BENCHMARKS FOR SPEAKER IDENTIFICATION USING COTINUANTS IN HINDI DIRECT MOBILE AND NETWORK RECORDING

Zakia Rida Ansari

Register Number: 12SLP031

A Dissertation Submitted in Part Fulfilment of Final Year

Master of Science (Speech Language Pathology)

University of Mysore, Mysore.

ALL INDIA INSTITUTE OF SPEECH AND HEARING

MANASAGANGOTHRI, MYSORE – 570006

MAY, 2014.

CERTIFICATE

This is to certify that this dissertation entitled "**Benchmarks for speaker identification using nasal cotinuants in Hindi in direct mobile and network recording**" is a bonafide work submitted in part fulfilment for the Degree of Master of Science (Speech Language Pathology) of the student (Registration No.: 12SLP031). This has been carried out under the guidance of a faculty of this institute and has not been submitted earlier to any of the University for the award of any other Diploma or Degree.

Mysore

May, 2014

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

CERTIFICATE

This is to certify that this dissertation entitled "**Benchmarks for speaker identification using nasal cotinuants in Hindi 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
Guide & Director
All India Institute of Speech and Hearing
Manasagangothri, Mysore -570 006.

DECLARATION

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

Mysore

May, 2014.

Register No.: 12SLP031

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

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. This is a vibrating system which is capable of producing acoustic signals. The acoustic signal makes use of the elastic property of air where it travels in the same manner as ripples carried in a pool away from the point of impact of a pebble dropped. The ripples carried to our ear, can be 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)

A voice is more than just a string of sounds. Identifying people on the basis of their voice is a common phenomenon. It is a media through which we identify other humans known to us like members of our family, friends, popular figures etc. We infer this information from the tone of the voice, rate of speaking, style of speaking etc. This is additional information apart from the intended linguistic message. Other characteristics of the individual can also be known by listening to their voice even if they are unfamiliar to us- age, gender, language, emotional state and so on. The complexity arises when these characteristics are to be distinguished. It is a difficult task because the linguistic and additional information is convolved together

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)

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 voice disguise. Out of the many possibilities available to an individual for voice disguise, falsetto, whisper, change in speaking rate, imitation, pinched nostrils and object in the mouth are popular favourites of perpetrators (Ramya, 2013)

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 in the form of speaker identification or speaker verification. Speaker identification, simply put, is the identification of a particular speaker from a group of unknown speakers. It demands 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.

Speaker verification refers to verifying if a particular voice sample of a particular 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.

The other classification of speaker recognition is *text- dependent and text-independent.* In the latter, voice characteristics are analyzed from the sample recording irrespective of the linguistic content of the recording. In the former, the identification is based on the speaker speaking a particular phrase like a password, pin code etc. (Rabiner, 1993). Each of the techniques employed have to be assessed for their advantages and disadvantages and these have to be considered. Whether text dependent or text independent, the choice of which technique to use is application-specific. At the highest level, all the modules contain two processes, feature extraction and feature matching, in the same order as stated.

Another group of problems that maybe faced by the analyst is system distortions and speaker distortion. System distortion maybe the result of reduced 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 material used for recording and storing, like microphones with limited capability or poor quality tape recorders, the information have a reduced frequency range. This may result in the loss of speaker characteristics which may be irrecoverable later (Hollien, 1990)

Criminals experience fear, may be anxious or stressed when they commit the crime which does give a different character to their speech. Ingested drugs, alcohol or even a cold can change the way a voice sounds in a recording. On the other hand, criminals rarely cooperate while providing the exemplars. Some may even try to disguise their

voice or may simply refuse to provide the sample. These are referred to as speaker distortions (Hollien, 1990)

The frequency with which the correct speaker identification is carried out gets 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. Hence, the present study will be focusing on usefulness of Mel frequency cepstral coefficients (MFCC) on speaker recognition.

Atal (1974) studied several different parameters using linear prediction model for their effectiveness for automatic recognition of speakers from their voices. The predictor coefficients, the autocorrelation function, the area function and the cepstrum function were used as input to an automatic speaker recognition system. The data consisted of six repetitions of 60 utterances by ten speakers. Result supported cepstrum as the most

effective parameter, providing an identification accuracy of 70% for 50 ms long speech, which increased to more than 98% for duration of 0.5s with the same date, verification accuracy was found to be in the whereabouts of 83% for duration of 50 ms increasing to 95% for duration of 1sec. The cepstrum is used as the inverse Fourier transform of the log magnitude Fourier spectrum. It is used to separate the transfer function and the excitation signal which exists in the low frequency and high frequency, respectively. The coefficients that make up the resulting cepstrum are known as the cepstral coefficients.

Most of the studies (Reich & Duke, 1979, Reich, Moll, & Curtis, 1976) that review effective disguise for speaker identification state that nasal disguise and slow rate of speech are the least effective disguises. Therefore, nasal continuants would be the best speech sounds to investigate speaker identification under disguise.

This brings us to the need of the study. Any expert should be competent in his/her field (Hollien, 2002). A certain amount of training which should be appropriate and advanced has to be taken in the field. For this reason, it is important to have studies that can aid the training as well as the analysis of experts so that when the need arises for them to provide a result, they can substantiate it with the general trend that has been studied in the past. 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. It is for this reason that a method to carry out these analyses becomes a must.

In this context, the present study examined speaker identification using nasal continuants in Hindi (*Hindi, an Indo-Aryan language, is one of the official languages of India. In the 2001 Indian census, 258 million people in India reported*

Hindi to be their native language. This makes Hindi approximately the sixth-largest language in the world) using Mel Frequency Cepstral Coefficients (MFCC).

Thus, the aim of the study was to establish Benchmark for speaker identification using nasal continuants in Hindi in direct mobile and network recording using Mel frequency Cepstral coefficients (MFCC). The objectives of the study were to provide benchmarks for (a) Mel-frequency cepstral coefficients for Hindi nasal continuants and (b) compare these in mobile and network recording conditions.

CHAPTER II

REVIEW OF LITERATURE

The ability to identify a person on the basis of their voice has long been investigated. For many years law enforcement personnel have tried to use forensic speaker identification to incriminate or confirm guilt or innocence of a suspect.

Hecker (1971) described three methods for speaker identification, namely

a) aural/ perceptual- listening to the speech

b) Spectograms- visual examination

c) Semi automatic identification by machines

Recognition of an individual from a forensic quality recording can prove to be an extremely challenging task. The methods used for the analysis can be automatic, semi-automatic or human based. The material of the recording may also differ significantly ranging from a yelling on the telephone to a whisper recording, recording under stress, drugs, sickness or disguise, and recording in the presence or absence of noise. These unknown and known variables make the discrimination between speakers a complicated and daunting task.

Kersta (1962) introduced the term *voiceprint identification.* This term has become the bane of a forensic analyst's existence as it equated with the fingerprint and is in vogue due to media. Kersta analyzed the spectograms of five clue words spoken in isolation using 12 talkers and closed set identification. Five days of training was given to high

school girls to identify the talkers from the spectrograms on the basis of eight ‘unique acoustic cues’. Results of this study indicate a high rate of identification accuracy which was inversely proportional to number of talkers. For 5 talkers, the identification rate was 99.6%. For 9 talkers, the rate was 99.2% and for 12 talkers, it was 99%. Another finding of the study was that ‘bar prints’ give a better identification score than ‘contour prints’.

This high estimation of correct identification has not been replicated in other studies. Scores reported by Kersta are 99%-100%, for short words spoken either in isolation or in context, when compared to (a) 81%-87%, for short words spoken in isolation (Bricker & Pruzansky, 1966), (b) 89% for short words taken from context (Pruzansky, 1963), and (c) 84%-92%, for short words spoken in isolation (Pollack, Pickett, & Sumby, 1954).

Differences in anatomy, physiology and acoustics are always present. A combination of these factors makes no two speakers the same. Even identical twins who may sound similar acoustically may have different implementation of single segment in their linguistic system.

Atal (1976), offers “any decision-making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance.” Another area wherein Atal’s characterization is that “it strongly suggests that an unambiguous, categorical outcome is expected: the person is either determined to be or determined not to be the speaker of a given utterance. In the forensic case the outcome should be a ratio of probabilities” (Rose, 1990). Figure 1 shows the schematic representation of speaker recognition.

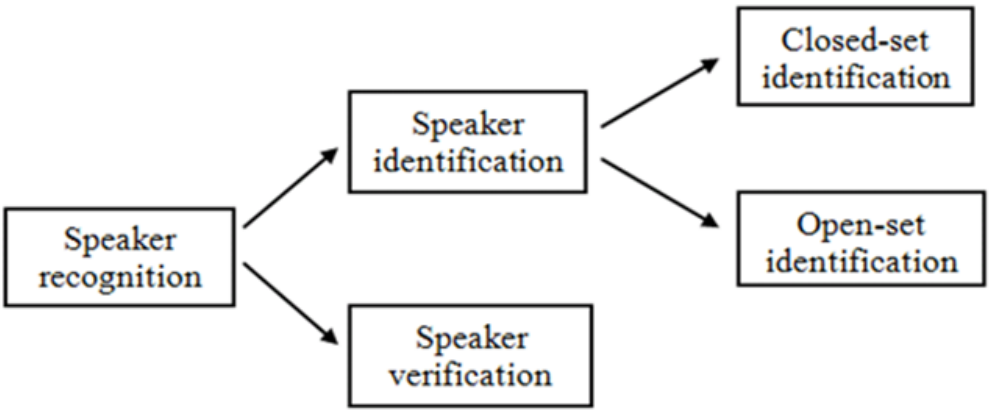


Figure 1: Schematic representation of speaker recognition.

Speaker identification, simply put, is the identification of a particular speaker from a group of unknown speakers. It demands 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. In Figure 2 below, the unknown sample on the left hand side is has to be compared with the known speaker 1 (A) and then with the next known speaker 1 (B) and so on. The question mark shows the question, "does the unknown sample match that particular known sample?" if the unknown sample matches any of the known samples, say the sample 3, then, the outcome shows that the unknown sample is identified as speaker C.

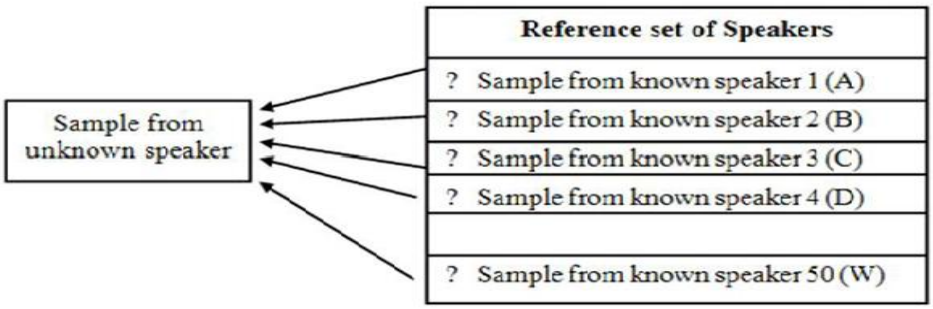


Figure 2: Illustration of speaker identification.

In speaker identification, the reference set of known speakers can be closed or open. Closed set refers to the process where the unknown speaker has to be identified from the given samples where the examiner has the knowledge that the speaker of the unknown sample is in the samples that have been given to him. An open set, on the other hand, refers to the process where the examiner is not aware if the speaker of the unknown sample is in samples that have been provided or not. This makes a closed set identification a lot easier than an open set identification. Since it is known that the unknown voice is one of the reference set, the closed set identification task lies in (a) estimating the distance between the samples of the unknown speaker and each of the known reference speakers, and (b) picking the known speaker using the sample who is separated by the smallest distance from the unknown speaker. The pair of sample separated by the smallest distance is assumed to be from the same speaker (Nolan, 1983). As speaker identification requires automatic selection of the unknown speaker from the samples given by selecting the one with the shortest distance from the test sample, it requires no threshold establishment. In the field of forensics, both closed and open set can occur but it has been seen that the frequency of open set is much more than that of closed set. Since the task becomes very much simpler with a closed set, the distinction between open and closed set is an important one in forensic speaker identification.

In speaker identification, the examiner can give only two responses, either the unknown speaker is among the test samples or it is not. Verification is a more complicated process, with four types of decision. The decision can be correct in two ways: the speaker is identified as being who they say they are, or not being who they say they are. And it can be incorrect in two ways: the claimant has been identified as

being who they say but it is not so and the claimant has been identified as not being who they say they are but it is so.

In the open set speaker identification task three types of errors are possible. Figure below shows the schematic representation of classification of errors. (a)Error A: A match did exist but the examiner selected the wrong choice (false identification), (b) Error B: A match did exist but the examiner failed to recognize it (false elimination), and (c) Error C: A match did not exist although the examiner selected one of the choices (false identification). Figure 3 shows types of errors in speaker identification.

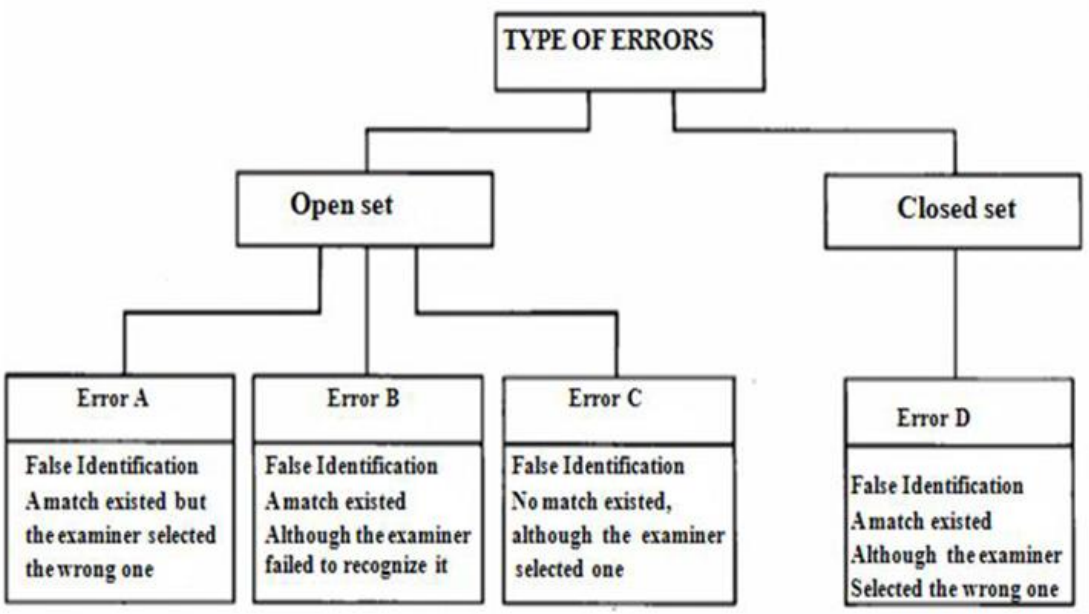


Figure 3: Types of errors in speaker identification.

Speaker verification, simply put, refers to verifying if a particular voice sample of a particular 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. The schematic representation of this procedure is shown in Figure below. Here speaker D wants to access and be verified. The system has samples

of speaker D`s voice in its storage, which it retrieves and compares with that of the sample provided by speaker D. If the two voice samples are judged similar enough, speaker D`s claim is verified and he is given the access.

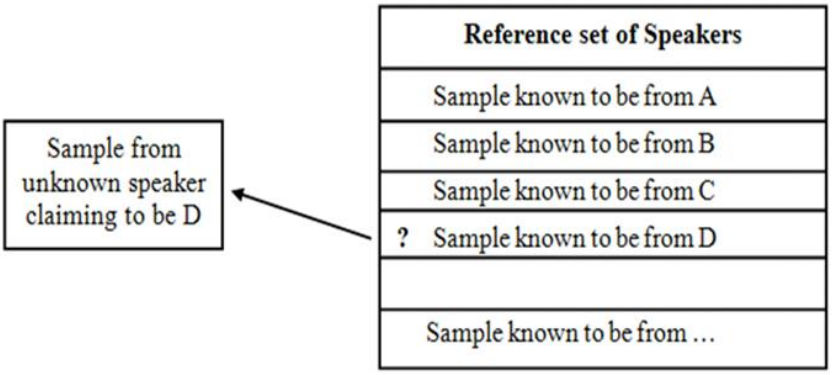


Figure 4: Illustration of speaker verification.

Campbell, Matrouf, Schwartz, Campbell, Wade, in 2009, state that another factor that plays a role in speaker identification is performance variability. Figure 5 shows the differences that can creep in the recognition if the time elapsed between the enrollment and test sample is too great. This has been referred to as the effect of voice aging on the different softwares used for speaker recognition.

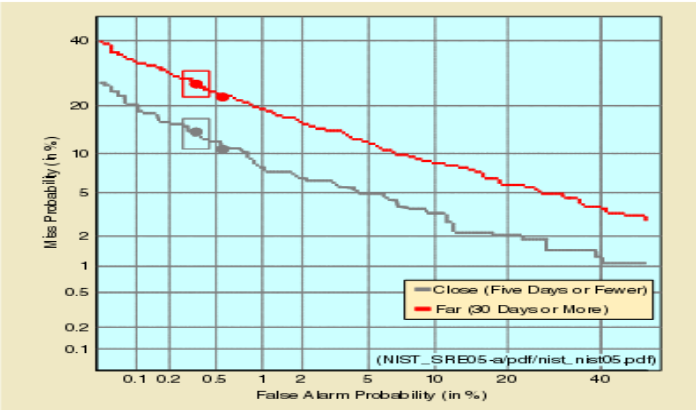


Figure 5: False alarm and miss probability (from Campbell et. al., 2009).

Another hurdle faced by the forensic analyst is the lack of availability of the speech material for analysis. Only short segments are present for training as well as analysis making the context in which speaker recognition has to be carried out very difficult. In 2007, Fuave, Evans, Pearson, Bonastre and Mason investigated the effect of short durations on a GMM-UBM baseline system and on a GSL-NAP system, using the ALIZE/SpkDet software. They found that the EER of the GMM-UBM increases about a factor of 3 when only the duration of both training and testing speech excerpts is 10 s (the most difficult situation).

Methods of speaker identification

Several studies have been reported on speaker identification using the listening method. In a study by McGehee (1937), five male voices were provided to the listeners and their task was to select a single target voice from the samples given. These were attempted after delays that ranged from 1 day to 5 months. The correct identification was seen to steadily decline from 83% after 1 day to 80.8% after a week, 68.5% after 2 weeks, 57% after a month, and to 13% after 5 months. Thompson (1985) used male voices in a six-voice line up task. The task of the listeners was to identify a voice that they had heard one week previously from the voices presented. The listeners were allowed to say that the voice that they had heard was not in the line up or that they were unsure if it was. However, the listeners could not say that the voice that they had heard previously was in the line up more than once. Thus, from the viewpoint of the listeners, the experiment was not an independent-judgment task even though it was an open set task. Such a task can be considered as an open-set, multiple-choice task with a decision threshold by the listener. The results were 62.1% correct identifications,

22.1% incorrect identifications, and 15.8% “not in line up" or "not sure if in line up" response.

Schwartz (1968) reported speaker identification of gender using voiceless fricatives like /s/, /ʃ/, /f/, /th^{h}/. A recording of nine females and nine males was made saying these fricatives in isolation. The stimulus was presented via a loud speaker to ten listeners in a random order. Their task was to identify the gender after listening. The results that were obtained said that the listeners could identify the gender of the speakers after listening to the isolated production of /s/ and /ʃ/. However this could not be done with accuracy from the isolated productions of /f/ and /th^{h}/. Subsequent spectrographic analysis of the /s/ and /ʃ/ stimuli revealed that the female spectra tended generally to be higher and parallel in frequency compared to that of male. Ingemann, in the same year, 1968, support the above results and reported that *listeners often identify the sex of a speaker from hearing voiceless fricatives in isolation and sex was better identified on fricative /h/.*

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

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

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

The term Voiceprint was introduced by Lawrence Kersta (1960); who studied if the patterns on sonograms exhibited features which could be used to identify speakers. He published a paper in 1962 on "voice identification" in which he initiated an erroneous idea that there is a close relationship between finger print and voice print. Kersta's identification method human observers visually matching spectrogram and to duplicate his investigation with what we believe are methodological and analytical improvements. The term, voiceprint, has become the bane of a forensic analyst's existence as it equated with the fingerprint and is in vogue due to media. Kersta analyzed the spectrograms of five clue words spoken in isolation using 12 talkers and closed set identification. Five days of training was given to high school girls to identify the talkers from the spectrograms on the basis of eight 'unique acoustic cues'. Results of this study indicate a high rate of identification accuracy which was inversely proportional to number of talkers. For 5 talkers, the identification rate was 99.6%. For 9 talkers, the rate was 99.2% and for 12 talkers, it was 99%.

Another finding of the study was that ‘bar prints’ give a better identification score than ‘contour prints’.

This high estimation of correct identification has not been replicated in other studies. Scores reported by Kersta are 99%-100%, for short words spoken either in isolation or in context, when compared to (a) 81%-87%, for short words spoken in isolation (Bricker & Pruzansky, 1966), (b) 89% for short words taken from context (Pruzansky, 1963), and (c) 84%-92%, for short words spoken in isolation (Pollack, Pickett, & Sumby, 1954).

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

Findings of a large scale study by Tosi and colleagues in 1972 were published. In these studies, an imitation of law enforcement conditions was made for identification. These were presented for spectrographic analysis only with no aural confirmation. An experiment which ran the course of two years was carried out. Using spectrograms, voice identification was done with the two-fold goal of a) checking Kersta's (1962) claims in this matter and b) testing models including variables related to forensic tasks. 25000 males speaking general American English were used as the population from which 250 participants were selected in a homogeneous manner. All of these were students at Michigan State University. 34996 experimental trials, for speaker

identification were carried out by 29 trained examiners. 10 - 40 known voices were included in each trial. These were in various conditions: contemporary and non-contemporary spectrograms, with closed and open trials, in a fixed context and in a random context, nine or six clue words spoken in isolation etc. A positive decision was asked to be made by the examiners giving them a time of 15 minutes. They had to either identify or eliminate, no other option was present to them. Their decisions were based solely on inspection of spectrograms; listening to and identification by voices was excluded from this experiment. A 4-point rating scale was used to judge the confidence level of the examiner in the task (1 and 2, uncertain, 3 and 4, certain).

Results of this experiment reinforced the results obtained by Kersta with her experimental data. Experimental trials of this study, correlated with forensic models, approximately 6% false identifications and approximately 13% false eliminations were noted to be the error in the study. This means that approximately 60% of their wrong answers and 20% of their right answers were judged as "uncertain" by the examiners. Main differences of conditions that could exist between models and real cases were hypothesized to be as follows:

(1) Population of known voices: In forensic cases, the catalog of known voices could theoretically include millions of samples. In these cases the catalog of known voices is open, true, but it is 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 decision. In addition, he is aware of the consequences that a wrong decision could mean to his

professional status as well as the consequences meant to the speaker whom he might erroneously identify. Availability of time and responsibility between experimental and professional conditions might help to improve the accuracy of the professional examiners.

(3) Type of decisions which the 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 have uncertainty 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 suspects, but more evidence is necessary in order to reach a positive identification; (d) Possibility that the unknown speaker is none of the available suspects, but more evidence is necessary to reach a positive elimination; and (e) Unable to reach any conclusion with the available voice samples. These possibilities of alternative decisions 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 six or nine clue words were available to the examiners for visual inspection. Rather, a professional examiner is entitled to request as many samples as he deemed necessary to reach a positive conclusion. In real forensic cases the professional examiner must 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 comparison enhances the accuracy of voice identifications.

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

Most of the speaker identifications are conducted in laboratory condition. The results may differ in actual forensic conditions.

Speaker identification by machine (objective method)

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

Automatic speaker verification was accomplished by Luck (1969) using cepstral measurement. The phrase 'My code is' was used to characterize short segments in each of the first two vowels. Additional parameters were also assessed like the speaker's pitch and the duration of the word 'my'. Like identification, verification also presents as having a black and white decision- the claimant is the authorized speaker or he is not.

A comparison of the reference data with the authorized speaker is carried out. This shows that if the reference data is collected over a period of time, say many days, then verification can be done as late as two months after the collection of the sample, whereas, if reference data was collected at one sitting, verification would be very

inaccurate as little as 1 h later. Four authorized speakers and 30 impostors were examined, with error rates obtained to 6% to 13%. When individuals tried to deceive the system by acting as impostors of the authorized speaker, they could not do so. It has been observed by many who have seen the system in operation that greater accuracy would be obtained if a *final decision were based on a series of two or three repetitions of the test phrase*. This is to say that ***increased accuracy depends on increasing the information available to the decision mechanism.***

Wolf (1972) suggested relations between the voice signal and vocal-tract shapes and gestures as an efficient approach to selecting parameters. It is desirable to use acoustic parameters for mechanical recognition of speakers as that are closely related to voice characteristics that distinguish speakers. Instead of measuring the entire utterance and giving general parameters, only the significant features of selected segments are used. Speech events can be manually located within the utterance after feeding it into a simulated speaker recognition system and then measuring the parameters at these locations to classify the speakers. Useful parameters found were word duration, features of vowel and nasal consonant spectra, F0, estimation of glottal source spectrum slope and voice onset time (VOT).

Atal (1972) examined the temporal variations of pitch in speech as a speaker identification characteristic. 60 utterances spoken by 10 speakers, consisting of six repetitions of the same sentence were recorded for analysis of the pitch. The pitch contours were linearly transformed so that the ratio of inter-speaker to intra-speaker variance in the transformed space was maximized. Again, the speaker sample with the least distance in the reference vector was taken as the correct identification. 97% of

correct identification was reported which led to temporal variations of pitch being suggested as a good and effective parameter for automatic speaker recognition.

In another experiment, Atal (1974), used linear prediction model, to assess its effectiveness in automatic recognition of speakers from their voices. He determined that speech sampled at 10 kHz, had predictors at approximately every 50ms giving a total of 12 predictor coefficients. The predictor coefficients, like the autocorrelation function, the impulse response function, the cepstrum function and the area function were used as input to an automatic speaker-recognition system. The data and method were the same as the previous study where 6 repetition of the same sentence spoken by 10 speakers was analyzed. For verification, the speaker was verified if the distance between the test sample vector and the reference vector for the claimed speaker was less than a fixed threshold amount. He reported that the **cepstrum was found to be the most effective parameter**, providing an identification accuracy of 70% for speech 50 ms in duration, which increased to more than 98% for an increase in duration to 0.5 sec. Using the same speech data, the verification accuracy was found to be approximately 83% for a duration of 50 ms, increasing to 98% for a duration of 1sec.

The results of this research suggest several conclusions. Firstly, it may be concluded that n- dimensional Euclidean distance among long-term speech spectra may be quite successfully utilized as criteria for speaker identification, at least under laboratory conditions. Moreover, this method displays a number of advantages: (a) It is relatively simple to carry out; (b) it eliminates such crucial factors as the time-alignment problem; (c) the data generated for the identifications does not depend on the overall power level of the speech samples used; and (d) the process does not depend on a human and, hence, subjective judgments. Finally, it appears that distortions created by

limited pass band and stress as these two factors are defined in these experiments have only minimal effects on the sensitivity of the LTS vector as a speaker identification cue.

Glenn & Kleiner (1968) describe an experiment involving identification based on the spectrum of nasal sounds in different environments in test and reference data. If just one speaker sample was correlated with the thirty reference vectors, a correct identification rate of 43% was obtained. This increased to 93% if the average of 10 speaker samples was used for correlation and further increased 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 segments of an utterance, in this case nasal phonemes. It is noted by the authors that no account was taken of the phonetic environment of the nasal phoneme. If the test had been restricted to exponents of /n/ in a single environment, or if the effects 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.

Pamela (2002) investigated the reliability of voiceprints by extracting acoustic parameters in the speech samples. Six normal Hindi speaking male participants in the age range of 20-25 years participated in the study. Twenty-nine bisyllabic meaningful Hindi words with 16 plosives, five nasals, four affricates and four fricatives in the word-medial position formed the speech material. Subject read the words five times. The results indicated no significant difference in F2, onset of burst and frication noise, F3 transition duration, closure duration, and phoneme duration between subjects. However, the results indicated high amount of intra-subject variability. High

intra-subject variability for F2 transition duration, onset of burst, closer duration, retroflex and F2 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 greater 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 belong to 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 made.

Reich & Duke. (1979) describe another experiment involving the effects of selected vocal disguises upon speaker identification by listening. The results of this experiment suggested that certain 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 in 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 (naïve listeners) and hoarse voice (sophisticated listeners) were the least effective disguises on the speaker identification by listening.

The power spectra of nasal consonants (Glenn and Kleiner, 1968) and coarticulated nasal spectra (Su; Li and Fu,, 1974) provide strong cues for the machine matching of speakers. It is interesting to know the listeners in the present study were unable to successfully utilize these seemingly speaking dependent cues.

Mel Frequency Cepstral Coefficients (MFCC)

Psychophysical studies of the frequency resolving power of the human ear has motivated modeling the non-linear sensitivity of human ear to different frequencies. The selective frequency response of the basilar membrane (hair spacing) acts as a bank of band pass filters equally spaced in the Bark scale. Figure 6 shows the linear spacing between 100 Hz to 1 kHz and the logarithmic spacing above 1 kHz further reduces dimensionality of frame/vector of speech. The low-frequency components of the magnitude spectrum are ignored and the useful frequency band lies between 64 Hz and half of the actual sampling frequency. This band is divided into 23 channels equidistant in Mel frequency domain. 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.

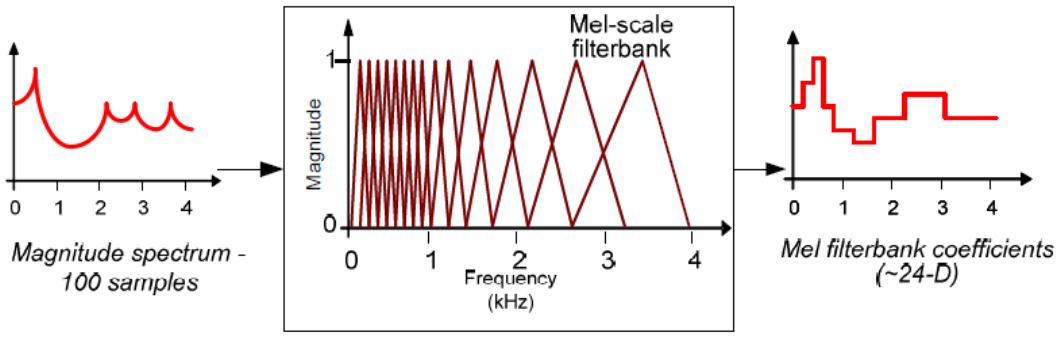


Figure 6: Illustration of Mel Frequencies and their coefficients.

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.

When MFCC feature extractor is used in speaker recognition system, one makes an implicit assumption that the human hearing mechanism is the best speaker recognizer. Authors aimed to find the critical parameters that affect the performance and tried to give some general guidelines about the analysis parameters. They conducted experiments on two speech corpora using vector quantization (VQ) speaker modeling. The corpora were a 100 speaker subset of the American English TIMIT corpus, and a Finnish corpus consisting of 110 speakers. Although noise robustness is an important issue in real applications, it is outside the scope of the thesis. The author's main attempt was to gain at least some understanding of what is individual in the speech spectrum. 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 on speaker identification. Database consists of 21 speakers, which included 13 males and 8 female speakers. Study showed 57.14% speaker identification for code book size of 1, 100% speaker identification for code book size of 16. Study reveals MFCC technique has been applied for speaker identification.

Chandrika (2010) compared the performance of speaker verification system using MFCCs while recording with mobile handsets over a cellular network as against digital recording. Ten subjects who participated in the study were provided with words containing long vowels /a: /, /i: / and /u: /. Speakers were provided with CDMA handset (Reliance, LG). MFCC values were extracted from the speech samples

obtained. Results revealed that the overall performance of speaker verification system using MFCCs is about 80% for the data base considered. The overall performance of speaker recognition is about 90% to 95% for vowel /i/. The accuracy of performance for vowel /i/ is marginally better compared to vowel /a/ and /u/.

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

Previous work suggests that nasal regions of speech are an effective speaker cue, because the nasal cavity is both speaker specific, and fixed in the sense that one cannot change its volume or shape. Various acoustic features have been proposed for detecting nasality. Glass and Zue (1995) used six features for detecting nasalized vowels in American English. Pruthi and Espy-Wilson (2007) extended Glass's work and selected a set of nine knowledge-based features for classifying vowel segments into oral and nasal categories automatically.

The review indicates that the effects of vocal disguises markedly interfere with spectrographic speaker identification as well as speaker identification by listening. In this context, the present study examined speaker identification using nasal continuants in Hindi using Mel Frequency Cepstral Coefficients (MFCC). The aim of the study was to establish Benchmark for speaker identification using nasal continuants in Hindi in direct mobile and network recording using Mel frequency Cepstral coefficients (MFCC). The objectives of the study were to provide benchmarks for (a) Mel-frequency cepstral coefficients for Hindi nasal continuants and (b) compare these in mobile and network recording conditions.

Chapter III

Method

Participants: Ten participants between the age range of 20 to 40 years with at least 10 years of exposure to Hindi language as a mode of oral communication were included in the study. The inclusion criteria of the speakers was

d) no history of speech, language and hearing problems,

(b) normal oral structure,

(c) no other associated psychological or neurological problem and

(d) reasonably free from cold and other respiratory illness and oral restructuring at the time of recording.

Stimulus: Commonly occurring forensically related Hindi meaningful words with nasal continuants – bilabial /m/, dental /n/ and palatal /ŋ/ - were selected. The nasal continuants were added in the initial, medial or final positions as may be the requirement of the forensically related word. These were embedded in 3-4 word sentences to maintain the naturalness of speech. There were a total of 5 /m/, 8 /n/, and 1 /ŋ/. Sentences used were as follows:

1) /**M**Udʒ^{h}e pæse ʧa:hIje/

2) /pulis ko **m**aţ baţa:**n**a/

3) /**n**ahi ha**m**e la:k^{h} de**n**a/

4) / agle fo**n** ka i**nţ**Eza:r kar**n**a/

5) /hu**m** larkI wa:pIs đE**ŋ**e/

6) / ţi**n** sE p**ãʧ** kE biʧ **mɛ̃** a**n**a/

Procedure: Speech samples of participants were recorded individually. Participants were informed about the nature of the study. Sentences were written on a card which was visually presented to the participants. Participants were instructed to read that sentences twice in a normal rate of speech. They were instructed to speak under two conditions, directly into the recording mobile (live) and through another mobile into the recording mobile phone (network). Each participant was instructed to utter the stimulus 3 times at an interval of 1 minute. The network used for making the calls was Vodafone (GSM 900/ GSM 1800 MHz frequency) and the receiving network was also Vodafone on a Sony Ericsson Xperia pro mobile phone. A speaker participating in an experiment was given a Vodafone on a Sony Ericsson Xperia pro mobile phone. A call was made to the speaker's handset from another Vodafone on a Sony Ericsson Xperia pro mobile phone with recording option held by the experimenter. Speech signal was recorded as the speaker uttered the test sentences. Later the recorded sentences were uploaded to a computer for further analysis. The message at the receiving end were recorded and saved by the experimenter in the microchip of Sony Ericsson Xperia Pro. The live recordings were made using a free software Smart Voice Recorder version 1.6. This recorded and stored the files in .wav format and hence did not require conversion. The network recordings were changed from .mpeg format to .wav files using an online downloader on the website Zamzar (www.zamzar.com) so that analysis can be carried out in an effective manner on a computer. All the files were opened in Praat software (Boersma and Weenink, 2009) and down sampled to 8 kHz.

Segmentation: The down sampled speech material was segmented manually using PRAAT software to obtain the nasal continuants in all positions of the target words. Figure 7 shows a segment of speech signal.

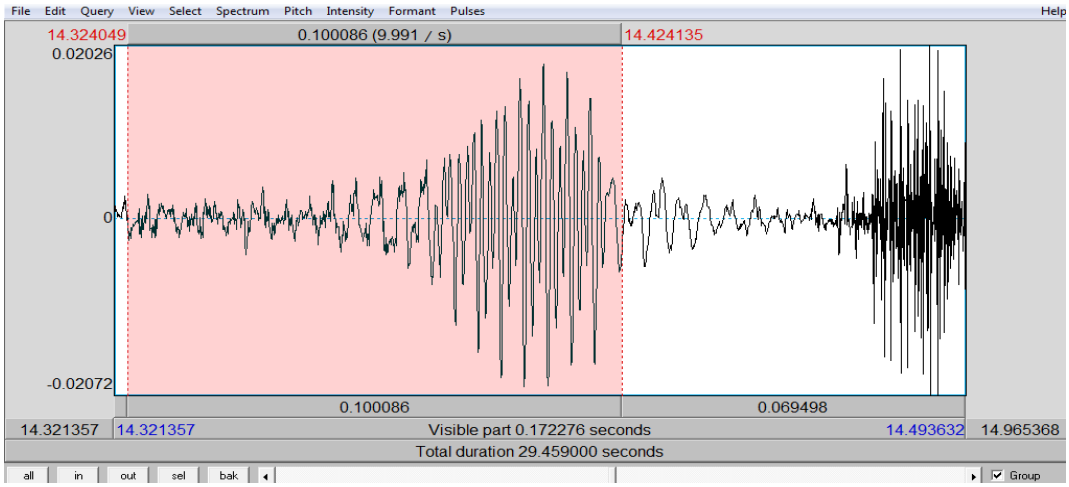


Figure 7: A segment of speech signal.

The segmented nasals were saved using a particular file name convention. For Ex: For speaker 1, first sample, first session, first occurrence was given the file name as ***"(speakers name)_call_1m.wav*** and saved in a folder with the name ***spk1***. There were 84 sample files (14 nasals * 3 repetitions * 2 conditions = 84) for each speaker. Similar pattern was followed for other participants. Converted samples were stored in separate folders for each participant and separate folders for each repeated recording (14 samples each). These were stored separately in two main folders by the name 'direct' and 'network' recordings.

Analyses: Analyses of the data was carried out using SSL Work Bench (Voice and Speech Systems, Bangalore, India). The nasal continuants were analyzed using SSL Work Bench at a sampling frequency of 8 kHz, to extract and compare its Mel frequency Cepstrum Coefficients (MFCC). Initially the file was specified using a

notepad and .dbs file that is extension of the notepad file was created. Figure 8 illustrates the note pad.

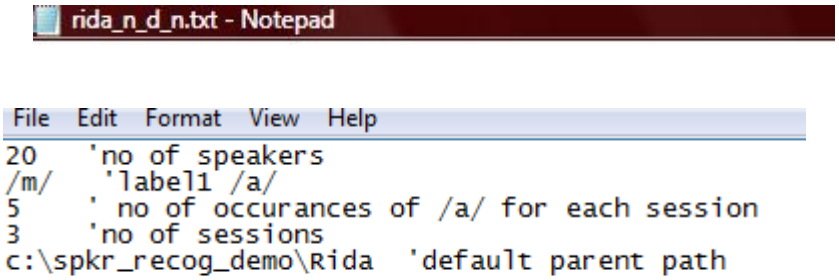


Figure 8: Illustration of the note pad.

MFCC computation: 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 9 and 10, respectively.

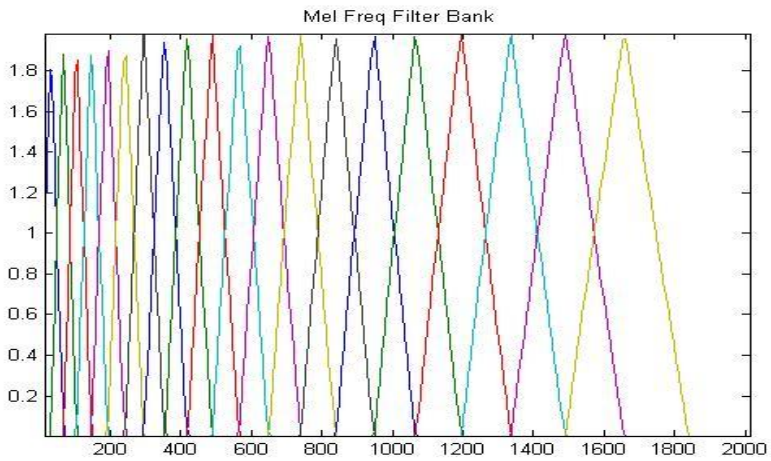


Figure 9: Mel frequency filter bank without normalization.

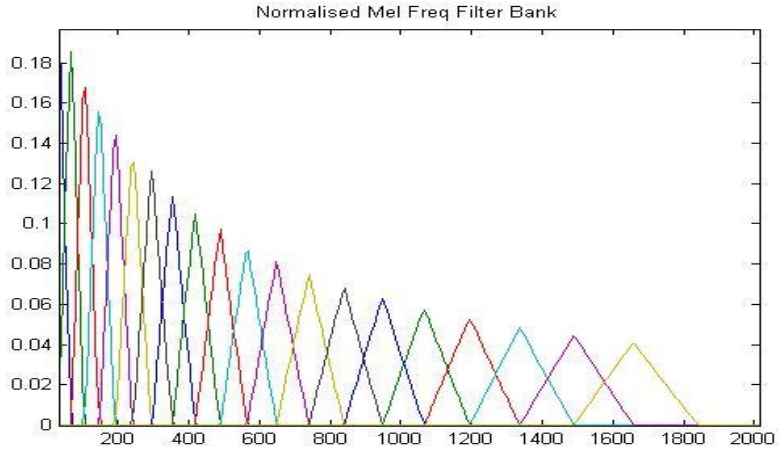


Figure 10: Mel frequency filter bank with normalization.

This notepad file was opened in SSL Workbench as in figure 11.

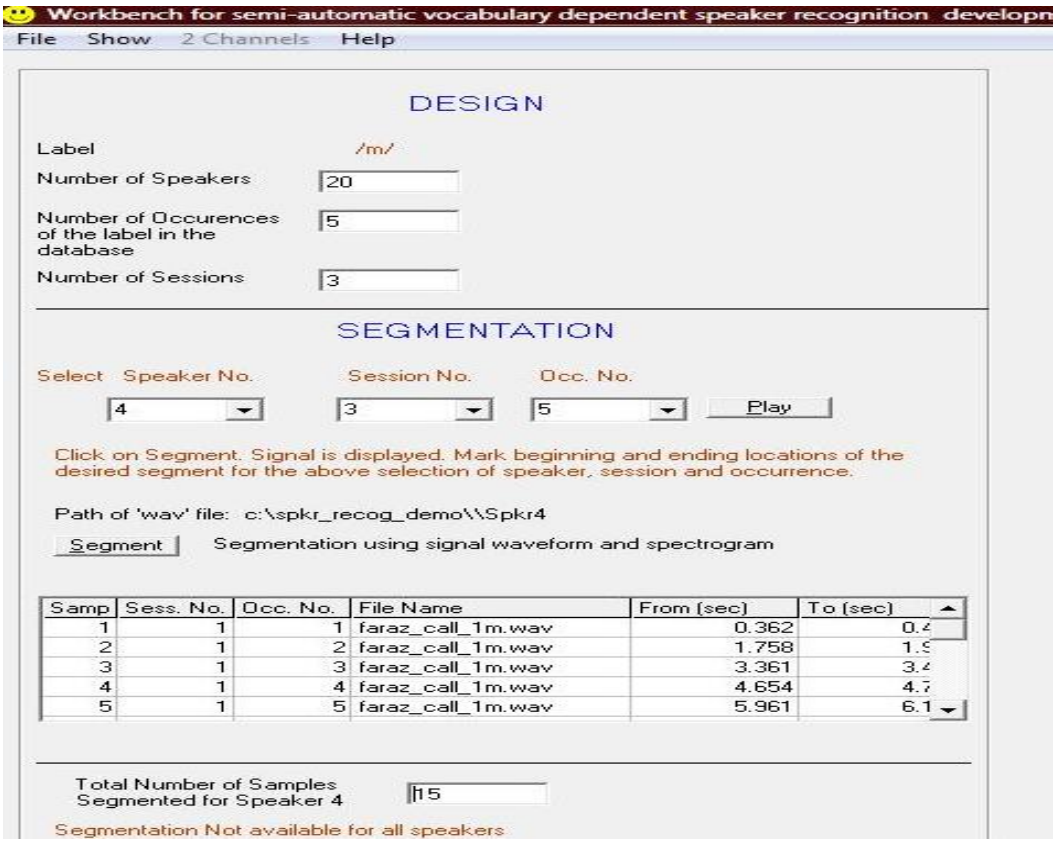


Figure 11: SSL workbench.

This file was created to obtain result no. 3. Although the number of speakers were 10 in the study, these were specified as 20 for this module because each speaker's live and network recording were treated as the recordings of two different speakers. For other results, this was specified as 10. The label of the file was kept as 'm' because the file was made for the nasal continuant /m/. This was changed according to nasal for which the file was being made, changed to /n/ or /ŋ/. The 'number of occurrence' was specified according to the nasal continuant being studied again. It was the number of times that nasal occurs in a single recording which was 5 /m/, 8 /n/ and 1 /ŋ/. The 'number of sessions' was specified according to the result under study again. This means the number of times the speaker's sentence was taken for analysis. It is important to note that although each speaker said the sentence 3 times in live and 3 times in network recording, it was not kept as 3 for each result. This was specified as 3 for the first three results, but was kept as two for the last result. 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. After making these specifications, the file was saved.

The notepad file was opened in SSL Workbench. When this was opened, the 'label', 'number of occurrence', and 'number of sessions' appeared on the window as they were already fed in to the software. The experimenter selected the recording to be analyzed and marked the segment according to the session number and occurrence number. This was done by clicking on the 'segment' button which opened the location specified in the parent file path of notepad file. Following this, the experimenters choose the file from the folder. Figure 12 shows the workbench window for analysis.

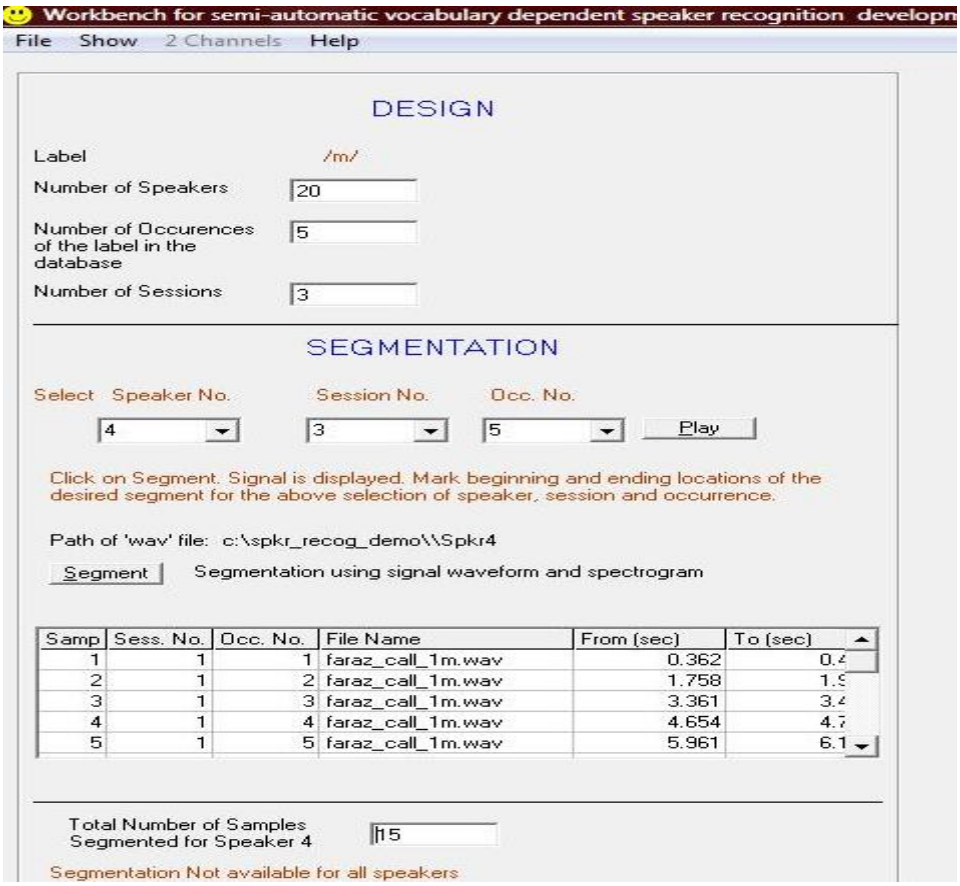


Figure 12: SSL Workbench window for analysis.

Followed by this samples for analysis 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 13 illustrated the speaker number being selected for segmentation.

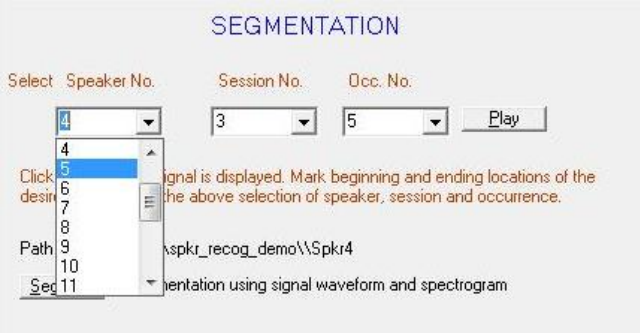


Figure 13: Illustration of speaker number being selected for segmentation.

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 14 illustrates selecting the session number and occurrence number.

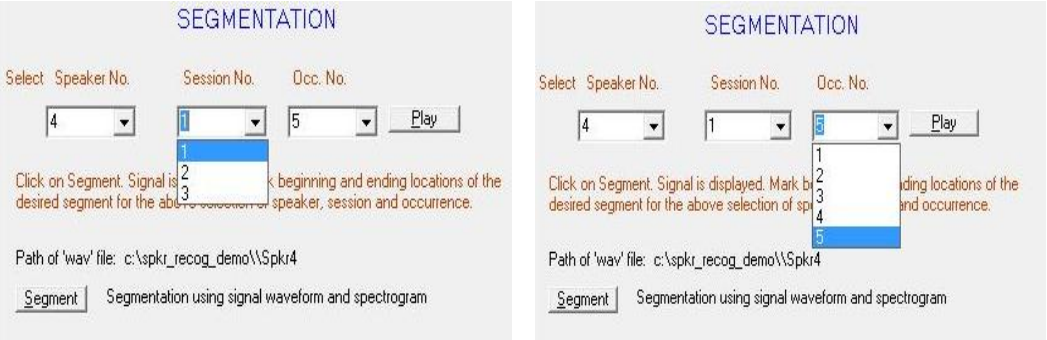


Figure 14: Illustration of selecting the session number and occurrence number.

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 opened for segmentation. Figure 15 illustrates segmentation window showing 5 occurrence of /m/ for a speaker.

Figure 15: 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’ was selected from the ‘Edit’ menu. After this, confirmation was done. Figure 16 shows the dialogue box asking for confirmation of the highlighted segment in the file.

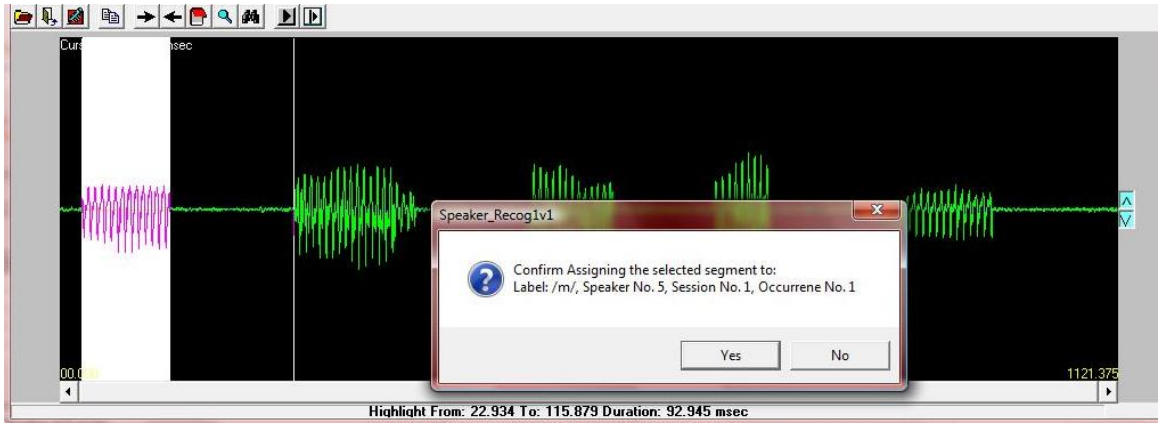


Figure 16: Showing dialogue box asking for confirmation of the highlighted segment in the file.

As soon as all files were segmented for all the speaker, ‘save segmentation’ option is 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 were selected and the sample for identification was tested. Figure 17 shows the analysis window of SSL Workbench.

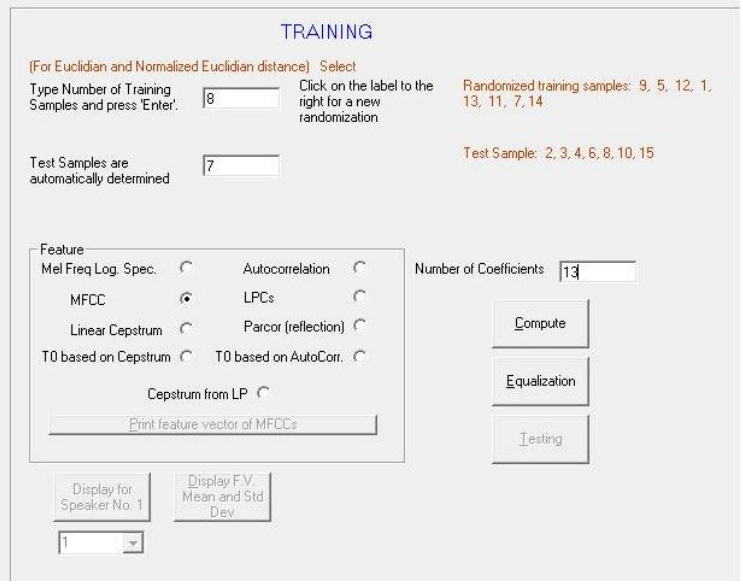


Figure 17: Analysis window of SSL Workbench.

Training samples number was specified and the rest were automatically selected as test samples. Once this was done, ‘compute’ is clicked on. This checked all the samples and compared them grossly and gave a qualitative analysis of each speaker. Next the ‘testing’ button was clicked on. This opened a window in which ‘compute score for identification’ was clicked on. This gave the diagonal matrix in the lower half of the window (figure 18) and a final percentage for correct speaker identification

This data was stored and the same procedure was repeated depending on the number of times according to the result. Live recordings were repeated 5 times; but network recordings were not repeated as they were taken as reference and compared with one live 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.

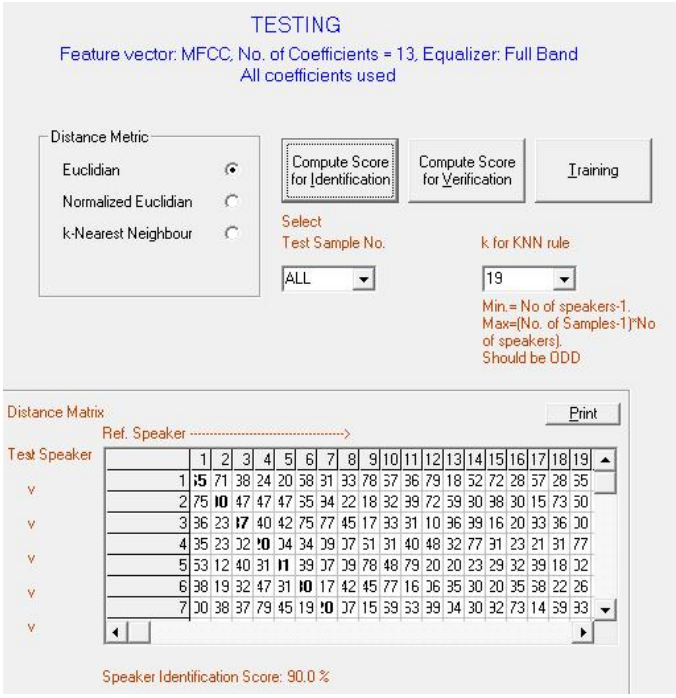


Figure 18: Analysis window of SSL Workbench showing diagonal matrix and the final speaker identification score.

Further telephone equalization was switched on and off which were considered as two conditions. 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, enhance particular aspects of an instrument's tone.

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 of the samples were averaged by the software separately for the test sample and the reference sample of the same speaker. These were then compared against all the speakers. The one with the minimum displacement from reference was identified as the test speaker. If the test and the reference speakers were the same then it was considered as correct identification; if not it was considered as incorrect identification. Percent correct identification was calculated by the formula Number of correct identification/ Total number of speakers * 100.

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

Chapter IV

Results

Results of the study will be discussed under following headings:

1) Comparison of MFCC of the speakers - live recording vs. live recording - for the three nasal continuants /m/, /n/ and /ŋ/

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

3) Comparison of MFCC of the speakers - live recording vs. network recording -but each treated as a different speaker for the nasal continuants /m/, /n/ and /ŋ/

4) Comparison of MFCC of the speakers - live recording vs. network recording - for the three nasal continuants /m/, /n/ and /ŋ/ considered as 2 different sessions.

1) Comparison of MFCC of the speakers - live recording vs. live recording - for the three nasal continuants /m/, /n/ and /ŋ/

Results indicated correct percent identification score for /m/, /n/ and /ŋ/ was seen to be 100%, 90% and 100%, respectively. The reference average is taken along the row and the test sample is taken along the column. The Euclidian distance of the samples were averaged by the software separately for the test sample and the reference sample of the same speaker. These

were then compared against all the speakers. The one with the minimum displacement from reference was identified as the test speaker. The green colour in the table indicates the correct identification of speaker sample as belonging to the same speaker as the reference sample. The red colour in the table indicates the error identification of test sample as belonging to a different reference speaker. The tables 1 to 3 depict the Euclidian distance as given by workbench. Sp refers to speaker in all the tables following.

Sp	1	2	3	4	5	6	7	8	9	10
1	1.345	4.499	4.037	8.163	3.399	5.072	4.886	4.74	5.329	4.239
2	6.584	2.806	7.439	7.092	7.401	7.29	6.148	8.048	5.303	4.824
3	3.668	4.964	0.967	8.875	4.601	3.552	5.28	4.89	5.205	3.64
4	8.753	7.19	9.658	1.274	9.753	8.558	7.993	8.564	5.801	7.497
5	3.03	5.804	4.7	9.636	1.925	5.965	4.345	4.057	6.287	4.508
6	4.967	5.247	3.645	7.775	5.727	2.359	5.323	4.991	3.711	4.028
7	4.821	5.6	5.233	8.31	4.174	5.436	1.434	4.944	4.316	3.215
8	4.046	5.708	5.15	7.836	4.785	5.787	5.205	1.277	5.514	4.424
9	6.033	4.942	6.548	5.271	6.095	4.573	4.532	5.924	2.278	4.218
10	3.786	3.389	4.117	6.572	3.515	4.267	2.964	4.322	3.517	1.561

Table 1: Diagonal matrix - live vs. live recording speaker identification of /m/.

Sp	1	2	3	4	5	6	7	8	9	10
1	2.24	6.496	4.468	4.974	4.767	3.8	4.915	6.258	6.366	5.038
2	4.376	4.007	3.291	4.138	2.853	3.999	3.423	4.32	4.24	2.729
3	3.882	5.299	0.727	3.682	3.4	2.864	2.5	6.078	4.068	4.568
4	5.466	6.285	5.083	2.582	4.451	5.121	4.314	4.263	2.694	3.959
5	4.156	4.448	3.577	4.344	1.083	3.547	3.602	4.049	4.137	2.563
6	3.447	5.226	3.294	4.231	3.503	1.439	2.948	6.059	5.554	3.637
7	6.309	5.628	5.241	5.645	6.368	5.044	4.234	8.132	7.137	6.99
8	5.341	6.013	5.81	5.014	3.995	5.881	5.442	1.069	5.166	3.99
9	6.513	5.506	4.308	3.31	4.488	5.474	3.816	5.83	1.145	5.156
10	4.331	4.602	4.384	3.686	2.878	3.616	3.561	4.193	4.314	0.835

Table 2: Diagonal matrix of live vs. live recording speaker identification of /n/.

Sp	1	2	3	4	5	6	7	8	9	10
1	1.077	4.419	3.943	7.288	5.81	4.324	9.844	3.507	7.58	6.341
2	4.915	0.934	3.479	5.641	3.274	3.075	7.881	3.595	4.478	5.031
3	3.568	3.832	1.942	6.321	4.819	4.046	9.619	4.43	5.343	6.574
4	7.62	5.263	5.379	1.328	7.504	4.694	6.505	6.767	3.076	6.928
5	5.776	3.954	5.408	8.186	0.762	4.513	9.628	4.288	6.432	7.685
6	4.533	2.968	3.567	4.163	5.327	1.158	7.08	3.517	4.7	5.009
7	11.344	8.529	9.738	7.952	10.004	8.515	2.198	9.367	8.1	8.741
8	2.901	4.417	4.967	7.679	4.475	3.603	9.488	1.9	7.333	7.395
9	7.64	4.529	4.649	3.986	6.179	4.433	6.334	6.35	1.445	5.925
10	6.229	4.138	4.516	6.5	6.902	4.881	8.182	5.609	6.091	1.088

Table 3: Diagonal matrix for live vs. live recording speaker identification of /ŋ/.

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

The results are discussed under one situation only. Results showed percent correct identification for /m/, /n/ and /ŋ/ to be 50%, 80% and 90%, respectively. The reference average is taken along the row and the test sample is taken along the column. The Euclidian distance of the samples were averaged by the software separately for the test sample and the reference sample of the same speaker. These were then compared against all the speakers. The one with the minimum displacement from reference was identified as the test speaker. The green and red colour in the table indicates the correct and error identification of speaker sample respectively. Table 4 to 6 depict Euclidian distance as given by Workbench.

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

Sp	1	2	3	4	5	6	7	8	9	10
1	8.297	10.067	12.299	12.243	11.079	10.581	8.021	9.084	5.136	5.912
2	4.626	1.883	5.382	5.88	5.275	4.858	6.853	5.305	7.353	5.076
3	9.102	9.641	5.188	6.638	10.133	9.875	10.112	9.626	11.905	10.102
4	6.142	3.824	9.249	6.384	4.674	4.468	5.334	4.674	5.467	4.866
5	8.658	6.05	10.004	7.424	3.935	5	5.89	3.514	7.548	6.209
6	7.862	5.52	6.819	2.585	4.367	4.518	5.293	4.562	8.714	6.881
7	5.751	7.059	7.979	7.574	7.81	7.32	4.32	5.698	3.946	3.277
8	5.063	3.133	6.458	5.19	3.575	3.419	5.25	2.805	6.181	4.32
9	7.835	8.797	12.463	12.902	11.49	10.56	10.157	10.519	7.132	7.56
10	4.693	5.906	6.478	6.551	7.254	7.12	5.237	5.584	4.638	2.429

Sp	1	2	3	4	5	6	7	8	9	10
1	1.641	10.075	7.858	9.908	11.185	12.408	11.385	10.824	11.483	9.969
2	11.639	2.536	8.161	7.321	6.235	7.265	7.369	6.357	3.9	5.28
3	8.935	3.932	3.352	5.844	4.513	6.667	8.704	5.62	6.997	4.905
4	13.53	8.859	9.659	6.222	6.543	6.773	7.106	5.683	8.94	6.716
5	9.551	6.077	5.972	6.322	2.563	5.085	7.025	3.1	7.411	4.167
6	10.584	6.062	7.595	5.933	4.102	2.38	5.477	2.822	6.201	3.678
7	11.237	7.193	9.72	6.055	7.023	7.352	1.626	5.207	6.177	4.792
8	7.386	7.153	7.019	8.084	6.344	7.431	8.108	6.292	7.73	6.276
9	9.529	2.822	7.181	5.936	6.362	7.289	6.402	5.833	2.793	4.921
10	8.679	4.173	5.729	4.176	3.48	4.766	4.507	2.529	5.481	1.371

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

Sp	1	2	3	4	5	6	7	8	9	10
1	2.265	5.477	7.686	12.452	10.28	10.202	11.097	7.628	9.489	5.228
2	9.058	3.679	10.702	10.378	5.938	5.917	9.776	5.134	12.009	7.132
3	6.282	8.836	3.4	12.417	11.633	11.305	10.591	9.359	10.936	8.769
4	13.017	10.413	10.019	3.517	7.859	8.458	5.939	8.786	11.479	11.479
5	10.65	7.135	9.058	8.014	3.986	4.364	3.995	4.452	9.002	7.967
6	14.121	9.242	12.671	6.699	4.204	4.172	5.93	6.461	12.092	11.105
7	11.601	10.279	7.895	7.847	8.521	8.631	3.976	8.184	9.247	10.087
8	12.819	8.791	11.736	7.368	4.541	5.092	4.891	6.896	9.443	9.803
9	9.048	7.895	11.678	13.573	10.139	9.617	9.558	8.591	4.061	5.458
10	5.135	3.955	9.773	13.137	8.746	8.553	10.232	5.939	7.896	2.26

Table 6: Diagonal matrix of network vs. network recording speaker identification of /ŋ/.

3) *Comparison of MFCC of the speakers - live recording vs. network recording -but each treated as a different speaker for the nasal continuants /m/, /n/ and /ŋ/*

The results are discussed under two situations. The Highest Percent Identification (HPI) and the Lowest Percent Identification (LPI) for each nasal continuant. The reference average is taken along the row and the test sample is taken along the column. The Euclidian distance of the samples were averaged by the software separately for the test sample and the reference sample of the same speaker. These were then compared against all the speakers. The one with the minimum displacement from reference was identified as the test speaker. The green colour in the tables indicates the correct identification of speaker sample as belonging to the same speaker as the reference sample. The red colour in the tables indicates the error identification of test sample as belonging to a different reference speaker.

It was found that the HPI for the continuants /m/, /n/ and /ŋ/ was 90%, 90% and 95%, respectively. The LPI for the continuants /m/, /n/ and /ŋ/ were found to be 60%, 65% and 60%, respectively. This indicated that /ŋ/ was the best nasal continuant for speaker identification through MFCC. Percent speaker identification was very poor when live recordings were compared with network recordings. Tables 7 to 13 show the results obtained under these conditions.

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.09	3.80	5.11	4.57	4.87	5.48	7.98	11.9	4.98	7.75	5.26	10.2	7.48	10.7	3.70	8.30	7.90	9.99	6.08	6.15
2	4.29	1.52	4.98	4.61	4.75	7.01	9.39	13.9	2.55	8.83	6.18	10.8	8.18	12.2	3.72	9.02	8.79	8.43	6.42	4.09
3	5.29	4.92	1.20	2.08	4.66	6.82	6.28	10.7	3.15	5.47	3.10	7.17	4.79	10.	4.11	5.44	5.47	7.45	4.32	4.54
4	5.27	5.56	2.26	1.73	5.18	7.52	7.00	10.5	4.41	5.55	3.95	6.98	5.66	10.5	3.24	5.13	6.48	8.79	5.23	4.95
5	3.73	4.01	5.22	4.17	1.70	4.48	7.12	11.4	4.53	5.86	4.82	9.05	6.63	8.22	5.16	6.83	5.85	7.14	6.46	4.33
6	6.61	5.70	6.34	5.91	5.20	2.33	8.81	11.	6.39	6.59	5.75	10.2	8.93	8.28	6.96	7.80	6.54	7.94	6.78	6.00
7	8.38	9.27	5.44	5.75	6.70	8.49	3.46	7.41	7.84	3.28	4.43	4.20	3.48	7.22	8.27	3.51	3.40	8.26	6.61	7.87
8	11.9	13.5	10.09	10.26	11.23	10.6	7.67	3.25	12.8	6.78	8.33	8.03	8.42	8.21	12.2	7.54	7.11	13.6	9.28	12.8
9	6.21	3.39	4.77	4.39	4.92	7.81	9.52	14.4	1.74	8.34	6.28	10.0	7.79	12.1	4.62	8.19	8.16	6.34	7.01	2.87
10	8.11	8.94	6.72	6.29	5.91	5.89	5.40	7.48	8.60	2.52	5.31	6.43	6.67	4.30	8.55	4.44	3.82	8.39	7.44	7.69
11	5.16	5.54	2.90	3.53	5.08	6.21	5.65	8.95	4.92	5.11	1.37	6.69	4.29	8.97	4.58	5.15	4.76	8.76	3.60	5.95
12	11.9	12.9	8.84	9.19	10.30	11.8	6.01	6.20	11.8	5.78	8.08	3.19	7.54	7.77	11.2	5.19	7.25	12.0	10.09	11.2
13	8.12	8.76	5.39	6.01	7.00	9.58	3.75	8.08	7.23	5.54	4.60	5.24	2.16	9.22	7.97	5.28	5.06	8.91	5.69	7.91
14	11.1	11.9	10.78	10.58	9.14	7.80	8.35	9.10	12.1	7.13	8.97	9.80	10.1	2.92	12.2	8.71	6.92	11.1	10.82	11.3
15	3.08	3.16	4.58	3.93	4.60	6.91	8.40	12.7	3.83	7.86	5.43	9.70	7.40	11.3	1.96	7.86	8.17	9.12	6.61	4.67
16	10.4	11.4	8.03	8.09	8.71	9.31	5.97	5.97	10.1	4.39	6.26	5.79	5.24	6.76	10.5	5.00	3.60	9.95	7.90	10.1
17	8.40	8.47	5.27	5.48	6.14	8.03	4.99	8.11	6.80	3.72	4.06	5.42	3.12	7.62	7.98	4.05	2.13	7.11	5.52	7.07
18	9.13	7.63	7.18	6.76	5.62	8.34	8.33	13.2	6.08	6.82	7.03	8.92	6.74	9.48	9.04	7.25	5.94	2.00	8.43	5.37
19	6.27	6.72	4.25	5.20	6.55	6.59	5.92	8.11	6.09	6.16	2.72	7.95	4.87	9.47	6.34	6.65	5.11	10.0	1.40	7.71
20	5.43	3.21	4.97	3.87	3.44	6.55	9.16	13.8	2.74	7.33	6.02	9.67	7.81	10.8	4.39	7.46	7.55	5.88	7.469	1.75

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

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.77	5.58	4.94	5.44	4.96	6.05	7.83	12.	4.85	7.95	5.27	7.13	7.46	11.4	5.01	7.19	8.24	10.6	6.54	6.74
2	3.87	4.05	5.30	4.60	5.33	8.13	9.55	14.5	3.01	8.91	5.50	6.00	8.47	12.4	5.24	7.56	9.21	9.8	6.94	5.13
3	5.20	6.56	1.91	2.44	3.86	7.67	6.65	11.2	2.92	5.02	2.67	3.54	5.51	10.2	3.99	4.94	5.91	7.77	4.92	4.98
4	5.44	7.84	2.15	2.64	4.81	8.24	7.27	11.0	4.04	4.80	3.66	4.02	6.51	10.6	2.24	4.17	6.95	8.43	6.01	5.33
5	3.26	4.49	5.24	4.53	2.44	5.07	7.32	12.2	4.26	6.81	4.41	5.76	5.99	8.95	6.03	5.98	5.95	7.65	7.15	4.57
6	6.32	6.02	6.27	6.36	4.44	3.63	9.24	12.9	6.36	8.04	5.18	7.49	8.35	9.38	7.30	6.86	6.78	7.97	7.42	5.63
7	8.36	10.0	5.39	6.06	5.40	8.57	4.09	7.85	7.40	2.75	4.86	5.64	3.49	7.46	7.58	5.17	3.51	7.15	6.93	7.63
8	12.2	14.7	9.5	11.08	9.77	10.0	7.33	4.67	12.3	6.90	9.16	11.2	8.28	8.60	11.1	9.09	7.94	11.1	9.63	12.4
9	5.76	4.36	5.47	3.76	5.43	9.01	10.0	15.0	2.31	8.4	5.23	4.40	8.16	12.3	5.57	7.07	8.30	8.19	7.53	3.85
10	8.01	9.52	6.44	6.81	4.71	5.77	5.99	8.70	8.21	4.50	5.43	7.18	5.72	5.34	8.05	4.94	3.96	6.61	8.08	7.02
11	5.31	7.37	2.58	4.38	4.11	6.54	5.72	9.41	4.68	4.70	2.04	4.81	5.01	9.53	4.40	4.84	5.56	8.62	4.28	6.20
12	12.0	14.3	8.53	9.45	9.35	11.7	6.82	6.27	11.5	4.70	8.64	8.85	7.92	7.20	9.5	6.57	7.23	9.82	10.56	10.7
13	8.11	9.92	5.41	6.19	5.96	9.81	3.8	8.22	6.89	4.38	5.14	5.54	3.39	8.79	7.46	6.29	5.74	8.38	5.84	8.05
14	10.9	11.7	10.55	11.00	8.32	7.00	8.87	10.1	11.9	8.77	9.29	11.1	8.88	5.96	12.0	9.27	6.65	9.84	11.31	10.6
15	3.07	5.47	4.63	4.33	5.25	7.7	8.54	13.2	3.88	7.76	4.97	5.78	7.75	11.7	4.04	6.43	8.56	9.96	7.22	5.46
16	10.5	12.1	7.77	8.57	7.39	9.02	6.09	6.89	9.64	4.64	6.87	8.15	5.03	7.46	9.69	7.28	3.99	7.96	8.32	9.82
17	8.28	9.14	5.39	5.65	4.85	8.27	5.42	8.83	6.39	3.76	4.14	5.03	3.26	7.89	7.47	5.61	3.42	6.05	6.02	6.90
18	8.57	6.68	7.75	6.12	5.40	9.11	8.96	14.0	5.91	7.70	6.20	5.36	6.05	9.71	9.35	7.59	5.57	4.31	8.81	5.19
19	6.45	8.40	3.90	6.09	5.06	6.80	5.62	8.62	5.88	5.83	3.63	6.63	5.53	9.94	5.98	6.77	6.47	9.67	2.02	7.97
20	4.93	3.96	5.42	3.41	4.42	7.70	9.66	14.6	2.79	7.83	4.94	4.47	7.79	11.2	5.43	6.17	7.48	7.31	8.14	2.59

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

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2.65	5.71	4.37	5.23	4.20	9.58	7.81	9.93	4.78	8.66	4.85	6.79	6.18	9.52	3.71	9.28	6.56	8.28	4.65	7.87
2	7.74	5	9.47	9.46	9.47	9.65	8.93	10.2	10.1	9.82	10.9	8.71	10.5	8.8	6.98	10.	8.14	7.72	9.49	6.77
3	3.86	7.22	2.87	4.39	5.42	9.74	6.77	9.45	4.16	7.93	4.81	6.1	4.95	8.98	5.16	9.19	5.92	8.36	2.99	8.29
4	5.34	7.22	5.01	2.19	7.04	9.33	6.08	8.07	6.61	7.31	5.39	4.48	7.32	7.77	4.91	7.23	5.20	8.81	4.76	8.41
5	3.53	7.12	4.39	5.81	2.01	9.89	8.07	10.0	3.77	8.47	4.79	7.20	5.20	10.2	5.28	9.32	6.89	8.17	4.02	8.30
6	8.97	8.18	10.8	8.47	9.81	3.30	8.17	4.42	11.4	10.7	10.1	8.05	11.8	9.29	8.19	7.34	7.68	5.22	9.26	7.22
7	7.99	7.37	7.87	5.79	10.4	9.18	3.20	8.07	9.15	6.69	9.63	4.99	9.03	5.3	6.91	7.72	4.14	7.69	7.93	7.26
8	10.3	9.20	10.7	9.55	12.1	7.82	8.09	6.93	11.8	10.3	12.1	9.09	12.4	8.71	9.11	10.1	8.90	7.40	10.4	7.47
9	5.23	7.97	3.85	7.14	5.69	11.6	8.63	12.5	2.35	10.6	4.80	8.93	3.80	11.8	6.25	12.5	7.81	10.2	4.67	10.2
10	9.69	10.0	10.2	9.5	10.9	12.1	8.98	9.79	11.4	4.85	12.7	7.34	11.7	7.35	9.73	7.31	8.96	8.97	10.3	7.61
11	4.86	7.30	4.8	5.40	4.85	9.65	8.32	10.3	4.70	9.90	2.63	7.04	5.34	11.0	5.45	9.93	6.65	9.47	4.04	9.89
12	7.38	7.84	8.01	4.80	8.91	8.37	6.36	5.18	9.66	5.86	8.91	3.83	10.0	5.73	6.83	3.79	5.80	7.37	7.05	7.01
13	5.21	7.61	4.24	7.03	5.60	10.9	7.96	11.9	2.78	9.76	5.30	7.95	1.25	11.0	6.72	11.5	6.64	9.34	4.07	9.82
14	9.72	7.81	10.7	9.81	12.1	10.0	7.6	8.34	12.1	8.17	13.4	8.28	12.3	5.09	9.04	9.15	8.29	6.81	10.5	5.18
15	4.83	5.01	5.35	5.42	6.95	9.51	7.38	9.48	6.46	9.02	6.36	6.63	7.59	8.84	3.46	9.57	6.41	8.69	5.96	7.70
16	8.05	7.91	9.06	6.28	8.94	8.48	7.38	6.89	10.1	7.00	9.12	4.7	10.3	7.71	7.35	4.14	6.00	8.27	8.24	8.41
17	7.07	6.94	7.03	5.75	8.55	7.39	3.81	7.77	7.35	8.29	7.53	5.79	7.15	7.56	6.45	8.44	4.00	6.34	6.40	7.28
18	12.6	12.0	13.4	12.8	14.0	9.71	11.1	10.2	13.6	14.2	14.2	13.0	13.9	12.2	12.3	13.4	11.9	9.27	12.9	10.4
19	4.82	8.45	3.73	3.88	5.89	10.7	6.93	9.73	4.76	7.33	5.18	5.96	5.35	8.79	6.19	8.63	6.23	9.17	3.48	9.01
20	7.59	6.62	9.07	7.63	9.57	6.73	5.93	5.05	10.0	7.57	10.7	6.75	10.6	5.67	6.86	7.06	6.59	3.58	8.41	3.15

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

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2.88	5.52	5.27	4.92	3.95	8.1	8.86	8.89	6.47	8.57	4.31	5.01	6.01	6.72	2.77	6.95	5.82	6.35	5.01	5.82
2	8.04	5.81	9.97	8.85	8.80	8.39	9.94	9.15	10.9	11.1	9.16	7.32	9.37	7.45	7.38	8.25	9.11	8.00	9.63	7.01
3	3.07	6.21	3.19	3.74	4.00	8.96	8.01	9.29	4.85	7.40	4.14	5.13	3.71	6.71	2.58	7.31	3.61	5.50	2.30	6.33
4	6.42	7.82	6.33	3.55	7.08	7.46	7.19	6.81	9.21	8.27	5.58	3.93	7.05	6.45	5.41	5.77	5.88	7.61	5.50	7.40
5	2.85	6.60	4.39	5.61	2.52	8.02	8.76	9.20	5.09	8.46	3.84	6.57	5.13	7.54	3.90	7.86	5.31	6.73	4.28	5.84
6	11.2	9.45	11.71	11.17	11.49	3.63	9.00	6.33	13.9	13.7	10.7	10.3	11.9	9.31	11.2	9.57	11.4	11.0	11.69	8.44
7	9.09	6.33	8.49	6.01	10.88	7.98	2.64	5.75	10.9	8.12	9.66	5.81	7.91	5.32	8.25	5.95	6.09	5.84	8.90	6.48
8	13.1	9.77	12.71	12.45	13.83	7.15	9.46	6.3	15.4	14.7	13.1	11.	13.8	9.55	12.8	10.8	12.7	11.0	13.30	9.07
9	4.54	7.60	3.56	6.12	5.13	10.1	9.61	11.0	3.14	10.2	4.18	8.12	4.01	9.49	4.40	10.5	4.81	6.64	3.20	8.05
10	10.6	8.11	10.57	10.18	11.75	10.7	8.43	8.42	13.4	6.98	12.6	8.51	12.1	6.03	11.0	6.19	10.3	9.14	11.73	6.60
11	5.24	8.38	5.52	5.73	5.15	8.41	9.80	10.3	6.16	10.6	2.87	7.1	4.80	9.59	4.17	9.70	5.94	8.28	3.63	8.75
12	9.13	8.05	8.73	6.72	9.59	6.26	6.10	4.69	12.1	7.85	8.86	5.08	9.30	5.33	8.41	4.23	7.82	8.63	8.44	6.92
13	4.81	7.35	4.72	6.76	5.02	9.80	9.23	11.2	4.01	9.56	5.12	7.99	2.54	8.97	4.82	9.99	4.45	6.57	3.65	7.69
14	12.8	8.20	12.36	11.31	14.37	11.0	8.13	7.39	15.2	11.0	14.4	9.89	13.3	6.29	12.4	8.45	11.5	8.71	13.21	7.97
15	5.78	5.87	6.78	5.27	7.26	7.51	8.13	7.44	8.70	9.98	5.90	5.21	7.63	6.97	4.91	7.32	6.69	6.77	6.70	6.53
16	9.58	9.49	10.53	7.97	9.40	6.33	8.57	7.20	13.0	9.05	8.86	5.97	10.0	7.66	8.89	5.20	9.26	10.7	9.73	8.57
17	8.28	5.67	7.96	6.72	9.78	5.60	2.74	5.14	9.92	9.40	8.70	6.93	7.14	5.85	7.90	6.88	6.1	5.63	8.41	5.26
18	15.1	12.9	15.07	14.88	15.61	8.56	11.5	8.77	17.2	17.3	15.1	14.4	15.6	12.1	15.4	13.1	15.1	13.6	15.24	11.5
19	5.37	7.70	4.51	4.01	5.91	8.39	7.03	8.02	7.36	7.30	5.39	5.32	5.67	6.43	5.30	6.43	4.78	6.76	4.08	6.80
20	9.81	6.55	9.88	9.32	10.88	6.48	6.58	4.07	12.7	10.6	10.8	8.16	11.0	5.11	9.97	6.70	9.67	7.72	10.51	5.10

Table 10: Diagonal matrix showing LPI for /m/.

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	4.04	12.4	3.01	6.44	4.25	7.14	4.48	8.86	3.39	6.76	4.15	8.74	4.46	8.97	5.35	10.2	5.66	11.7	3.26	8.74
2	9.49	3.18	12.23	9.25	11.49	6.77	14.2	9.43	13.0	11.9	11.4	11.4	14.4	11.1	10.1	9.66	12.9	8.37	11.62	8.20
3	3.27	11.1	2.48	5.7	3.31	6.23	4.07	6.76	3.41	4.71	2.51	6.39	4.97	7.36	4.63	8.29	4.99	10.6	2.49	6.91
4	5.54	9.86	6.29	2.39	6.82	7.09	7.73	9.33	6.52	6.83	6.21	7.39	6.83	7.42	4.36	7.85	6.00	6.53	7.20	6.83
5	3.35	11.1	3.60	5.34	2.11	6.53	5.04	7.39	4.40	5.23	2.90	6.89	4.81	7.58	5.04	7.99	4.12	9.71	3.98	6.86
6	6.30	6.44	7.89	6.69	8.07	4.52	10.0	7.99	8.82	8.88	7.75	9.52	10.3	9.21	6.43	9.01	9.46	8.86	7.42	7.28
7	4.20	12.5	2.55	5.57	4.06	7.97	2.52	7.70	3.71	4.30	2.96	6.03	3.58	6.42	4.01	8.12	3.66	10.5	3.31	7.01
8	8.38	8.15	9.8	9.69	9.34	6.90	10.8	4.58	10.9	8.15	8.56	7.97	12.2	7.99	8.46	7.67	10.7	10.7	8.96	6.63
9	4.86	12.5	3.4	5.57	4.94	7.68	4.79	9.38	2.68	6.41	4.38	8.17	4.38	9.07	5.20	9.99	5.37	10.9	4.29	8.76
10	8.22	13.4	7.42	8.96	7.95	10.1	7.38	6.83	8.69	3.95	6.43	4.31	8.94	5.95	6.75	6.53	7.28	11.5	7.92	7.03
11	4.08	11.8	3.42	5.25	3.59	7.51	3.86	6.86	4.23	3.4	2.21	5.07	4.27	5.83	4.25	7.31	3.55	9.99	4.15	6.47
12	8.70	12.9	8.67	9.61	9.08	10.5	8.19	5.72	10.0	5.19	7.47	3.96	10.2	4.63	7.66	6.33	8.73	11.9	8.62	6.41
13	6.55	15.0	4.65	7.17	5.50	10.2	4.22	11.1	4.26	7.30	5.45	9.20	2.51	9.60	6.89	11.1	4.62	12.3	5.83	10.3
14	7.72	12.2	8.00	7.38	8.33	10.0	7.66	7.43	9.25	5.32	7.03	4.67	8.65	3.60	5.97	5.54	6.93	9.24	8.37	5.69
15	4.00	9.79	4.69	4.56	5.72	5.78	6.08	6.87	5.59	5.47	4.60	6.64	6.56	6.38	2.86	7.39	5.94	8.81	4.71	5.84
16	5.74	9.25	6.17	5.95	5.73	6.45	7.61	5.67	7.05	4.2	4.97	4.60	8.25	6.63	5.29	5.42	6.24	8.09	6.41	5.00
17	4.57	11.9	4.01	4.61	3.71	7.85	4.80	8.48	4.54	4.92	3.72	6.57	3.77	7.06	4.26	7.57	1.95	8.61	4.98	6.89
18	11.0	7.56	13.10	8.69	12.49	9.98	14.8	12.3	13.6	12.3	12.4	11.8	14.1	11.5	10.1	9.90	12.1	5.56	13.32	9.42
19	3.70	11.8	2.24	6.19	3.49	6.79	3.29	6.92	3.77	5.02	2.94	6.79	4.58	7.18	4.63	8.48	4.76	11.0	1.99	7.06
20	5.24	8.69	6.79	5.53	6.43	6.41	7.91	5.30	8.07	5.16	5.69	5.00	8.46	4.79	4.47	4.61	6.47	7.33	6.73	3.36

Table 11: Diagonal matrix showing LPI for /ŋ/.

Sp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	5.33	12.0	3.53	6.96	4.47	8.55	4.04	8.49	3.59	7.75	3.64	8.00	3.64	8.86	5.34	10.7	5.05	10.2	4.39	8.00
2	9.40	3.72	12.00	9.30	11.63	6.48	14.3	9.92	13.6	12.3	12.5	11.7	15.2	13.5	11.1	9.05	12.0	8.89	11.24	10.1
3	4.09	10.4	2.12	6.06	3.13	7.22	3.52	6.34	2.80	5.31	2.05	5.56	4.51	6.98	4.38	8.50	4.18	9.27	3.06	6.35
4	6.78	9.99	6.59	2.87	6.79	8.46	7.63	8.34	7.21	7.08	7.62	6.44	7.23	9.08	5.29	8.37	5.48	5.29	7.08	5.59
5	4.38	10.5	3.83	5.49	2.14	7.57	4.43	6.86	4.23	5.96	3.86	6.29	4.74	7.68	5.06	8.08	2.71	8.28	3.97	5.89
6	5.40	7.07	6.14	5.88	6.98	5.05	8.33	6.66	8.11	8.17	6.89	7.96	9.10	9.49	5.53	8.46	7.29	7.87	5.66	6.85
7	5.38	11.9	3.22	6.00	4.57	9.33	2.66	7.07	3.68	5.13	4.13	5.32	3.58	6.05	3.90	8.86	4.06	9.05	4.12	5.88
8	7.23	8.41	8.11	9.49	8.59	7.01	9.46	4.36	9.79	7.24	8.03	7.43	11.5	7.55	7.62	7.03	9.37	10.8	7.45	7.76
9	6.55	12.1	4.28	6.29	5.51	8.90	5.19	9.11	3.86	7.56	4.94	7.73	4.33	9.64	5.29	10.8	5.35	9.56	5.44	7.92
10	8.56	13.3	7.25	8.96	8.14	11.5	7.14	6.29	8.28	3.94	7.57	4.60	8.58	4.85	6.15	7.67	7.58	10.7	7.84	6.5
11	5.04	11.1	3.85	5.13	3.74	8.72	3.45	5.88	4.07	3.89	3.85	3.88	4.13	5.33	3.76	7.36	2.89	8.05	4.58	4.79
12	8.98	12.8	8.45	9.83	9.30	11.8	8.12	5.78	9.51	4.58	8.70	4.52	10.2	3.13	7.49	7.10	8.98	11.2	8.62	6.93
13	7.43	13.8	5.51	6.97	5.57	10.9	4.27	9.82	4.45	7.74	5.71	7.86	2.53	8.82	6.4	11.1	4.63	9.98	6.40	8.05
14	7.81	11.0	8.10	6.17	8.43	10.6	8.08	6.25	9.28	5.11	8.95	4.25	9.18	4.87	5.78	5.69	7.03	6.82	8.22	4.01
15	4.31	8.81	4.59	4.05	5.78	6.7	6.09	6.07	6.14	6.28	5.69	5.93	6.75	7.46	2.98	7.70	5.32	6.80	4.81	5.01
16	5.85	9.54	4.95	5.43	5.01	7.53	6.36	4.96	6.03	2.78	5.72	3.56	7.58	6.64	4.45	5.28	4.71	7.23	5.10	4.09
17	5.82	11.8	4.30	5.06	3.83	9.31	4.01	7.6	4.32	5.46	5.16	5.92	3.83	7.40	4.27	8.31	1.83	7.58	4.74	5.20
18	10.8	7.86	12.53	7.71	11.93	9.66	14.2	11.2	13.7	11.8	13.5	11.2	14.4	13.7	10.6	9.08	11.0	5.90	12.16	9.01
19	4.46	11.3	1.90	6.46	3.63	8.23	2.34	6.76	2.96	5.73	2.61	6.07	3.66	6.61	4.16	9.12	4.19	9.63	2.71	6.47
20	4.30	7.42	6	5.25	5.96	6.49	7.36	3.89	7.75	5.06	6.68	4.58	8.81	5.90	4.86	4.34	5.97	6.66	5.53	3.86

Table 12: Diagonal matrix showing LPI for /n/.

	/ŋ/	/m/	/n/
Highest Percent Identification	95%	90%	90%
Lowest Percent Identification	60%	60%	65%

Table 13: Percent correct identification for all nasal continuants.

4) Comparison of MFCC of the speakers - live recording vs. network recording - for the three nasal continuants /m/, /n/ and /ŋ/ considered as 2 different sessions

The reference average (network recording) is taken along the row and the test sample (live recording) is taken along the column. The Euclidian distance of the samples were averaged by the software separately for the test sample and the reference sample of the same speaker. These were then compared against all the speakers. The one with the minimum displacement from reference was identified as the test speaker. The green values indicate the correct identification of speaker sample as belonging to the same speaker as the reference sample. The red values indicate the error identification of test sample as belonging to a different reference speaker. These values have been expressed under one condition, namely Highest Percent Identification (HPI). These will be depicted as the MFCC for the 10 speakers. The identification was recorded with telephonic equalization switch on and off. Therefore the HPI for /m/, /n/ and /ŋ/ with telephone equalization switched on was 80%, 70% and 100%, respectively. On the other hand the HPI with telephone equalization switched off for /m/, /n/ and / n˜/ was 90%, 90% and 100%, respectively. Tables 14 to 21 show the HPI diagonal matrix of Euclidian distance for 3 nasal continuants under two conditions (telephone equalization switched on and switched off).

Sp	1	2	3	4	5	6	7	8	9	10
1	4.092	7.182	7.584	11.269	4.651	6.202	10.535	7.965	10.527	9.467
2	6.003	4.038	7.001	8.048	3.396	3.804	7.966	3.388	6.599	8.499
3	6.821	7.009	3.911	8.514	8.602	6.786	8.202	8.404	8.608	8.086
4	13.752	10.422	10.289	6.849	14.478	11.89	9.63	11.353	10.948	9.572
5	8.045	4.693	5.813	4.461	8.063	5.758	5.568	5.164	5.618	6.979
6	8.412	6.133	5.435	6.045	8.157	6.337	4.914	6.936	4.193	7.924
7	11.033	9.729	8.429	9.039	11.56	10.118	6.99	10.394	7.542	11.382
8	6.522	4.049	4.149	5.317	7.128	4.68	5.726	4.886	5.539	7.105
9	10.972	8.619	7.912	8.355	9.124	8.349	7.452	9.268	5.796	10.934
10	5.619	6.083	6.322	9.68	4.021	4.363	8.442	6.057	6.61	10.027

Table 14: Diagonal matrix (HPI) of speaker identification for /ŋ/ - telephone equalized condition.

Sp	1	2	3	4	5	6	7	8	9	10
1	5.482	10.223	9.96	11.687	9.643	8.366	13.602	9.646	11.705	11.446
2	5.686	4.976	6.481	7.764	4.791	4	8.316	3.283	6.294	8.075
3	8.139	10.864	8.707	10.534	12.113	9.26	12.27	10.399	11.212	10.244
4	13.142	12.268	11.527	8.79	15.234	12.079	11.95	12.252	12.177	11.156
5	8.177	6.893	6.77	5.289	8.977	5.864	5.544	5.927	6.888	7.951
6	8.405	7.537	6.429	5.809	9.121	6.188	5.851	6.661	5.915	8.173
7	10.359	10.203	8.796	7.12	12.029	8.888	8.266	9.695	7.694	11.256
8	6.038	6.181	5.27	5.424	8.25	4.744	7.04	5.532	6.28	7.705
9	11.306	11.698	10.753	9.317	11.997	9.526	10.758	11.508	8.491	13.564
10	5.284	6.615	6.712	8.793	5.646	4.439	10.325	6.421	6.336	10.079

Table 15: Diagonal matrix (HPI) of speaker identification for /ŋ/ - telephone equalization off condition.

Sp	1	2	3	4	5	6	7	8	9	10
1	2.34	8.05	9.28	8.21	10.15	7.21	9.67	10.25	5.74	4.80
2	6.48	3.5	7.83	5.48	8.23	5.93	7.51	7.86	6.60	5.49
3	10.19	11.96	2.87	9.17	14.97	12.58	14.65	13.84	12.06	9.90
4	9.15	5.59	12.70	6.86	6.18	7.12	4.96	7.60	5.87	8.28
5	6.30	5.34	11.37	7.47	2.60	2.74	4.46	4.85	4.05	5.48
6	5.94	5.13	11.22	8.17	5.07	2.65	6.07	7.35	4.58	6.02
7	10.94	6.55	15.64	9.95	5.30	6.93	3.64	5.92	8.19	9.67
8	7.73	4.03	11.86	6.4	3.88	4.31	2.47	2.72	5.35	5.90
9	4.67	7.33	11.31	7.35	7.70	6.43	7.15	8.90	3.16	5.61
10	3.63	6.30	5.91	5.29	8.80	5.94	8.36	7.71	5.74	2.57

Table 16: Diagonal matrix (HPI) of speaker identification for /m/ - telephone equalized condition.

Sp	1	2	3	4	5	6	7	8	9	10
1	2.73	6.96	8.53	9.09	10.14	6.56	8.83	9.68	5.66	6.16
2	5.61	3.42	7.02	8.89	9.67	6.50	8.10	8.86	6.65	7.11
3	9.64	10.65	3.01	10.19	14.60	11.73	13.51	13.10	11.56	10.29
4	9.46	9.20	11.90	5.68	8.44	8.84	6.36	7.55	6.89	8.07
5	6.47	6.78	11.07	9.17	2.34	3.62	5.07	4.76	4.17	5.80
6	4.87	5.75	9.96	8.32	5.20	2.85	5.56	6.34	3.38	4.99
7	8.85	6.65	13.38	9.20	5.54	6.19	2.38	4.75	5.98	8.23
8	6.78	5.81	10.74	6.94	4.07	4.37	2.29	2.39	3.90	5.26
9	4.63	6.86	10.45	7.78	6.63	5.28	5.80	7.11	2.74	4.64
10	5.00	7.46	6.06	5.25	8.96	6.71	7.61	6.90	5.63	2.70

Table 17: Diagonal matrix (HPI) of speaker identification for /m/ - telephone equalization off condition.

Sp	1	2	3	4	5	6	7	8	9	10
1	2.70	9.30	6.03	10.38	11.45	10.65	13.26	7.48	11.58	9.39
2	5.13	3.09	8.13	5.58	6.03	4.83	7.349	6.38	6.45	4.18
3	5.44	9.42	2.48	10.47	11.69	11.17	13.79	9.99	13.53	10.81
4	8.42	4.60	9.39	2.70	3.12	2.50	4.40	8.32	7.26	4.19
5	8.91	4.54	10.88	3.95	3.25	2.45	4.38	7.73	4.86	3.53
6	6.88	5.55	9.18	3.38	3.60	2.69	5.04	4.81	5.5	1.93
7	11.18	7.02	14.04	6.21	5.24	4.31	3.65	9.06	5.82	4.98
8	6.22	7.75	8.34	5.18	6.01	5.66	7.42	3.28	7.39	4.39
9	11.09	6.79	12.87	6.32	5.32	5.29	6.14	8.44	2.60	5.07
10	7.87	5.42	10.10	3.37	3.58	2.63	4.24	6.18	4.78	1.10

Table 18: Diagonal matrix (HPI) of speaker identification for /n/ - telephone equalized condition.

Sp	1	2	3	4	5	6	7	8	9	10
1	3.13	10.38	7.23	9.56	12.16	11.69	12.6	7.64	10.70	9.51
2	6.18	2.90	8.18	5.39	5.96	4.90	6.66	6.83	7.28	3.80
3	4.85	8.9	2.39	9.08	11.39	10.83	12.37	9.99	12.65	9.99
4	7.74	5.19	8.69	2.19	4.36	3.86	4.08	7.65	6.94	3.86
5	10.42	5.09	11.5	5.85	2.96	2.35	4.66	8.26	6.99	4.37
6	8.90	5.46	10.04	4.73	2.70	2.17	4.55	6.06	6.88	3.04
7	10.80	7.65	13.53	5.88	5.48	5.31	3.76	7.82	4.95	4.72
8	7.06	8.08	9.48	5.04	6.13	6.21	6.59	2.94	6.53	4.47
9	10.18	7.06	12.29	5.64	5.11	5.10	5.00	7.29	3.06	4.27
10	8.11	5.07	9.90	3.37	3.65	2.96	3.61	6.20	5.48	0.95

Table 19: Diagonal matrix (HPI) of speaker identification for /n/ - telephone equalization off condition.

Condition	/m/	/n/	/ŋ/
Telephone equalization	80%	70%	50%
No telephone equalization	90%	90%	30%

Table 20: Percent correct speaker identification in both conditions.

To summarize, the percent correct speaker identification was 100, 90, 100 for /m/, /n/, and /ŋ/, respectively when live recoding was compared with live recording; 50, 80 and 90 when network recording was compared with network recording; HPI was 90, 90, 95 when live recording was compared with network recording; LPI was 60, 65, 60 when live recording was compared with network recording; 80, 70, 50 when live recording was compared with network recording under telephone equalized condition; 90, 90, 30 when live recording was compared with network recording under telephone not equalized condition. The results indicated that the nasal continuant /ŋ/ had the best percent correct speaker identification among the nasals except under telephone equalized/ not equalized conditions. Under these conditions /m/ had the best percent correct speaker identification. Table 21 shows the summary of percent correct speaker identification. Figure 19 shows a graphical representation of percent identification under three conditions.

Condition	/m/	/n/	/ŋ/
Live vs. Live recording	100	90	100
Net work vs. network recording	50	80	90
Live vs. network recording – HPI	90	90	95
Live vs. network recording – LPI	60	65	60
Live vs. network recording considered as two sessions – telephone equalization	80	70	50
Live vs. network recording considered as two sessions – No telephone equalization	90	90	30

Table 21: Summary of percent correct speaker identification.

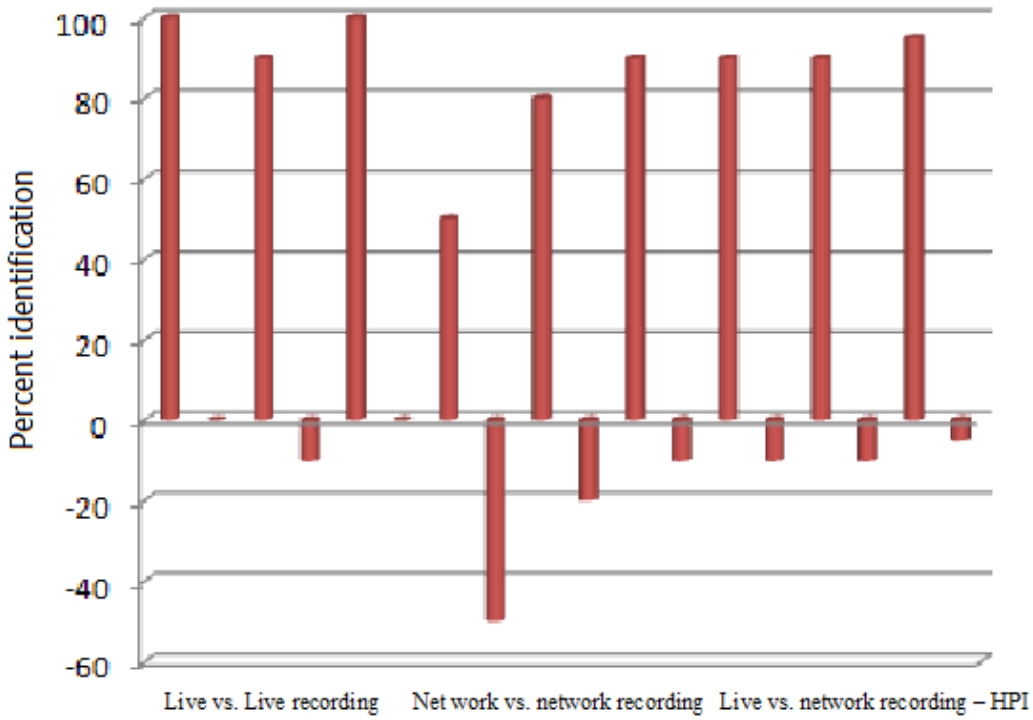


Figure 19: Percent identification under 3 conditions.

Chapter V

Discussion

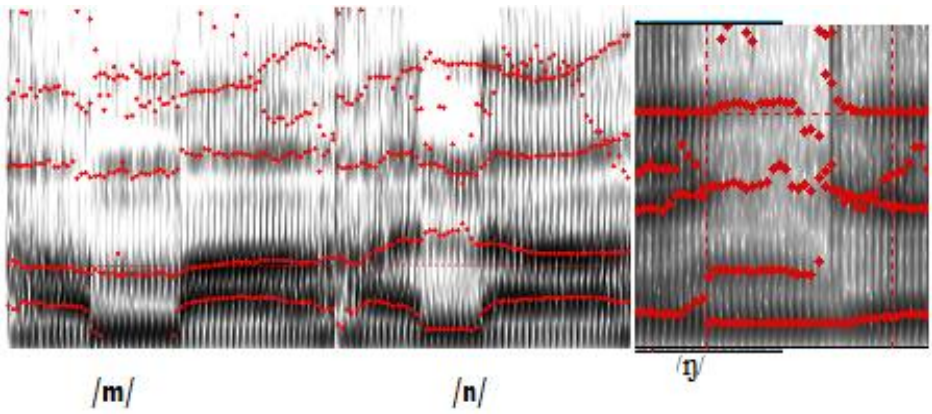
Results indicated that the ***percent correct speaker identification was 100, 90, 100 for /m/, /n/, and /ŋ/, respectively when live recoding was compared with live recording using MFCC***. The results are in agreement with those of the earlier studies. Hasan et al (2004) showed 57.14% speaker identification for code book size of 1, and 100% speaker identification for code book size of 16. Rajsekhar (2008) using the word "zero" reported 75% identification in MFCC. Tiwari, (2010) reported improvement in percent correct speaker identification with increase in number of filters in MFCC with 85% for 32 filters. Chandrika (2010) reported the overall performance of speaker verification system using MFCCs as about 80%. The overall performance of speaker recognition is about 90% to 95% for vowel /i/. Ramya (2011) used Mel frequency Cepstral coefficients (MFCC) for speaker identification and reported that percent correct identification was above chance level for electronic vocal disguise for females. Interestingly vowel /u:/ had 96.66%, /a:/ 93.33 %, and /i:/ 93.33%. Patel and Prasad (2013) reported 13% error rate for the word "hello" using MFCC. While these studies used vowels and words the current study has used nasal continuants. The percent correct identification in the present study, interestingly, is very high. This could be attributed to the characteristics of nasal continuants. Nasal continuants require two movements for its correct articulation- movement of tongue or lips to occlude the oral tract and lowering of the velum. This gives a unique quality to the spectrum produced (Pickett, 1980).

Percent correct speaker identification was 50, 80 and 90 when network recording was compared with network recording. The percent correct identification drastically decreased when network recording was compared with network recording. Further, HPI was 90, 90, 95 when live recording was compared with network recording; LPI was 60, 65, 60 when live recording was compared with network recording. It was assumed that the network frequency bandwidth (900/1800 for vodafone) would mask the characteristics of the nasals that would have helped for identification in the direct to direct identification. These characteristics would not have been found or have been eliminated altering the spectra of the nasal continuants in the direct vs. network comparison of the same speaker. However in previous studies also it has been found that nasal consonants, despite having a lower frequency and being more prone to masking by noise prove as a good cue for speaker identification. The HPI could again be attributed to the nature of nasal continuants. Jyotsna (2011) states that nasal coarticulation leads to better speaker identification (>90%) in Malayalam. Pickett (1980) says nasalization effect stays for 100ms preceding and following the nasal consonant leading to maintenance of nasal characteristics for a longer duration than any of the other speech sounds. A more stable spectrum would be obtained with lesser variation and more chances of correct speaker identification.

Percent correct speaker identification was 80, 70, 50 when live recording was compared with network recording under telephone equalized condition; 90, 90, 30 when live recording was compared with network recording under telephone not equalized condition. "In telecommunications, equalizers are used to render the frequency response—for instance of a telephone line—*flat* from end-to-end. When

a channel has been "equalized" the frequency domain attributes of the signal at the input are faithfully reproduced at the output. Telephones, DSL lines and television cables use equalizers to prepare data signals for transmission (http://en.wikipedia.org/wiki/Equalization). In fact one should expect higher percent identification scores when the equalization is on. However, interestingly, it was the reverse in the present study. Percent speaker identification was better when the telephone was not equalized.

The results indicated that the ***nasal continuant /ŋ/ had the best percent correct speaker identification among the nasals except under telephone equalized/ not equalized conditions***. The velar nasal continuant has a mid frequency spectra, the bilabial has a low frequency spectra and the dental has a high frequency spectra. Most often energy in the nasal continuants is damped. Following spectrogram shows wide band bar type of spectrogram of all the three nasals.



The reason that velar /ŋ/ had highest percent correct identification may be attributed to the acoustic properties of this nasal continuant. In the production of bilabial /m/, tongue anticipates or retains the position of adjacent vowels. The continuant is voiced except when partially devoiced by a preceding unvoiced consonant. The first resonance occurs at around 250 Hz, the second at around 1000Hz. The oral resonance may show some continuity with

the vowel's second formant. The oral resonance is weak in the murmur itself but as the closure is released it will increase substantially as it moves into the vowel F2 transition. In the production of dental /n/, the lip shape is dependent on adjacent vowels e.g. "coon", "keen". The continuant is voiced except when partially devoiced by a preceding unvoiced consonant. The first resonance occurs at around 250 Hz, the second at around 800Hz. The first antiresonance is higher than that of /m/ as a result of shortened oral tract. The oral resonance is about 1.4 k Hz. Lack of energy on spectrograms of /n/ above 250Hz up to at least 2 kHz is usual. In the production of velar /ŋ/, point of closure depends on following vowel. More fronted for "sing" than "sung". The first resonance occurs at around 250 Hz. The first antiresonance is highest at above 3 kHz. There is very little side branching. The frequency of first antiresonance, little side branching, and lowest number of occurrence (1) may be reason for high percent correct identification of /ŋ/.

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

Condition	/m/	/n/	/ŋ/
Live vs. Live recording	100	90	100
Net work vs. network recording	50	80	90
Live vs. network recording – HPI	90	90	95
Live vs. network recording – LPI	60	65	60
Live vs. network recording considered as two sessions – telephone equalization	80	70	50
Live vs. network recording considered as two sessions – No telephone equalization	90	90	30

The study was restricted to 10 participants and 14 occurrences of nasal continuants and Hindi 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

A voice is more than just a string of sounds. Identifying people on the basis of their voice is a common phenomenon. 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.

The review indicates that the effects of vocal disguises markedly interfere with spectrographic speaker identification as well as speaker identification by listening. This calls for a need for benchmark to be established in such a way that would not be independent of manipulation by the speaker. Thus, the aim of the study was to establish Benchmark for speaker identification for nasal continuants in Hindi using Mel frequency cepstral coefficients (MFCC). The objectives of the study are to provide benchmarks for Mel-frequency cepstral coefficients for Hindi nasal continuants in mobile and network conditions.

Ten participants between the age range of 20 to 40 years with at least 10 years of exposure to Hindi language as a mode of oral communication were included in the study. Material included 6 Hindi sentences with bilabial, dental and velar nasals embedded in words in all positions. Participants were instructed to speak the sentences

under two conditions- directly into the recording mobile (live) and through another mobile into the recording mobile phone (network) - 3 times at an interval of 1 minute. The network used for making the calls was Vodafone (GSM 900/ GSM 1800 MHz frequency) and the receiving network was also Vodafone on a Sony Ericsson Xperia pro mobile phone. The message at the receiving end were recorded and saved in the microchip of Sony Ericsson Xperia Pro. Analyses of the data were carried out using SSL Work Bench (Voice and Speech Systems, Bangalore, India) to extract Euclidian distances. A speaker was presumed to be identified correctly when the Euclidian distance between the training and test sample was the least. Percent correct identification was calculated by using the formula total number of correct identification/ number of samples * 100.

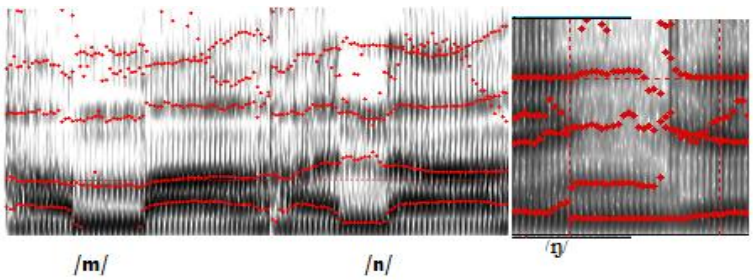
Results indicated that the ***percent correct speaker identification was 100, 90, 100 for /m/, /n/, and /ŋ/, respectively when live recoding was compared with live recording using MFCC***. The results are in agreement with those of the earlier studies. Hasan et al (2004) showed 57.14% speaker identification for code book size of 1, and 100% speaker identification for code book size of 16. Rajsekhar (2008) using the word "zero" reported 75% identification in MFCC. Tiwari, (2010) reported improvement in percent correct speaker identification with increase in number of filters in MFCC with 85% for 32 filters. Chandrika (2010) reported the overall performance of speaker verification system using MFCCs as about 80%. The overall performance of speaker recognition is about 90% to 95% for vowel /i/. Ramya (2011) used Mel frequency Cepstral coefficients (MFCC) for speaker identification and reported that percent correct identification was above chance level for electronic vocal disguise for females. Interestingly vowel /u:/ had 96.66%, /a:/ 93.33 %, and /i:/ 93.33%. Patel and Prasad (2013) reported 13% error rate for the word "hello" using MFCC. While these studies

used vowels and words the current study has used nasal continuants. The percent correct identification in the present study, interestingly, is very high. This could be attributed to the characteristics of nasal continuants. Nasal continuants require two movements for its correct articulation- movement of tongue or lips to occlude the oral tract and lowering of the velum. This gives a unique quality to the spectrum produced (Pickett, 1980).

Percent correct speaker identification was 50, 80 and 90 when network recording was compared with network recording. The percent correct identification drastically decreased when network recording was compared with network recording. Further, HPI was 90, 90, 95 when live recording was compared with network recording; LPI was 60, 65, 60 when live recording was compared with network recording. It was assumed that the network frequency bandwidth (900/1800 for vodafone) would mask the characteristics of the nasals that would have helped for identification in the direct to direct identification. These characteristics would not have been found or have been eliminated altering the spectra of the nasal continuants in the direct vs. network comparison of the same speaker. However in previous studies also it has been found that nasal consonants, despite having a lower frequency and being more prone to masking by noise prove as a good cue for speaker identification. The HPI could again be attributed to the nature of nasal continuants. Jyotsna (2011) states that nasal coarticulation leads to better speaker identification (>90%) in Malayalam. Pickett (1980) says nasalization effect stays for 100ms preceding and following the nasal consonant leading to maintenance of nasal characteristics for a longer duration than any of the other speech sounds. A more stable spectrum would be obtained with lesser variation and more chances of correct speaker identification.

Percent correct speaker identification was 80, 70, 50 when live recording was compared with network recording under telephone equalized condition; 90, 90, 30 when live recording was compared with network recording under telephone not equalized condition. "In telecommunications, equalizers are used to render the frequency response—for instance of a telephone line—*flat* from end-to-end. When a channel has been "equalized" the frequency domain attributes of the signal at the input are faithfully reproduced at the output. Telephones, DSL lines and television cables use equalizers to prepare data signals for transmission (http://en.wikipedia.org/wiki/Equalization). In fact one should expect higher percent identification scores when the equalization is on. However, interestingly, it was the reverse in the present study. Percent speaker identification was better when the telephone was not equalized.

The results indicated that the ***nasal continuant /ŋ/ had the best percent correct speaker identification among the nasals except under telephone equalized/ not equalized conditions***. The velar nasal continuant has a mid frequency spectra, the bilabial has a low frequency spectra and the dental has a high frequency spectra. Most often energy in the nasal continuants is damped. Following spectrogram shows wide band bar type of spectrogram of all the three nasals.



The reason that velar /ŋ/ had highest percent correct identification may be attributed to the acoustic properties of this nasal continuant. In the production of bilabial /m/, tongue

anticipates or retains the position of adjacent vowels. The continuant is voiced except when partially devoiced by a preceding unvoiced consonant. The first resonance occurs at around 250 Hz, the second at around 1000Hz. The oral resonance may show some continuity with the vowel's second formant. The oral resonance is weak in the murmur itself but as the closure is released it will increase substantially as it moves into the vowel F2 transition. In the production of dental /n/, the lip shape is dependent on adjacent vowels e.g. "coon", "keen". The continuant is voiced except when partially devoiced by a preceding unvoiced consonant. The first resonance occurs at around 250 Hz, the second at around 800Hz. The first antiresonance is higher than that of /m/ as a result of shortened oral tract. The oral resonance is about 1.4 k Hz. Lack of energy on spectrograms of /n/ above 250Hz up to at least 2 kHz is usual. In the production of velar /ŋ/, point of closure depends on following vowel. More fronted for "sing" than "sung". The first resonance occurs at around 250 Hz. The first antiresonance is highest at above 3 kHz. There is very little side branching. The frequency of first antiresonance, little side branching, and lowest number of occurrence (1) may be reason for high percent correct identification of /ŋ/.

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

Condition	/m/	/n/	/ŋ/
Live vs. Live recording	100	90	100
Net work vs. network recording	50	80	90
Live vs. network recording – HPI	90	90	95
Live vs. network recording – LPI	60	65	60
Live vs. network recording considered as two sessions – telephone equalization	80	70	50
Live vs. network recording considered as two sessions – No telephone equalization	90	90	30

The study was restricted to 10 participants and 14 occurrences of nasal continuants and Hindi 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, *Journal of Acoustic Society of America*, Vol. 55, 1304-1312

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

B. Fauve, N. Evans, N. Pearson, J. F., Bonastre, and J. S. D. Mason. (2007). Influence of task duration in text-independent speaker verification, Proceeding Inter-speech, 794-797.

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.

Campbell, J. P., Matrouf, D., Schwartz, R., Campbell, W. M., Wade, S. (2009). Forensic Speaker Recognition. Signal Processing Magazine, IEEE, Vol 21 (2). 95-103.

Chandrika.. (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.

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

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

Glass, J.R., Zue, V.W. (1985). Detection of nasalized vowels in American English, in Proc. of ICASSP.

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

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

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

Hindi (n.d.) in Wikipedia Online. Retrieved from *http://en.wikipedia.org/wiki/Hindi*

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

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

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.

Kent, R. D., Charles, R. (2002). *The Acoustic Analysis of Speech, 2nd Edition.* 83- 95.

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, K., Aggarwal, R. K. (2011). Hindi speech recognition system using HTK. *International journal of Computing and Business Research,* Vol. 2.

Lei, H. (2010). Structured Approaches to Data Selection for Speaker Recognition, Ph.D Thesis, UC Berkeley.

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

Lei, H., Mirghafori, N. (2011). Data selection with kurtosis and nasality features for speaker recognition, Proceeding Inter-speech.

Liu, M., Xie, Y., Dai, B., Yao, Z. (2006). Kurtosis Normalization in Feature Space for Robust Speaker Verification, in Proc. Of ICASSP.

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

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

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

Medha, 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.

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

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

Patel, K., & Prasad, R. K. (2013). Speaker recognition and verification using MFCC & VQ. *International Journal of Emerging Science and Engineering*, 1 (7), 33-37.

Pickett, J. M. (1980). *The sounds of Speech Communication: A Primer of Acoustic Phonetics and Speech Perception*. California: University Park Press.

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.

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

Pruthi, T and Espy-Wilson, C. Y. (2007). Acoustic parameters for the automatic detection of vowel nasalization, in Proc. of Interspeech,

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.*

Rajsekhar, A.(2008). *Real time speaker recognition using MFCC and VQ.* Thesis submitted in fulfillment of Master of Technology degree in Electronics and Communication Engineering to National Institute of Technology, Rourkela. Downloaded from http://ethesis.nitrkl.ac.in/4151/1/2.pdf on 29.4.2013.

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

Ramya. B.M. (2013). Bench mark for speaker identification under electronic vocal disguise using Mel Frequency Cepstral Coefficients. *Unpublished project of Post graduate Degree in Masters of Science submitted to University of Mysore, Mysore.*

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

Reich, A. R., Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *Journal of the Acoustical Society of America,* Vol. 26, 403-406.

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

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

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

Savithri, S. R. (1987). Degree of nasalization. An experimental verification. *Journal of Acoustical Society of India*, Vol. 15, 243 -254.

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

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.

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.

Su, L. S., Li, K. P., Fu, K. S. (1974). Identification of speakers by use of nasal coarticulation, *Journal of the Acoustical Society of America,* Vol. 56,1876-1883.

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.

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