functions. Results of the experiment indicate that verification error rate of one percent or less can be obtained even if the reference and test utterances are subjected to different transmission conditions. But, this study did not address the issue if the transmission system is over mobile phones.

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

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

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

Meltzer and Lehiste (1972) investigated the relative quality of synthetic speech. They selected three speaker one man, one women and one child. They recorded a set of 10 monophthong English vowels 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.

Wolf (1972) describes an investigation of an efficient approach to selecting such parameters, which are motivated by known relations between the voice signal and vocal-tract shapes and gestures. In a scheme for the mechanical recognition of speakers it, is desirable to use acoustic parameters that are closely related to voice characteristics that distinguish speakers. Useful parameters were found in F0, features of vowel and nasal consonant spectra, estimation of glottal source spectrum slope, word duration, and voice onset time. These parameters were tested in speaker recognition paradigms using simple linear classification procedures. When only 17 such parameters were used, no errors were made in speaker identification from a set of 21 adult male speakers. Under the same condition, speaker verification errors of the order of 2% were also obtained.

Atal (1972) examined the temporal variations of pitch in speech as a speaker identifying characteristics. The pitch data was obtained from 60 utterances, consisting of six repetitions of the same sentence, spoken by 10 speakers. The pitch data for each utterance was represented by a 20-dimensional vector in the Karhunen-Loeve coordinate system. The 20-dimensional vectors representing the

pitch contours were linearly transformed so that the ratio of inter-speaker to intra-speaker variance in the transformed space was maximized. The percentage of correct identifications was reported 97% and suggested that temporal variations of pitch could be used effectively for automatic speaker recognition.

In another experiment Atal (1974) examined several different parameters using linear prediction model for their effectiveness for automatic recognition of speakers from their voices. He determined twelve predictor coefficients approximately once every 50 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 consisted of 60 utterances, consisting of six repetitions of the same sentence spoken by 10 speakers. 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.

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

Jakkar (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 only direct recording . Results of this 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.

Doddington et al. (1974) 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 demonstrated high

levels of correct speaker identification for normal speech, slightly reduced scores for speech during stress and markedly reduced correct identifications for disguised speech. Finally, it appears that distortions created by limited pass band and stress as these two factors are defined in these experiments have only minimal effects on the sensitivity of the LTAS vector as a speaker identification cue.

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.

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

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

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

Conclusions of above results of initial studies on speaker identification by semi-automatic methods indicate that quite accurate speaker identification can be achieved on the basis of spectral information taken from individual segment of an utterance of nasal sound. The effect of coarticulation could have been factored out and it might be expected that within-speaker variation would have been reduced and as a result some of the errors eliminated.

Kinnunen (2003) indicated that the ***Mel-frequency Cepstral Coefficients*** (MFCC) is the most evident example of a feature set that is extensively used in speaker recognition. In using MFCC feature extractor, one makes an assumption that the human hearing mechanism is the optimal speaker recognizer. 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 ***MFCC***s 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% Net work vs. network recording.

Speech communication considers the *perception* of speech, i.e. how human listener's auditory system processes speech sounds. The discipline of sound perception in general is referred to as *psychoacoustics*. Techniques adopted from psychoacoustics are extensively used in audio- and speech processing systems for reducing the amount of perceptually irrelevant data. Psychoacoustics aims at finding connections between the physical, objectively measurable auditory stimuli, and the subjective impression about what the listener has about the stimuli.

The *loudness* of a sound is not linearly proportional to the measured sound intensity. For instance, if the sound intensity is doubled, it is not perceived "twice as loud" in general.

Fundamental frequency (*F*0) is defined as the rate at which the vocal folds vibrate during voiced phonation. Psycho acousticians call perceived *F*0*pitch*. Even if a speech signal is filtered so that the frequency region of the fundamental is not present in the signal, humans can perceive it.

The human ear processes fundamental frequency on a logarithmic scale rather than a linear scale. It has been observed that in the high frequencies, the *F*0 must change more that 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*. If the fundamental frequency is kept the same and the relative amplitudes of the upper harmonics are changed, the sound were perceived as having different *timbre*. Thus, timbre is the perceptual attribute of the spectral shape, which is known to be an important feature in speaker recognition. For instance, the widely used mel-cepstrum feature set measures the perceptual spectral shape. Studies of the human hearing mechanism show 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 impetus for designing perceptually motivated filter banks as front-ends for speech and speaker recognition systems. One should question the usefulness of perceptual frequency scales in speaker recognition. Perceptually motivated representations have been used successfully in speech recognition, and a little ironically, in speaker recognition as well, despite the opposite nature of the tasks. The implicit assumption made when using psycho acoustical representations is that the human ear is the optimal recognizer. If this is not true, then we are throwing useful information away! (Tomi Kinnunen, 2003).

Psychophysical studies of the frequency resolving power of the human ear has motivated modeling the non-linear sensitivity of human ear to different frequencies. MFCC's are based on the known variation of the human ears critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies. In addition, MFCC's are shown to be less

susceptible to the variation of the speaker's voice and surrounding environment. Initially, Fast Fourier Transformation (FFT) of a speech sample is extracted which is converted to Mel frequency. Cepstral coefficients are extracted on Mel frequencies. Figure 2 illustrates Mel filtering.

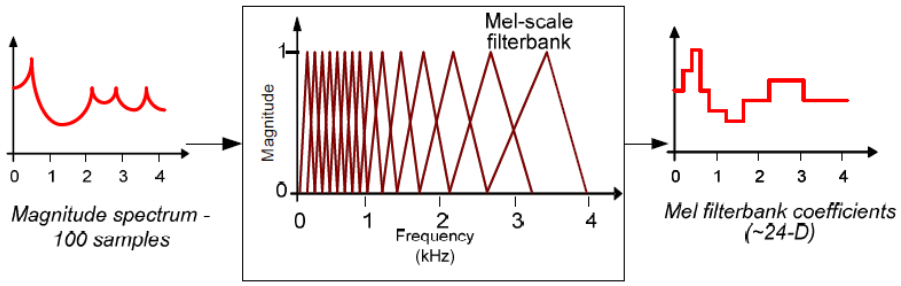


Figure 2: Illustration of Mel filtering [Taken from Milner, 2003].

It is evident from the review that MFCCs is, perhaps, the best parameter for speaker identification. Also, nasal continuants may be the most suitable, among speech sounds, for speaker identification. However, 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 analysis becomes a must. ***In this context, the present study is planned.*** The aim of the study is to establish ***Benchmark for speaker identification for nasal continuants in Kannada using Mel Frequency Cepstral Coefficients in Kannada***. In the Mysore dialect of spoken Kannada the frequency of occurrence of bilabial /m/ is 2.76% , dental /n/ is 7.59% and retroflex / n̪ / is 0.29% (Sreedevi- 2013).

The objectives of the study are two-fold and as follows:

1) To find out the Mel frequency Cepstral Coefficients for Kannada nasal continuants in direct and mobile recording, thus providing benchmark for speaker identification, and to
2) Compare the MFCCs across three age groups of 20≤30 years, 30≤40 years, and 40≤50 years.

CHAPTER III

METHOD

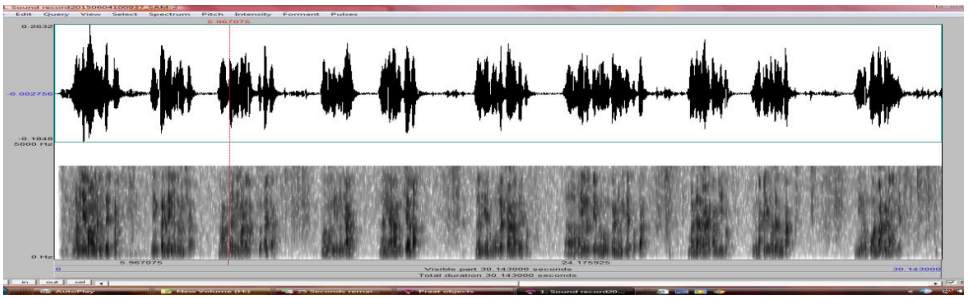
Participants: Male participants, 10 each in the age range of 20 ≤30 years, 30≤40 years, and 40≤50 years with at least 10 years of exposure to Kannada language as a mode of oral communication were included in the study. The inclusion criteria of the participants were (a) no history of speech, language and hearing problems, (b) reasonably free from cold and other respiratory illness and oral restructuring at the time of recording.

Stimulus: Three Kannada nasal continuants - bilabial /m/, dental /n/ and retroflex /n̪/ - as occurring in initial, and medial positions in 30 meaningful Kannada words were selected. Using these words, 10 meaningful 3/4-word sentences were formed to maintain the naturalness of speech. The sentences used were as follows:

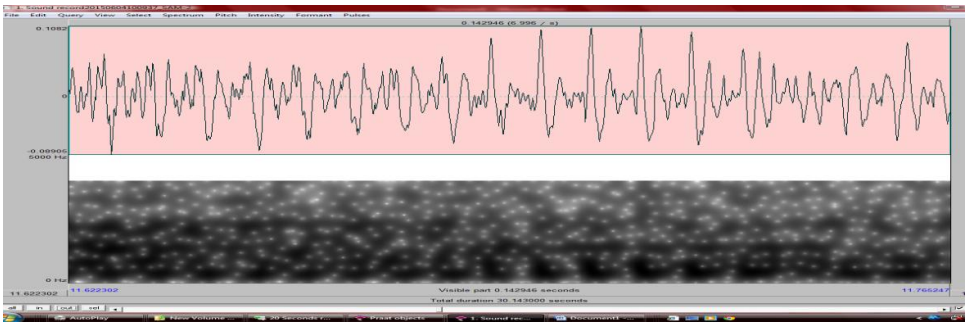
1. /**m**agu **n**i:**n**u ja:**n̪**a /
2. /idu **n**akali ba**n̪** **n̪**ada **m**uka/
3. /**n**a:vu dhara**n̪i** **m**a:dalla/
4. /ba**n̪** **n̪**ada **n**a:taka **m**ugijitu/
5. /ni:nu karun̪e illada manushya/
6. /**n**a**n**age **m**u:ru laks̪a ha **n̪**a be:ku/
7. / **m**aju:rakke **n**u:ru ka**n̪** **n̪**u ide /
8. /**n**a**n**age ha **n̪**a **m**ukhja/
9. /nanage ma:vina han̪n̪ /
10. /**n**i:**n**u fo:**n**u **m**a:d̪abe:d̪a/

Recording Procedure: The participant were given the written material to familiarize themselves to utter the sentences at a normal rate of speech into a mobile phone. They were instructed to speak under two conditions, directly into the recording mobile (direct) and through another mobile into the recording mobile phone (network). Each participant was instructed to utter the sentences 3 times. The network used for making the calls was Airtel 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 mobile. Later the .gpp format files were converted to .wav files using Total video converter and Praat software (Boersma and Weenink, 2009) so that analysis could be carried out in an effective manner on a computer.

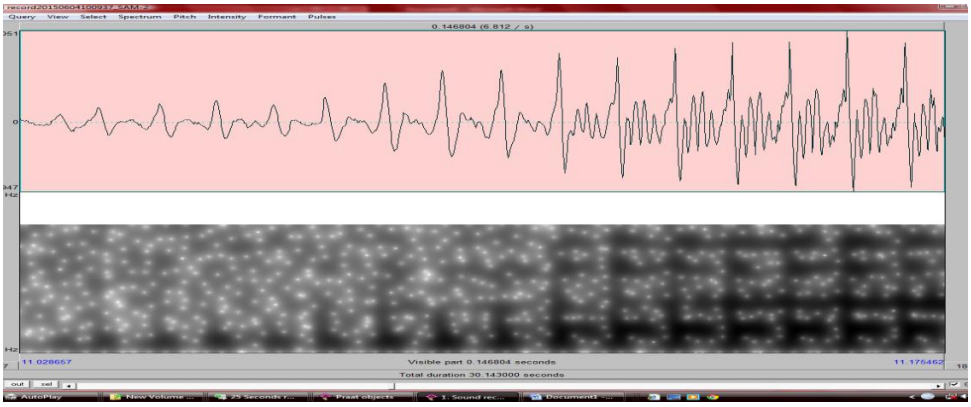
Speech Segmentation: The .wav converted speech sample wave opened with Praat software and identified the words with nasal continuants at word - initial, medial and final positions were identified and segmented. Segmented words were saved as .wav file for each speaker for all the nasal continuants. Figure 3 illustrates segmentation.



(a)



(b)



(c)

Figure 3: Segmentation of samples for (a) /m/, (b) /n/ and (c) /n̪/.

In the present project, stimulus contained totally 10 sentences out of which only 9 sentences were taken for final analyses. A total of 27 nasal continuants occurred in these 9 sentences. Thus, the total number of samples for each speaker was 162 (27 * 3 * 2), and the total number of samples stored for 30 speakers were 4860.

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 4 illustrates the note pad.

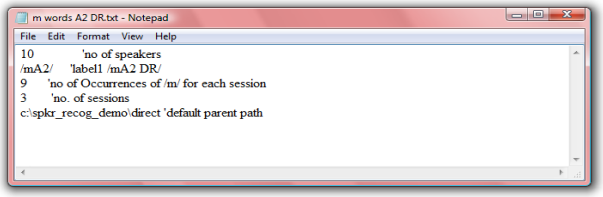


Figure 4: Illustration of the note pad.

The segmented material was analyzed to extract MFCCs. 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 5 and 6, respectively.

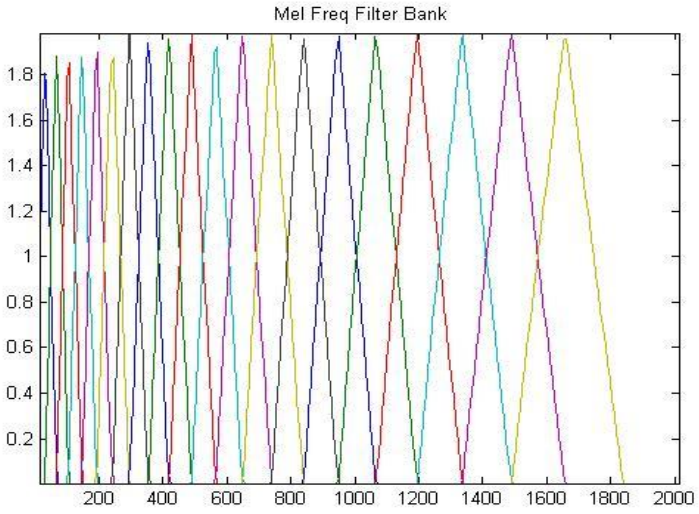


Figure 5: Mel frequency filter bank without normalization.

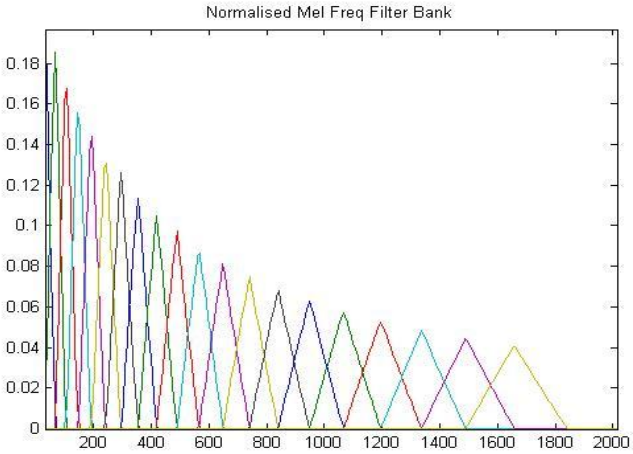


Figure 6: Mel frequency filter bank with normalization.

The notepad file was opened in SSL Workbench as in figure 7.

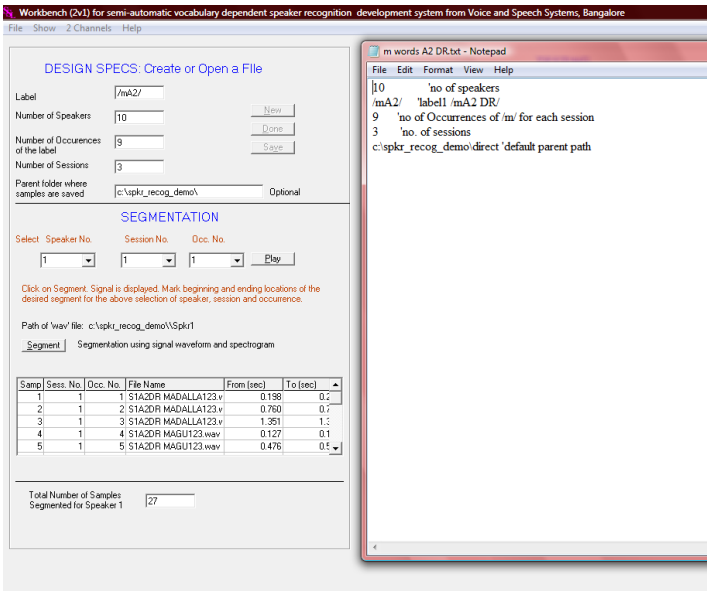


Figure 7: Notepad of SSL workbench.

The ‘number of occurrence’ was specified according to the occurrence of nasal continuant being studied. The ‘number of sessions’ was specified as 3 for the first three results, but was kept as two for the last result, as the participants will utter each sentence thrice. The parent file name was also specified in the notepad file. This is

the file where the recordings were saved and is 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 opens the location specified in the parent file path of notepad file. Following this, the experimenter chose the file from the folder. Figure 8 shows the workbench window for analyses.

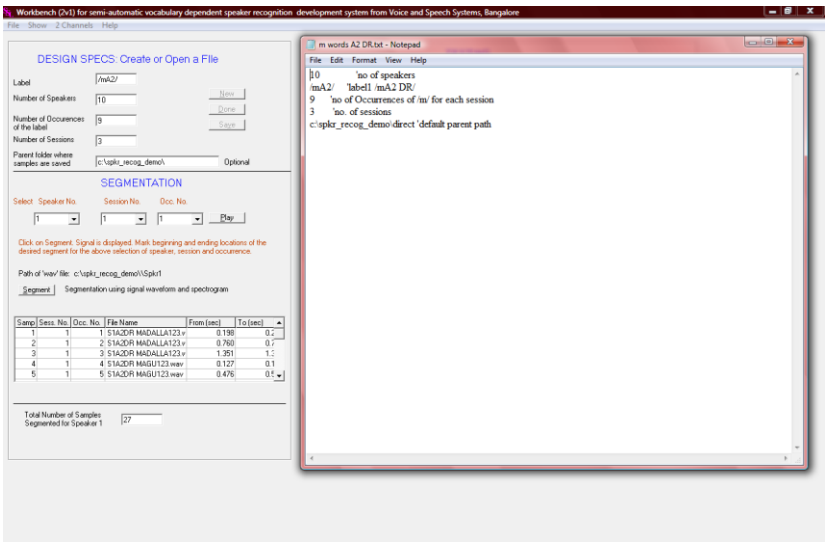


Figure 8: SSL Workbench window for analysis.

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 9 illustrates the speaker number being selected for segmentation.

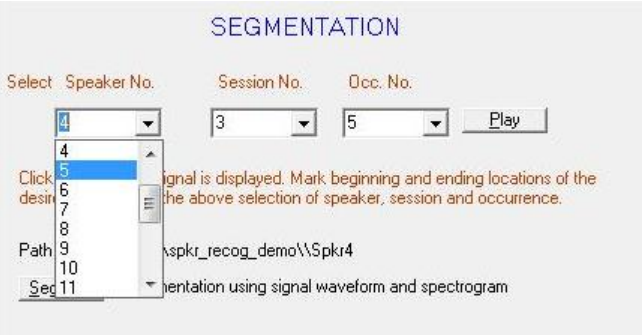


Figure 9: Illustration of speaker number being selected for segmentation.

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 10 illustrates selecting the session number and occurrence number.

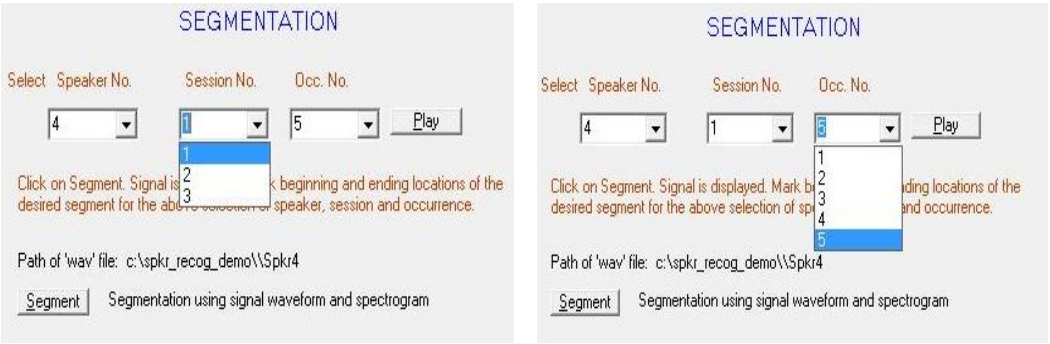


Figure 10: Illustration of selecting the session number and occurrence number.

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 11 illustrates segmentation window showing 5 occurrence of /m/ for a speaker.

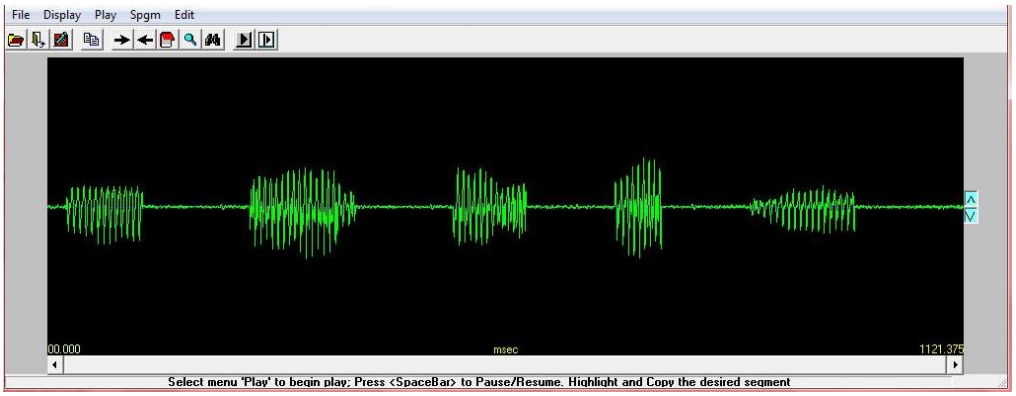


Figure 11: Depiction of segmentation window showing 5 occurrence of /m/ for a speaker.

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 12 shows the dialogue box seeking for confirmation of the highlighted segment in the file.

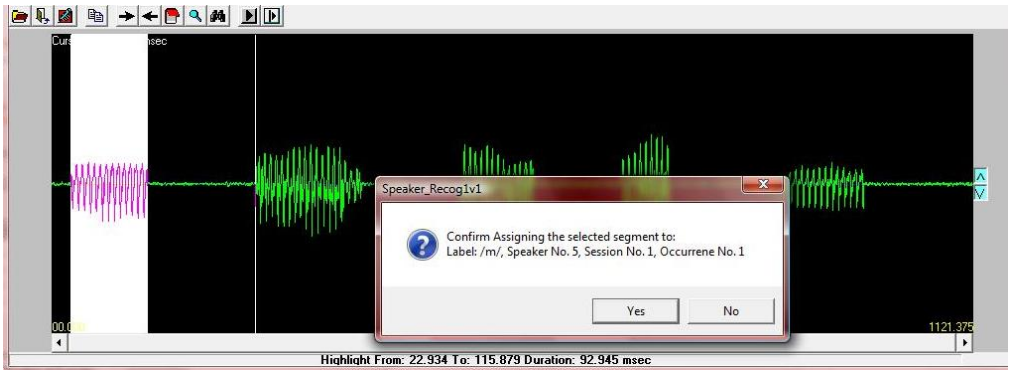


Figure 12: Showing dialogue box asking for confirmation of the highlighted segment in the file.

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 13 shows the analysis window of SSL Workbench.

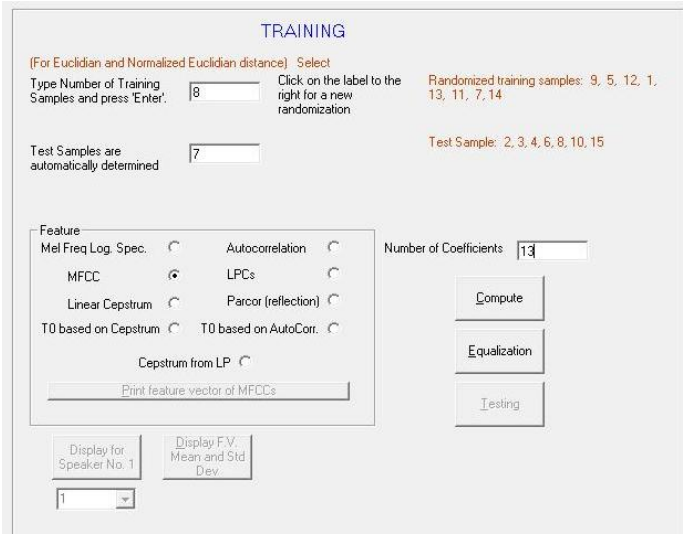


Figure 13: Analysis window of SSL Workbench.

Telephone equalization: Equalization is the process commonly used to alter the frequency response of an audio system using linear filters. Equalization may be used to eliminate unwanted sounds, make certain instruments or voices more prominent, and enhance particular aspects of an instrument's tone.

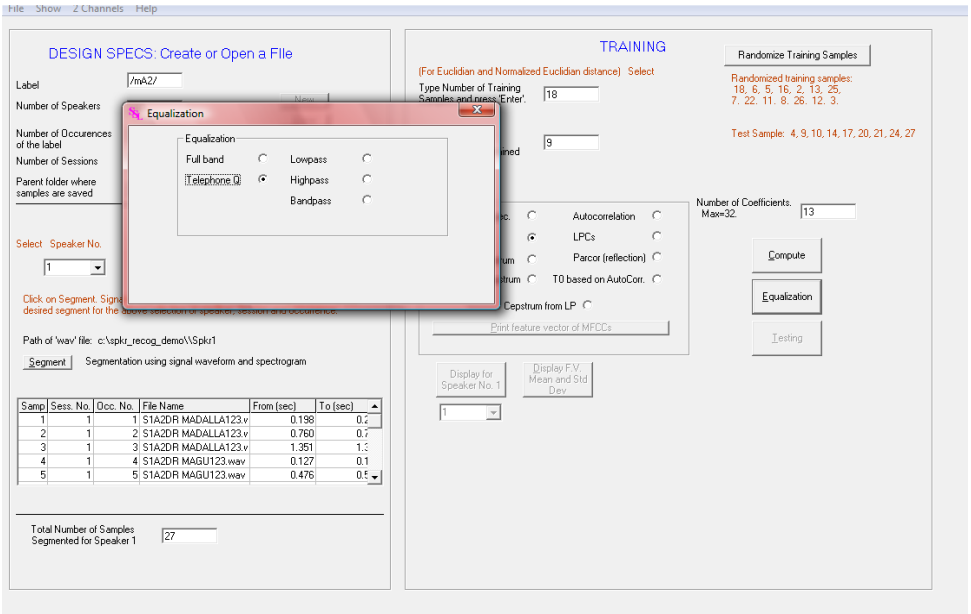


Figure 14: Telephone equalization selection window.

Training sample numbers was specified and the rest were automatically selected as test samples. Once this was done, ‘compute’ was clicked on. This 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 15) and a final percentage for correct speaker identification.

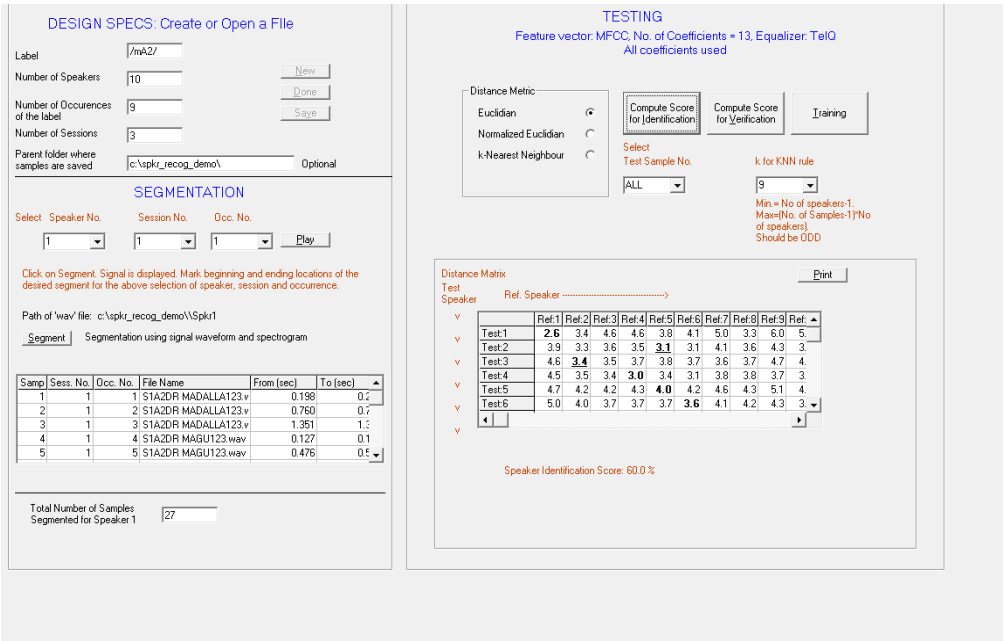


Figure 15: Analysis window of SSL Workbench showing diagonal matrix and the final speaker identification score.

This data was stored and the same procedure was repeated. Direct recordings were repeated 5 times; but network recordings were not as they were taken as reference and compared with one direct recording of the same speaker as test sample. 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 points p and q is the length of the line segment connecting them ($\overline{\mathbf{pq}}$). In Cartesian coordinates, if p = (p1, p2,..., pn) and q = (q1, q2,..., qn) are two points in Euclidean n-space, then the distance from p to q, or from q to p is given by:

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

The Euclidian distance between 13 MFCCs was extracted and also within and between participants was noted. 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 were calculated using the following formula:

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

In this study, all the speech samples were contemporary, as all the recordings of a participants were carried out in same session. Closed set speaker identification tasks were performed, in which the experimenter is aware that the 'unknown speaker' were the 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

The results are discussed under the following headings:

(1) Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 20≤30 years in direct mobile recordings.

(2) Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 20≤30 years in network recordings.

(3) Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 30≤40 years in direct mobile recordings.

(4) Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 30≤40 years in network recordings.

(5) Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 40≤50 years in direct mobile recordings.

(6) Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 40≤50 years in network recordings.

(7) Percent speaker identification in direct and network recoding.

(1) ***Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 20≤30 years in direct mobile recordings:*** The percent correct identification was 82 %, 89 % and 93 % for /m/ , /n/ and /n̩/, respectively. Tables 1 to 3 show the distance matrix for /m/ , /n/ and /n̩/, respectively. In all the tables red colour shows incorrect speaker identification and green colour

shows correct speaker identification.

sp	1	2	3	4	5	6	7	8	9	10
1	**2.505**	3.581	3.65	3.848	3.778	3.414	3.813	3.728	6.152	5.219
2	4.008	**3.018**	3.22	3.263	3.41	3.485	3.398	5.203	4.276	3.714
3	**3.835**	4.163	4.161	4.242	3.938	4.021	3.87	4.451	5.97	5.209
4	4.733	3.729	3.624	**3.375**	3.833	4.506	3.804	6.097	4.181	3.844
5	3.573	3.105	3.436	3.199	**2.212**	3.157	2.567	4.301	4.682	3.761
6	3.776	3.438	3.305	3.706	3.27	**2.627**	3.171	3.965	4.87	4.149
7	3.984	3.648	3.574	3.501	3.047	3.785	**3.007**	4.919	4.636	3.954
8	4.367	5.382	5.454	5.797	5.293	4.608	5.115	**3.439**	7.624	6.858
9	6.845	5.002	5.002	4.888	4.852	5.641	4.802	7.947	**3.67**	3.764
10	5.134	4.049	3.916	4.312	3.95	3.87	3.878	5.653	4.421	**3.844**

Table 1: Distance matrix for speaker identification for /m/ in direct recording in the age range 20≤30 years.

sp	1	2	3	4	5	6	7	8	9	10
1	**3.66**	4.481	4.092	4.142	5.665	4.828	5.04	4.651	6.397	6.739
2	3.532	**2.202**	3.013	2.262	3.73	3.15	3.097	4.535	3.856	4.088
3	4.617	4.601	**4.585**	4.652	5.785	5.124	5.099	5.541	5.911	6.406
4	5.703	5.059	5.298	**4.951**	5.892	5.745	5.337	6.121	5.828	6.222
5	5.505	3.136	3.626	3.865	**2.331**	3.616	2.839	5.371	3.23	3.05
6	4.354	3.731	3.754	3.827	4.593	**3.369**	4.289	4.464	5.373	5.269
7	5.655	3.674	3.677	4.045	3.665	4.258	**2.878**	5.611	3.658	4.117
8	3.967	4.863	3.575	4.638	4.834	4.35	4.807	**2.537**	6.885	6.933
9	5.376	4.301	4.299	4.562	4.422	4.581	**4.041**	5.661	4.496	4.844
10	6.916	4.077	5.213	4.901	4.055	4.833	3.881	7.218	2.999	**2.569**

Table 2: Distance matrix for speaker identification for /n/ in direct recording in the age range 20≤30 years.

sp	1	2	3	4	5	6	7	8	9	10
1	**3.186**	4.267	3.488	4.087	5.426	4.241	4.656	4.958	5.699	5.556
2	4.002	**2.364**	3.334	3.174	3.358	3.006	2.968	6.028	3.505	2.931
3	4.135	4.455	**3.65**	4.592	5.015	4.319	4.236	5.462	5.405	5.167
4	4.516	3.998	4.004	**3.546**	4.612	4.71	3.947	6.352	4.259	4.739
5	4.845	3.366	4.088	4.051	**2.603**	4.042	2.98	6.433	3.244	2.957
6	3.347	2.847	3.193	4.067	3.73	**2.237**	3.433	4.541	4.892	3.748
7	3.962	2.985	3.353	3.417	2.795	3.522	**2.435**	5.495	3.299	3.274
8	4.087	5.545	5.403	5.781	6.357	4.82	5.965	**2.888**	7.763	6.914
9	6.012	4.37	4.755	4.627	4.013	5.302	3.746	8.044	**2.939**	3.807
10	4.992	3.262	4.013	4.116	3.821	3.816	3.677	7.051	3.64	**2.766**

Table 3: Distance matrix for speaker identification for /n̪/ in direct recording in the age range 20≤30 years.

(2) ***Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̪/ in the age range 20≤30 years in network recordings:*** The percent correct identification was 96 %,90 % and 84 % for /m/ , /n/ and /n̪/, respectively. Tables 4 to 6 show the distance matrix for /m/ , /n/ and /n̪/, respectively. In all the tables red colour shows incorrect speaker identification and green colour shows correct speaker identification.

sp	1	2	3	4	5	6	7	8	9	10
1	**4.404**	5.496	5.444	5.343	5.315	4.936	5.998	7.246	6.303	6.462
2	6.114	**3.901**	5.343	4.351	5.151	6.659	6.31	10.402	5.118	5.241
3	4.909	4.703	**4.167**	4.554	5.245	4.995	4.746	8.584	5.934	5.438
4	6.581	5.627	6.322	**5.589**	6.057	7.485	7.31	10.158	6.465	6.011
5	5.154	5.298	4.367	4.2	**3.592**	5.261	5.016	8.096	5.121	4.384
6	5.529	6.416	5.296	5.847	5.687	**4.454**	5.187	8.108	6.72	6.797
7	6.978	6.646	5.793	6.565	6.794	6.345	**5.629**	9.894	7.303	7.002
8	6.057	8.052	7.167	8.181	7.488	6.867	6.99	**3.598**	9.754	7.902
9	6.81	5.705	6.129	5.152	5.195	6.861	7.121	11.028	**4.353**	5.969
10	5.665	5.254	4.547	4.6	4.523	6.416	5.248	8.567	5.627	**3.948**

Table 4: Distance matrix for speaker identification for /m/ in network recording in the age range 20≤30 years.

sp	1	2	3	4	5	6	7	8	9	10
1	**5.479**	6.337	6.862	5.885	6.375	7.294	7.778	7.981	7.743	6.447
2	5.417	**4.176**	4.79	5.001	4.841	5.423	5.421	7.224	5.582	4.59
3	6.298	6.475	**5.247**	5.958	6.128	5.494	5.473	7.634	6.716	5.971
4	6.119	6.815	6.765	**5.745**	6.214	6.57	7.351	8.296	7.88	6.468
5	5.52	5.308	4.793	5.159	**3.408**	5.186	5.878	7.396	4.518	4.281
6	7.235	6.775	5.697	6.836	6.508	**5.568**	6.019	8.313	6.932	6.451
7	6.589	6.911	4.555	6.278	6.326	4.53	**4.127**	6.83	6.73	6.205
8	6.01	8.366	7.289	7.462	7.429	7.314	7.685	**3.404**	9.302	8.27
9	7.914	6.899	6.512	7.311	5.991	7.478	7.56	9.966	**5.499**	6.133
10	4.328	4.95	4.43	**3.893**	4.084	5.25	5.244	6.756	5.541	4.117

Table 5: Distance matrix for speaker identification for /n/ in network recording in the age range 20≤30 years.

sp	1	2	3	4	5	6	7	8	9	10
1	**4.549**	5.109	5.649	5.729	5.057	6.311	6.076	8.713	4.732	4.936
2	4.978	**4.715**	5.866	6.101	5.68	6.68	5.949	8.44	5.73	5.223
3	5.675	5.791	**4.698**	5.337	5.85	5.146	4.833	7.951	6.097	5.904
4	6.698	6.722	6.423	**6.233**	6.742	6.868	6.399	8.436	7.171	6.721
5	5.994	5.878	6.262	6.059	**3.847**	6.375	6.091	8.545	4.749	5.676
6	7.178	7.518	6.079	6.669	6.661	**4.767**	5.715	6.424	7.83	7.865
7	5.72	5.04	4.092	4.82	5.299	4.718	**3.719**	7.786	5.583	5.384
8	7.616	8.532	7.882	7.384	7.604	7.376	7.36	**3.917**	9.418	8.505
9	5.584	5.326	5.308	5.63	**4.477**	5.617	5.417	8.855	4.573	5.348
10	5.338	5.254	5.457	4.658	5.472	6.76	5.37	7.87	6.068	**4.44**

Table 6: Distance matrix for speaker identification for /n̩/ in network recording in the age range 20≤30 years.

(3) ***Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 30≤40 years in direct mobile recordings:*** The percent correct identification was 66 %, 89 % and 88 % for /m/ , /n/ and /n̩/, respectively. Tables 7 to 9 show the distance matrix for /m/ , /n/ and /n̩/, respectively.

sp	1	2	3	4	5	6	7	8	9	10
1	**3.141**	3.801	4.815	4.732	4.237	4.363	5.374	3.892	6.255	4.932
2	4.201	**3.281**	3.719	3.682	3.717	3.571	3.957	3.557	4.619	4.323
3	4.753	3.697	**3.445**	3.876	3.787	3.755	3.977	3.663	4.64	4.312
4	4.554	3.12	2.806	**2.636**	3.198	2.84	3.043	3.435	3.226	3.557
5	4.219	**3.654**	3.925	3.93	3.705	3.821	4.199	3.914	4.85	4.05
6	4.784	3.535	3.341	3.26	3.496	**3.25**	3.564	3.901	3.96	3.673
7	5.182	4.169	4.162	**4.115**	4.354	4.151	4.18	4.494	4.752	4.694
8	4.293	3.734	4.067	4.241	4.196	4.097	4.35	**3.602**	5.236	4.965
9	6.585	5.143	4.495	4.344	4.975	4.76	4.401	6.01	**3.809**	4.45
10	5.153	4.242	4.439	4.205	**3.955**	4.088	4.589	4.779	4.876	4.052

Table 7: Distance matrix for speaker identification for /m/ in mobile recording in the age range 30≤40 years.

sp	1	2	3	4	5	6	7	8	9	10
1	**2.846**	3.388	4.005	3.311	3.125	3.583	5.105	4.107	4.94	3.947
2	3.077	**2.766**	3.87	3.264	3.122	3.533	4.282	3.212	5.045	3.817
3	4.064	3.591	**3.048**	3.637	3.553	3.417	3.592	4.076	3.498	3.427
4	3.611	3.571	3.754	**3.028**	3.538	3.337	3.699	4.276	4.101	3.816
5	3.515	3.545	3.675	3.575	**3.292**	3.564	4.453	4.097	4.512	3.694
6	3.572	3.539	3.239	3.042	3.205	**2.898**	3.801	4.393	3.083	3.241
7	5.831	5.208	4.728	5.189	5.332	5.002	**4.397**	5.868	4.731	4.544
8	3.779	3.635	4.599	4.179	4.129	4.143	4.926	**2.838**	5.896	5.076
9	5.362	4.734	3.935	4.492	4.646	4.292	3.763	5.47	**3.492**	4.031
10	3.706	3.919	3.806	**3.517**	3.527	3.716	4.588	4.746	4.145	3.702

Table 8: Distance matrix for speaker identification for /n/ in mobile recording in the age range 30≤40 years.

sp	1	2	3	4	5	6	7	8	9	10
1	2.464	3.193	3.258	3.215	2.979	2.903	3.375	4.272	4.721	3.28
2	3.315	2.528	2.84	2.69	3.181	2.856	2.555	3.391	4.431	3.254
3	2.923	3.131	2.496	2.728	2.515	2.583	2.76	3.824	3.748	3.322
4	3.708	3.337	3.504	2.768	3.525	2.999	3.045	4.468	4.149	3.627
5	3.275	3.277	2.836	3.32	2.691	2.907	3.105	3.833	4.717	3.447
6	3.127	3.001	3.269	3.093	3.221	2.873	3.125	4.111	4.376	3.224
7	4.749	4.332	4.233	3.875	4.354	4.017	3.905	4.968	4.599	4.596
8	4.265	3.567	3.505	3.899	4.076	4.186	3.6	3.192	4.954	4.725
9	5.181	4.806	4.291	3.908	4.561	4.329	4.282	5.889	3.023	4.127
10	4.524	4.165	4.064	4.012	3.803	3.731	4.006	5.51	4.924	3.284

Table 9: Distance matrix for speaker identification for /n̩/ in mobile recording in the age range 30≤40 years.

(4) ***Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 30≤40 years in network recordings:*** The percent correct identification was 86 %, 91 % and 84 % for /m/ , /n/ and /n̩/, respectively. Tables 10 to 12 show the distance matrix for /m/ , /n/ and /n̩/, respectively.

sp	1	2	3	4	5	6	7	8	9	10
1	**4.389**	6.308	7.387	7.184	8.194	7.096	6.421	6.505	9.444	8.633
2	7.583	**5.735**	6.02	5.997	5.885	6.242	5.77	6.873	6.59	5.995
3	7.302	4.909	**4.166**	4.632	4.998	5.764	4.889	5.723	5.604	5.968
4	7.375	5.777	6.1	**4.747**	5.079	5.507	5.832	6.672	5.543	6.281
5	8.146	5.998	5.777	4.841	**4.67**	5.742	5.87	7.021	4.873	6.252
6	7.639	6.39	7.032	6.425	6.898	**5.992**	6.802	7.614	7.761	7.146
7	6.567	4.903	4.645	4.95	4.724	5.724	**4.205**	5.952	5.549	5.547
8	6.227	5.307	5.55	**5.031**	5.817	5.8	5.428	5.326	6.429	7.489
9	9.65	7.123	7.498	6.004	5.656	6.057	7.15	9.061	**5.099**	5.689
10	7.662	5.555	5.866	5.312	4.858	4.984	5.448	7.505	5.656	**4.516**

Table 10: Distance matrix for speaker identification for /m/ in network recording in the age range 30≤40 years.

sp	1	2	3	4	5	6	7	8	9	10
1	2.676	3.259	4.084	3.443	3.094	3.588	5.011	3.787	5.039	3.801
2	3.021	3.008	4.136	3.172	3.273	3.759	4.595	3.719	5.134	3.767
3	3.781	3.286	3.089	3.541	3.436	3.576	3.762	3.991	3.978	3.237
4	3.506	4.084	3.887	3.444	3.456	3.308	4.344	4.857	4.036	3.663
5	3.279	3.254	3.654	3.395	3.196	3.602	4.321	3.803	4.607	3.391
6	3.631	3.507	3.687	3.428	3.474	3.412	4.079	4.165	4.192	3.609
7	5.961	5.268	5.12	5.081	5.576	5.394	4.554	5.912	5.388	5.054
8	4.027	3.694	4.652	4.039	4.253	4.461	5.154	3.219	5.833	4.831
9	4.96	4.493	3.614	4.207	4.223	3.827	3.658	5.37	3.264	3.714
10	3.646	3.843	3.736	3.797	3.443	3.768	4.341	4.861	4.182	3.325

Table 11: Distance matrix for speaker identification for /n̪/ in network recording in the age range 30≤40 years.

sp	1	2	3	4	5	6	7	8	9	10
1	**4.765**	5.626	5.101	5.643	5.299	7.401	5.525	5.142	6.649	6.621
2	5.273	**3.363**	5.036	3.976	5.285	4.979	5.445	4.627	5.634	5.942
3	5.819	5.366	4.979	5.008	**4.525**	7.823	5.223	4.833	5.734	6.977
4	5.992	5.285	6.056	**5.041**	5.359	6.656	5.758	5.902	5.363	6.073
5	6.111	5.677	6.068	5.263	**3.694**	7.277	5.1	5.543	4.662	6.176
6	6.642	6.526	7.871	6.634	6.917	**5.118**	6.845	8.039	6.11	5.597
7	5.835	5.745	5.458	5.363	5.372	7.236	**5.321**	6.082	5.812	6.57
8	5.962	5.197	5.883	5.505	5.287	7.043	6.084	**4.329**	6.17	7.13
9	7.337	7.293	7.399	6.609	5.583	8.534	6.41	7.029	**5.072**	7.305
10	6.236	6.441	7.486	6.054	6.045	5.971	5.76	7.903	5.373	**4.323**

Table 12: Distance matrix for speaker identification for /n̪/ in network recording in the age range 30≤40 years.

(5) ***Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 40≤50 years in direct mobile recordings:*** The percent correct identification was 86 %, 78 % and 93 % for /m/ , /n/ and /n̩/, respectively. Tables 13 to 15 show the distance matrix for /m/ , /n/ and /n̩/, respectively.

sp	1	2	3	4	5	6	7	8	9	10
1	**3.148**	3.596	3.31	4.117	3.956	5.24	4.343	5.051	3.488	4.97
2	3.155	**2.618**	3.193	3.688	3.746	5.666	4.257	4.52	3.3	4.558
3	3.086	4.078	**2.999**	3.167	3.265	3.845	3.335	4.363	3.749	3.874
4	3.413	3.936	3.436	**2.788**	3.325	3.722	3.141	3.517	4.063	3.374
5	3.682	4.271	3.856	3.201	**2.917**	4.186	3.375	3.845	4.624	3.211
6	4.486	5.436	4.496	3.992	4.194	**3.207**	3.979	4.855	4.962	4.464
7	3.241	3.751	3.298	**3.089**	3.283	4.434	3.219	3.878	3.844	3.55
8	4.359	4.674	4.687	4.11	4.496	4.623	4.241	**3.051**	4.865	3.881
9	3.147	3.102	3.31	3.826	4.213	4.359	4.243	4.496	**2.61**	5.078
10	4.742	5.563	4.721	4.163	4.006	5.196	3.819	4.562	5.825	**3.344**

Table 13: Distance matrix for speaker identification for /m/ in direct recording in the age range 40≤50 years.

sp	1	2	3	4	5	6	7	8	9	10
1	**2.604**	3.55	3.091	3.239	3.395	4.515	3.605	4.252	2.684	3.97
2	3.817	**3.132**	4.389	4.895	4.952	6.578	5.623	6.117	4.424	5.759
3	3.937	4.117	**3.151**	3.301	3.512	4.211	3.597	4.505	3.47	4.129
4	3.928	4.028	**3.484**	3.502	3.642	4.746	4.016	4.604	3.611	4.307
5	3.937	4.511	3.551	3.051	**3.009**	3.998	3.474	3.487	3.592	3.512
6	5.891	6.153	5.062	4.479	4.226	**4.133**	4.87	4.441	5.089	5.123
7	4.464	4.693	3.518	**3.433**	3.616	4.434	3.444	4.58	3.898	4.133
8	5.248	6.226	5.391	4.467	4.264	4.365	4.911	**3.267**	5.027	4.606
9	3.084	3.638	2.936	2.846	2.731	3.452	3.279	3.709	**2.356**	4.033
10	5.576	5.957	4.837	4.214	4.15	4.896	4.595	4.118	5.084	**3.921**

Table 14: Distance matrix for speaker identification for /n/ in direct recording in the age range 40≤50 years.

sp	1	2	3	4	5	6	7	8	9	10
1	**2.459**	3.297	3.278	3.416	3.622	4.865	4.415	4.712	3.528	4.58
2	3.502	**2.967**	3.774	3.419	3.82	4.289	4.878	4.5	3.103	5.255
3	3.588	4.012	**2.878**	3.109	3.972	4.219	3.751	4.698	4.005	4.185
4	3.713	4.048	3.414	**2.679**	3.513	3.189	3.572	4.14	3.227	4.197
5	4.041	3.874	3.388	3.051	**3.143**	4.212	3.439	4.329	4.206	3.732
6	4.455	4.595	4.261	3.355	4.113	**3.261**	4.119	3.808	3.863	4.596
7	4.91	5.144	4.506	4.101	4.288	5.196	**3.828**	5.117	5.12	4.277
8	4.796	5.364	4.852	4.375	4.544	4.328	4.306	**2.797**	4.994	4.119
9	3.239	3.39	3.747	3.309	4.252	3.337	4.826	4.528	**2.147**	5.299
10	5.353	5.718	4.666	4.373	4.068	5.465	3.258	4.834	5.921	**3.051**

Table 15: Distance matrix for speaker identification for /n̩/ in direct recording in the age range 40≤50 years.

(6) ***Percent speaker Identification score for the Nasal continuant /m/ /n/ and /n̩/ in the age range 40≤50 years in network recordings:*** The percent correct identification was 90 %, 87 % and 88 % for /m/ , /n/ and /n̩/, respectively. Tables 16 to 18 show the distance matrix for /m/ , /n/ and /n̩/, respectively.

sp	1	2	3	4	5	6	7	8	9	10
1	**4.279**	4.001	6.1	5.247	4.883	4.345	5.323	6.2	5.021	5.266
2	4.003	**3.755**	5.67	4.996	4.611	4.171	4.96	5.855	4.568	4.983
3	5.554	5.986	**3.754**	6.733	5.873	6.831	5.405	5.125	5.495	6.984
4	6.108	5.402	7.229	**4.106**	5.219	7.712	6.533	6.329	6.941	5.732
5	5.108	5.061	5.254	5.642	**4.673**	5.97	6.006	5.05	4.839	6.388
6	3.946	3.753	5.991	5.657	5.094	**2.089**	4.654	6.21	4.697	4.581
7	5.345	5.015	5.498	5.407	5.553	5.651	**4.129**	5.932	5.694	5.401
8	6.038	5.76	4.471	5.18	4.953	6.989	5.005	**3.482**	5.223	6.404
9	5.082	4.815	5.689	5.425	5.522	4.814	5.499	5.907	**3.699**	6.401
10	6.211	5.665	7.506	5.823	5.727	6.502	5.715	7.054	7.62	**4.53**

Table 16: Distance matrix for speaker identification for /m/ in network recording in the age range 40≤50 years.

sp	1	2	3	4	5	6	7	8	9	10
1	**5.696**	7.374	6.28	6.496	6.064	6.748	6.104	6.202	6.008	7.152
2	4.476	**3.078**	4.965	4.655	3.575	3.082	4.397	4.617	3.586	4.277
3	6.419	8.225	**4.888**	6.813	6.042	7.269	5.735	6.166	6.268	7.398
4	5.175	4.987	4.906	**3.358**	4.046	4.237	4.584	3.806	3.889	4.648
5	6.701	6.981	6.156	5.699	**5.606**	6.203	6.344	5.797	6.164	6.491
6	5.256	3.487	5.488	4.594	4.003	**3.399**	4.776	4.484	4.086	4.562
7	5.376	5.82	4.656	4.633	**4.312**	4.79	4.496	4.754	4.748	4.549
8	6.281	5.585	6.271	4.812	5.081	5.094	5.822	**4.521**	5.189	5.556
9	4.996	5.311	5.018	4.538	4.429	4.739	4.905	4.66	**3.833**	5.466
10	5.911	5.28	5.04	4.626	4.326	4.36	4.517	4.546	4.994	**3.579**

Table 17: Distance matrix for speaker identification for /n/ in network recording in the age range 40≤50 years.

sp	1	2	3	4	5	6	7	8	9	10
1	6.218	7.536	7.236	5.996	6.563	7.13	6.483	6.276	6.312	6.874
2	6.518	3.211	8.02	5.462	5.486	4.937	5.987	6.101	5.71	5.586
3	5.264	7.412	4.44	6.083	5.467	6.095	5.621	5.239	5.888	6.074
4	4.75	5.63	5.646	3.855	4.164	4.571	4.965	3.931	4.83	4.911
5	4.54	5.473	5.176	4.204	3.624	5.077	4.863	3.7	4.842	4.876
6	4.681	4.844	5.017	4.386	4.447	3.628	4.714	4.375	4.188	4.394
7	5.982	5.67	6.943	5.393	5.487	5.776	4.658	5.684	6.138	4.952
8	3.978	6.095	5.153	4.176	3.865	4.785	4.759	3.73	4.899	5.048
9	3.965	4.385	5.221	3.89	4.453	3.941	4.746	4.502	2.679	4.748
10	6.663	6.259	6.636	5.94	5.706	5.846	5.318	5.768	6.513	4.698

Table 18: Distance matrix for speaker identification for /n̪/ in network recording in the age range 40≤50 years.

(7) ***Percent speaker identification in direct and network recoding:*** It was observed that in direct recording, % Speaker identification (SPID) for /m/ was better in the age range of 40≤50 years compared to that in the age range of 20≤30 years and 30≤40 years. % SPID for /n/ was better in the age range of 20≤30 years compared to that in the age range of 30≤40 years and 40≤50 years; for /n̪/ % SPID was better in the age range of 20≤30 years compared to that in the age range of 40≤50 years and 30≤40 years.

In network recording, % SPID for /m/ was better in the age range of 20≤30 years compared to that in the age range of 40≤50 years and 30≤40 years. %

SPID for /n/ was better in the age range of 30≤40 years compared to that in the age range of 20≤30 years and 40≤50 years; for /n̪/ % SPID was better in the age range of 20≤30 years compared to that in the age range of 40≤50 years and 30≤40 years. No specific age preferences were observed.

% SPID in direct recoding was78, 83, and 91.33 for /m/, /n/, and /n̪/, respectively. In network recording it was 90.67, 89.67, and 88.33 for /m/, /n/, and /n̪/, respectively. % SPID was higher in network recording compared to direct recording for /m/ and /n/, and higher in direct recording for /n̪/.

Overall % SPID was 84.11 and 88.56 in direct and network recoding, respectively. Table 19 shows percent speaker identification for 3 nasal continuants in direct (DR) and network (NR) recordings, and table 20 shows benchmark for speaker identification.

Phneme	1	2	3	4	5	6	7	8	9	Average
mA1DR	90	80	70	80	100	90	70	90	70	**82.222**
nA1DR	90	100	90	90	90	80	90	80	90	**88.889**
NA1DR	90	90	90	90	100	90	90	100	100	**93.333**
mA1NR	90	100	90	100	100	90	90	100	100	**95.556**
nA1NR	90	80	80	90	90	100	90	100	90	**90**
NA1NR	70	90	90	100	80	80	70	100	80	**84.444**
mA2DR	60	60	70	60	80	80	50	50	80	**65.556**
nA2DR	90	100	80	100	70	80	90	90	100	**82.22**
NA2DR	90	90	90	80	80	70	90	100	100	**87.778**
mA2NR	70	90	90	90	80	90	70	100	90	**85.556**
nA2NR	100	80	100	90	100	100	90	90	70	**91.111**
NA2NR	90	70	70	80	90	90	80	90	100	**84.444**
mA3DR	90	90	80	90	90	70	90	90	80	**85.556**
nA3DR	80	70	70	90	80	90	80	70	70	**77.778**
NA3DR	100	90	90	100	90	90	90	100	90	**93.333**
mA3NR	100	90	80	100	90	90	90	80	90	**90**
nA3NR	100	100	100	70	80	90	80	80	90	**87.778**
NA3NR	90	90	80	80	100	90	80	80	100	**87.778**

Table 19: Percent speaker identification for 3 nasal continuants in direct (DR) and network (NR) recordings.

Age Range In years	Recording condition	% SPID for nasal continuant		
		/m/	/n/	/ n̪ /
20≤30	**Direct**	82	89	**93**
	Network	**96**	90	84
30≤40	**Direct**	66	82	**88**
	Network	86	**91**	84
40≤50	**Direct**	86	78	**93**
	Network	**90**	88	88

Table 20: Benchmark for speaker identification.

CHAPTER V

DISCUSSION

The results showed several interesting points. First of all ***percent correct speaker identification for /m/, /n/ and /n̪/ were 82, 89, 93 in the age range of 20≤30 years , 66, 82, 88 in the age range of 30≤40 years, and 86, 78, 93 in the age range of 40≤50 years, respectively for direct recordings. In network recording it was 96, 90, 84 in the age range of 20≤30 years, 86, 91, 84 in the age range of 30≤40 years and 90, 88, 88 in the age range of 40≤50 years using MFCC.*** The results are in consonance with those of Hasan, Jamil, Rabbani, & Rahman (2004), Mao et al., (2006), Wang et al., (2009) etc. Hasan, Jamil, Rabbani, & Rahman (2004) using MFCCs for feature extraction and vector quantization in security system based in speaker identification reported 57.14% speaker identification for code book size of 1, 100% speaker identification for code book size of 16. 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%. Wang et al., (2009) reported that 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. 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/. Tiwari (2010) found 85% of efficiency using MFCC with 32 filters in speaker recognition task i.e. increase in the number of MFCC filters is directly proportional to the improvement in the percent correct speaker identification. Ramya (2011) found 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%. Results of study by Rida (2014) on speaker identification for nasal continuants using MFCC, indicated 90 to 100% speaker recognition in live vs. live recording and 50% to 90% Net work vs. network recording. This study was in Hindi language. Whereas, in the present study is in Kannada language.

Second, it was observed that in direct recording, % Speaker identification (SPID) for /m/ was better in the age range of 40≤50 years compared to that in the age range of 20≤30 years and 30≤40 years. % SPID for /n/ was better in the age range of 20≤30 years compared to that in the age range of 30≤40 years and 40≤50 years; for /n̪/ % SPID was better in the age range of 20≤30 years compared to that in the age range of 40≤50 years and 30≤40 years. In network recording, % SPID for /m/ was better in the age range of 20≤30 years compared to that in the age range of 40≤50 years and 30≤40 years. % SPID for /n/ was better in the age range of 30≤40 years compared to that in the age range of 20≤30 years and 40≤50 years; for /n̪/ % SPID was better in the age range of 20≤30 years compared to that in the age range of 40≤50 years and 30≤40 years. ***No specific age preferences were observed***.

Third, % SPID in direct recoding was78, 83, and 91.33 for /m/, /n/, and /n̪/, respectively. In network recording it was 90.67, 89.67, and 88.33 for /m/, /n/, and /n̪/, respectively. ***% SPID was higher in network recording compared to direct recording*** for /m/ and /n/, and higher in direct recording for /n̪/. Overall % SPID was 84.11 and 88.56 in direct and network recoding, respectively. The results are interesting and contrast the earlier results. The reason as to why % SPID was better in network recording compared to direct recording is unknown and need further investigation.

Fourth, ***% SPID was highest for nasal continuant /n̪ / i.e. 93 for age range 20≤30 years and 40≤50 years whereas 88 for 30≤40 years age range***; in case of network recording samples, the highest score for speaker identification is 96 for 20≤30 years age group and 90 for 40≤50 years age group for nasal continuant /m/; 91 for 30≤40 years age group for the nasal continuant /n/.Percent correct identification is increased in case of network recorded samples. The results indicate that nasal continuant /n̪ / has got highest percent of correct speaker identification score in case of direct recording and /m/ and /n/ has got highest score in case of network recorded samples. It is well known that the nasal consonants are produced with the closure of the oral cavity and radiation of the sound through nasal cavity until the oral obstruction is maintained. Acoustic features of the nasal continuants are nasal murmur, F1 at around 300 Hz, damped formants, wide band widths and formant transitions. Hence, the frequency spectra for nasal continuant varies according to the type of nasal continuant. Bilabial /m/ will show low frequency spectra, dental /n/ shows high frequency and retroflex /n̪/ shows mid frequency spectra.

The results indicate a high bench mark for nasal continuants when MFCC is used. The bench mark is as follows:

Age Range In years	Recording condition	% SPID		
		/m/	/n/	/ n̪ /
20≤30	**Direct**	82	89	93
	Network	96	90	84
30≤40	**Direct**	66	82	88
	Network	86	91	84
40≤50	**Direct**	86	78	93
	Network	90	88	88

The study was restricted to 30 participants and 27 occurrences of nasal continuants and Kannada speakers. Future studies on large number of speakers, in other Indian languages and more number of occurrences of nasal continuants are warranted.

CHAPTER VI

SUMMARY AND CONCLUSIONS

The present study established Benchmark for speaker identification for nasal continuants in Kannada using Mel Frequency Cepstral Coefficients in Kannada. Specifically the objectives of the study were (a) to find out the Mel frequency Cepstral Coefficients for Kannada nasal continuants in direct and mobile recording, thus providing benchmark for speaker identification, and (b) to compare the MFCCs across three age groups of 20≤30 years, 30≤40 years, and 40≤50 years.

Male participants, 10 each in the age range of 20 ≤30 years, 30≤40 years, and 40≤50 years with at least 10 years of exposure to Kannada language as a mode of oral communication were included in the study. Three Kannada nasal continuants - bilabial /m/, dental /n/ and retroflex /n̪/ - as occurring in initial, and medial positions in 30 meaningful Kannada words were selected. Using these words, 10 meaningful 3/4-word sentences were formed to maintain the naturalness of speech. The participant were instructed to speak these sentences under two conditions, directly into the recording mobile (direct) and through another mobile into the recording mobile phone (network) thrice. The speech communicated at the receiving end were recorded and saved in the SD CARD of mobile. Later the .gpp format files were converted to .wav files using Total video converter and Praat software (Boersma and Weenink, 2009) so that analysis could be carried out in an effective manner on a computer. The .wav converted speech sample wave opened with Praat software and identified the words with nasal continuants at word - initial, medial and final positions were identified and segmented. Segmented words were saved as .wav file for each speaker for all the nasal continuants. A total of 27 nasal continuants occurred in these

9 sentences. Thus, the total number of samples for each speaker was 162 (27 * 3 * 2), and the total number of samples stored for 30 speakers were 4860. SSL Work Bench (Voice and Speech Systems, Bangalore, India) was used for analyses. The nasal continuants were segmented. The segmented material was analyzed to extract MFCCs. Further telephone equalization was done. The diagonal matrix and a final percentage for correct speaker identification were obtained. The Euclidian distance between 13 MFCCs was extracted and also within and between participants was noted. 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 results indicated that the percent correct speaker identification for /m/, /n/ and /n̪/ were 82, 89, 93 in the age range of 20≤30 years , 66, 82, 88 in the age range of 30≤40 years, and 86, 78, 93 in the age range of 40≤50 years, respectively for direct recordings. In network recording it was 96, 90, 84 in the age range of 20≤30 years, 86, 91, 84 in the age range of 30≤40 years and 90, 88, 88 in the age range of 40≤50 years using MFCC. It was observed that in direct recording, % Speaker identification (SPID) for /m/ was better in the age range of 40≤50 years compared to that in the age range of 20≤30 years and 30≤40 years. % SPID for /n/ was better in the age range of 20≤30 years compared to that in the age range of 30≤40 years and 40≤50 years; for /n̪/ % SPID was better in the age range of 20≤30 years compared to that in the age range of 40≤50 years and 30≤40 years. In network recording, % SPID for /m/ was better in the age range of 20≤30 years compared to that in the age range of

40≤50 years and 30≤40 years. % SPID for /n/ was better in the age range of 30≤40 years compared to that in the age range of 20≤30 years and 40≤50 years; for /n̪/ % SPID was better in the age range of 20≤30 years compared to that in the age range of 40≤50 years and 30≤40 years. No specific age preferences were observed. Percent SPID in direct recoding was78, 83, and 91.33 for /m/, /n/, and /n̪/, respectively. In network recording it was 90.67, 89.67, and 88.33 for /m/, /n/, and /n̪/, respectively. **%** SPID was higher in network recording compared to direct recording for /m/ and /n/, and higher in direct recording for /n̪/. Overall % SPID was 84.11 and 88.56 in direct and network recoding, respectively. The results are interesting and contrast the earlier results. The reason as to why % SPID was better in network recording compared to direct recording is unknown and need further investigation. Percent SPID was highest for nasal continuant /n̪ / i.e. 93 for age range 20≤30 years and 40≤50 years whereas 88 for 30≤40 years age range; in case of network recording samples, the highest score for speaker identification is 96 for 20≤30 years age group and 90 for 40≤50 years age group for nasal continuant /m/; 91 for 30≤40 years age group for the nasal continuant /n/. Percent correct identification is increased in case of network recorded samples. The results indicate that nasal continuant /n̪ / has got highest percent of correct speaker identification score in case of direct recording and /m/ and /n/ has got highest score in case of network recorded samples. The results indicate a high bench mark for nasal continuants when MFCC is used. The bench mark is as follows:

Age Range In years	**Recording condition**	**% SPID**		
		/m/	**/n/**	**/ n̪ /**
20≤30	**Direct**	82	89	93
	Network	96	90	84
30≤40	**Direct**	66	82	88
	Network	86	91	84
40≤50	**Direct**	86	78	93
	Network	90	88	88

The study was restricted to 30 participants and 27 occurrences of nasal continuants and Kannada speakers. Future studies on large number of speakers, in other Indian languages and more number of occurrences of nasal continuants are warranted.

References

Amino, K., Sugawara, T., & Arai, T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties, *Acoustic Science and Technology*, Vol. 27 (4). 233-235

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

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, The *Journal of the Acoustical Society of America*, Vol. 55, 1304-1312

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

Bennani. Y & Gallinari, P (1991), "On the use of TDNN- extracted features information in talker identification", IEEE ICASSP, 385-388

Boersma and Weeninck.D 2009- Institute of Phonetics Sciences, university of Amsterdam, Retrieved July,20, 2009 from http://www.praat.org/

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

Brouardel P.C.H. (Late 19th century French Medico-Legist)(Reproduced in "Forensic Radiology" by B.G. Brogdon, at page 364. Also quoted in "The American Journal of Forensic Medicine and Pathology", Vol 20, Number 1, March 1999 at page 17, where it is attributed to Paul H. Broussard, Chair of Forensic Medicine, Sorbonne, 1897)

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*.

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

Eatock, J.P. and Mason, J.S, (1994), "A quantitative assessment of the relative speaker discrimination properties of phonemes", Proc. ICASSP, 133-136,

Furui, S. Itakwa, F, & Saito S (1972), Talking recognition by long time averaged speech spectrum , IEEE, 55-A, No. 10, 549-556

Furui, S. (1978). "Effects of long-term spectral variability on speaker recognition". *The Journal of the Acoustical Society of America, 64*, S183 (A).

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

Gish, H. (1985), "Investigation of text-independent speaker identification over telephone channels", IEEE, 379-382

Glen, J. W., & Kleiner, N. (1968). Speaker Identification Based on Nasal Phonation, the *Journal of the Acoustical Society of America*, Vol. 43, 368-372.

Grey CHG, Kopp.GAS (1944), Voice print identification, Bell Telephone Laboratory annual report, New York, pp 1-14.

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

Hazen, B. M. (1973). "Effects of differing phonetic contexts on spectrographic speaker identification". *JASA, 54*, 650-660.

Hecker, M. H. (1971). Speaker recognition. An interpretative survey of literature, *ASHA Monograph*, Vol. 16, 103.

Helms, R.E. (1981), "Speaker recognition using linear predictive vector codebooks", Southern Methodist University

Higgins, AL Bahler, L.G and Porter, J.E (1993), "Voice identification using nearest neighbor distance measure", IEEE ICASSP, 375-378

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

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

Hollien, H. (1990). "The Acoustics of Crime". In P. Rose, 2002, (Ed.), *Forensic Speaker Identification.* Taylor and Francis, London.

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

Hollien, H & Schwartz, R (2000) Aural Perceptual Speaker Identificaiton: Problems with non contemporary samples " *Forensic Linguistics* 7, 199-211.

Hollien, H & Schwartz, R (2001) Speaker identification utilizing non contemporary speech *Journal of Forensic Sciences* 46, 63-37.

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

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

Jakhar, S.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.

Jyotsna. (2011). *Speaker identification using Cepstral Coefficients and Mel Frequency Cepstral Coefficeints in Malayalam nasal Co-articulation. Project of Post graduate Diploma in Forensic Speech Sciences and Technology submitted to University of Mysore, Mysore*.

Kannada language details downloaded from Wikipedia website http://en.wikipedia.org/ wiki/Kannada

Kent, R. D., & Charles, R. (2002). *The Acoustic Analysis of Speech, 2nd Edition.*

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

Kinnunen, T. (2003). Spectral features for automatic text-independent speaker recognition. *Unpublished thesis University of Joensuu, Department of Computer Science. Finland.*

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

Kumar. P, Rao. P, (2004), "A study of frequency-scale warping for speaker recognition", presented at National Conference on Communications, Bangalore.

Lei, H., Lopez-Gonzalo, E. (2009). Importance of Nasality Measure for Speaker Recognition Data Selection and Performance Prediction, in Proc. of Interspeech.

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

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

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

Matsui , Tand Furui, S (1991) "A text-independent speaker recognition method robust against utterance variations", IEEE ICASSP, 377-380

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

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.*

Milner B. (2004). *MAP prediction of pitch from MFCC vectors for speech reconstruction*. Proc. ICSLP.

Miyajima, T (2001), "A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction", *Speech Communication*, 203 –218

Naik, J. (1994). "Speaker verification over the telephone network: database, algorithms and performance assessment". In P. Rose, 2002, (Ed.), *Forensic Speaker Identification.* Taylor and Francis, London.

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

Nolan, F. (1997). "Speaker recognition and forensic phonetics". In P. Rose, 2002, (Ed.), *Forensic Speaker Identification.* Taylor and Francis, London.

Oglesby, J & Mason , J. S (1991), "Radial basis function networks for speaker recognition", IEEE ICASSP, 393-396

Orman,N (2000), "Frequency Analysis of Speaker Identification Performance", M S Thesis, Electrical and Electronics Engg. Bogazici University, Turkey

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.*

Pamela, S. (2002). *Reliability of voice prints*. Masters dissertation, University of Mysore, Mysore.

Pottter, R (1945), Visible pattern of Speech Sciences 102: 463-470

Potter, R. K. (1946). "Introduction to technical discussions of sound portrayal". *JASA,* 18, 1-3.

Pollack, I., Pickett, J. M. and Sumby, W. A. (1954). "On the identification of speakers by voice". *JASA,* 26, 403-406.

Plumpe, M. D., Quatieri, T. F., Reynolds, D. A. (1999). Modeling of the glottal flow waveform with application to speaker identification, *Proc. IEEE 7,* 569- 586.

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

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

Ramya. B.M. (2013). *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.*

Rao, Y.H , Rajasekaran P.K and Baras, J.S (1992), "Free-text speaker identification over long distance channel using hypothesized phonetic segmentation", IEEE ICASSP, 177-180

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

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

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,* Vol. 26, 403-406.

Reich, A. R. (1981). Detecting the presence of vocal disguise in male voice. *Journal of Acoustical Society of Americ*a, Vol. 69, 1458-1461

Reynolds. D. A, (1995), "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication,* 91-108.

Rudasi , L and Zahorian, S.A (1991), "Text-independent talker identification with neural networks", IEEE ICASSP, 389-392

Richard Safferstien, (1994) Criminology and Forensic Science Prentice-Hall; Fifth edition edition (1994) chapter document and voice examination

Rida, Z, A., (2014). *Benchmarks for speaker identification using nasal continuants in Hindi in direct mobile and network recording. Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysore.*

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

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

Sambur, M.R (1975), "Selection of Acoustics Features for Speaker Identification", *IEEETrans. Acoustics, Speech and Signal Processing*, 176-182.

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

Shuzo Saito, Kazuo Nakata, Fundamentals of Speech Signal Processing. Tokyo (1985) 74-83.

Sir Arthur Conan Doyle. (2015). The Biography.com website. Retrieved 04:31, Jul 27, 2015, from Arthur Conan Doyle. (2015). The Biography.com website. Retrieved 04:31, Jul 27, 2015, from http://www.biography.com/people/arthur-conan-doyle-9278600.

Soong, F.K, Rosenberg, A.E , Rabiner and Juang , B.H (1985), "A vector quantization approach tospeaker recognition", IEEE ICASSP, 387-390.

Speech Science Lab, developed by Ananthapadmanabha, T.V, 2015, CEO, Voice & Speech System, Bengaluru, India.

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.

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

Steve Cain – (2015)Tape Expert President - Forensic Tape Analysis, Inc., - Forensic Audio/Video Tape Examiner/ Examiner of Questioned Documents, Lake Geneva, WI and Diplomat and Fellow, American

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*, Vol. 56,1876-1883.

Thevenaz, P & Hugli H (1995), Usefulness of the LPC-residue in text independent speaker verification, *Speech communication*, 17, 145-157.)

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

Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies* 1(1): 19-22. Retrieved from http://www.researchtrend.net/ijet/4_Vibha.pdf on 29.4.2014.

Tiwari, R., Mehra, A., Kumawat, M., Ranjan, R., Pandey, B., Ranjan, S. & Shukla, A. (2010). Expert system for speaker identification using lip features with PCA. *Intelligent Systems and Applications* (ISA), 2010 2nd International Workshop, 1-4.

Tosi, O., Oyer, H. J., Lashbrook, W., Pedrey, C., Nicol, J., Nash, E. (1972). Experiments on voice identification. *The Journal of the Acoustical Society of America*, 51, 2030-2040.

Wang, L., Ohtsuka, S., & Nakagawa, S. (2009). High improvement of speaker identification and verification by combining MFCC and phase information, *Proc. IEEE* 4529- 4532.

Wolf, J. J. (1972), Efficient acoustic parameter for speaker recognition, *The Journal of the Acoustical Society of America* 51, 2044–2056.

Yarmey, A.D (1991) Description of distinctive and non distinctive voices over time. *Journal of Forensic Sciences society*, 31, 421-428.

Yarmey A.D & Matthys.E (1992), Voice identification of an abductor, *Applied Cognitive Psychology*, 6 367-377.

Young, M. A. & Campbell, R. A. (1967). "Effects of context on talker identification". *The Journal of the Acoustical Society of America*, 42, 1250-1254.