Benchmark for Speaker Identification using Tamil Nasal Continuants in Live Recording and Mobile Network Recording

Nithya K

Register No: 14FST003

An Independent Project Submitted in Part Fulfillment of PG Diploma in

Forensic Speech Science and Technology (PGDFSS&T)

University of Mysore

Mysuru

ALL INDIA INSTITUTE OF SPEECH AND HEARING

MANASAGANGOTHRI, MYSURU-570006

JULY, 2015

CERTIFICATE

This is to certify that this independent project titled **"Benchmark for Speaker Identification using Tamil Nasal Continuants in Live Recording and Mobile Network Recording"** is the bonafide work submitted in part fulfilment for the Post Graduate Diploma in Forensic Speech Science and Technology by the student (Registration No. 14FST003). This has been carried out under the guidance of a faculty of this institute and has not been submitted earlier to any other University for the award of any other Diploma or Degree.

Mysuru
July, 2015

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

CERTIFICATE

This is to certify that this independent project titled **"Benchmark for Speaker Identification using Tamil Nasal Continuants in Live Recording and Mobile Network Recording"** has been prepared under my supervision and guidance. It is also certified that this has not been submitted earlier in any other University for the award of any Diploma or Degree.

Mysuru
July, 2015

Dr. Jayakumar T.
Guide
Lecturer in Speech Sciences
Department of Speech-Language Sciences
All India Institute of Speech and Hearing
Mysuru - 570006

DECLARATION

This is to certify that this independent project titled **"Benchmark for speaker identification using Tamil nasal continuants in live recording and mobile network recording"** is the result of my own study under the guidance of Dr. Jayakumar T., Lecturer, Department of Speech-Language Sciences, All India Institute of Speech and Hearing, Mysuru, and has not been submitted earlier in any other university for the award of any diploma or degree.

Mysuru
July, 2015

Register No. 14FST003

ACKNOWLEDGEMENTS

First and foremost I would like to thank the Dr. S. R. Savithri, Director, All India Institute of Speech and Hearing, Mysuru, for permitting me to conduct this study.

Words cannot express my gratitude for the guidance, help and support my guide, Dr. Jayakumar T., has provided me with. Thank you very much sir.

I would also like to thank Dr. Sreedevi, Dr. Hema, Dr. Santosh and Dr. Rajsudhakar for their valuable suggestions during the course of this program.

I would like to extend my heartfelt gratitude to all the participants, without whose co-operation this study would not have been possible.

TABLE OF CONTENTS

List of Tables

List of Figures

INTRODUCTION

Biometrics refers to the identification of a person's identity based on his/her traits. Such traits may vary from simple factors such as height, weight, build, facial complexion, colour of the eyes, etc to the more sophisticated factors such as finger prints, DNA etc. With the merging of telephony and computers, and with the extensive use of speech in man-machine communications, the need to recognize a person by his or her voice is constantly increasing. Applications of speaker recognition are wide ranging, including computer access control (Naik and Doddington, 1987; Higgins et al., 1991), telephone voice authentication for banking access, intelligent answering machines and law enforcement (Forensic speaker identification).

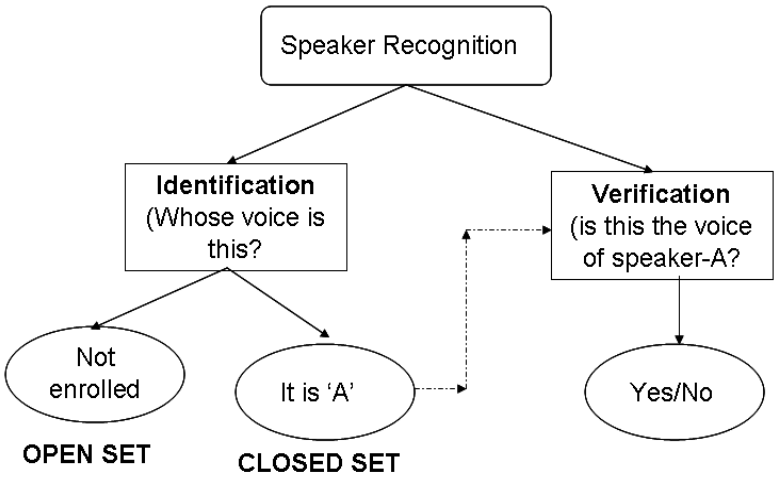

Figure 1: Schematic representation of speaker recognition.

Speaker recognition is defined as any decision making process that uses the speaker dependent features of the speech signal (Hecker, 1971). It can broadly be categorized into two specific tasks:

- Speaker identification
- Speaker verification

In speaker identification, the goal is to determine which one of a group of voice samples (reference sample) best matches the input voice sample (test sample). Whereas, the goal of a speaker verification task is to determine if a person is who he claims to be, from his voice sample (Figure 1).

Based on the content used for speaker identification or verification, the tasks can further be classified as

- Text-dependent, where the speaker's identity is dependent on the text uttered
- Text-independent, where no constraints are placed on the text uttered

Speaker recognition, according to Hecker, (1971) can be accomplished by the following methods

- Aural perceptual method
- Visual examination of spectrograms
- Machines

Identifying a speaker by listening refers to ***aural*** / ***perceptual speaker identification.*** Mc Gehee (1937) conducted one of the first studies in the area of aural speaker identification and reported an 83% correct identification of the individual up to 2 weeks, which slowly deteriorated to 35% after 3-5 months. Subsequently, similar studies were carried out by Bricker and Pruzansky (1966), the results of which followed a similar trend. Aural perceptual speaker identification is influenced by several factors such as familiarity, training, disguise, length and quality of sample, etc. Extensive studies conducted by Hollien, Mc Gehee, Schwartz state that a listener identifies a voice/differentiates one voice from another based on: Pitch, dialect, voice quality, articulation, rate of speech, stress, intonation, rhythm.

Speaker identification by visual comparison of spectrograms came into use after the invention of Sonagraph (is) an instrument which converts speech signal into a visual display. Here,

spectrograms of different utterances of a given phrase are presented to a trained observer, who attempts to determine whether some utterances were produced by a common speaker. Kersta (1962) claimed that identification of a speaker using spectrograms was an extremely efficient method, yielding an error rate of 1 %. However, further studies conducted by Stevens (1971), failed to validate this claim, and it was found that identification by visual comparison of spectrograms was influenced by several speaker related factors, phonetic context, length of the utterance, and other environmental/recording related factors.

Speaker identification using machines can be categorized into automatic and semi-automatic methods. Semi-automatic methods require human involvement for the decision making process, whereas in automatic methods, the decision is arrived at by the program. An automatic speaker recognition system goes through the following process

- Feature extraction
- Pattern Matching and
- Classification

In the primary process of feature extraction, several feature vectors are extracted from the speech samples for comparison. Researchers, in the past, have employed Fundamental frequency (F0), Formant Frequencies, F0 contour, Linear Prediction Coefficients (LPC) (Atal, 1974; Imperl, Kacic & Hovert, 1997), Cepstral Coefficients (CC) (Jakkar 2009; Medha, 2010 and Sreevidhya, 2010) and Mel-Frequency Cepstral Coefficients (MFCC) (Plumpe, Quateri & Reynolds, 1999; Hassan, Jamil, Rabbani & Rahman, 2004; Chandrika, 2010; Tiwari et al., 2010) for speaker identification.

Atal (1974) examined several parameters using linear prediction model for their effectiveness for automatic recognition of speakers from their voices, and deduced that cepstral parameters produced an identification accuracy of 70% for 50 msec of speech data. The accuracy further increased to 98% for duration of 0.5 sec. Several studies on Indian languages using cepstral parameters produced correct identification of around 80%.

A further improvement on the Cepstral Coefficients, namely Mel-frequency cepstral coefficients, is being used widely in automatic speaker identification. Cepstral coefficients are derived from the log spectrum represented on a Hz scale, whereas Mel-frequency cepstrum is mapped onto the Mel-scale prior to obtaining the log of a Fourier transform. The Mel scale is modelled based on the human auditory system. Several studies have been conducted by Eatock and Mason, 1994; Miyajima, 2001; Plumpe, Quateri and Reynold, 1991; Tiwari, 2010 based on MFCC. Although several feature vectors have been used in the past for speaker identification, CC and MFCC have been found to be the most efficient ones.

Nasals are a class of consonant sounds that comprise 11% of the phonemic content of English. They are produced when the glottal source is further modified by the resonance characteristics of an open nasal tract and a closed oral tract. All nasals are considered to be voiced, and they can be either released (word-initial and word-medial positions) or unreleased (word-final position). It is known that the availability of the speech contents used for speaker identification differs depending on the types of sounds they contain, and it is reported that voiced sonorants, such as vowels and nasals, are most effective for speaker identification by both humans (Matsui, Pollack and Furui, 1993; Sambur, 1975; Amino, 2004) and machines (Nakagawa and Sakai, 1979).

The effectiveness of the nasals in speaker identification can be explained by the uniqueness of the morphology of the resonators. It is reported that the shapes of the nasal cavity and paranasal sinuses are different among individuals (Dang and Honda, 1996). Also, the shapes of these resonators cannot be altered voluntarily. Differences in the timing of the velic action may be another factor that differentiates the nasals from oral sounds (Engwall, Delvaux and Metens, 2006), and this is something that the speakers cannot intentionally or voluntarily control by themselves. This is why the acoustical properties of the nasal sounds are of relatively stable nature, and thus stably reflects speaker's individuality. Nevertheless, it is also to be noted, that production of nasal sounds can be affected because of an upper respiratory tract infection, accumulation of mucus or pus in the nasal cavity, etc.

Perceptual studies conducted by Amino and Arai (2009) and Amino, Sugawara and Arai (2006) showed that stimuli including a nasal were effective cues for speaker identification. Glenn and Kleiner (1968) hypothesized that the power spectrum produced during nasal phonation is idiosyncratic to an individual. The results of their experiment revealed a 97% correct identification with the nasal /n/. Indian studies conducted using MFCCs on nasal continuants in Hindi and Malayalam (Ridha, 2014; Lekshmi devi, 2012), and nasal coarticulation in Malayalam (Jyothsna, 2011) have shown 100%, 95% and 90% correct identification using MFCC, respectively.

Of the four literary languages of the Dravidian group, Tamil enjoys the greatest geographical extension, has the richest and most ancient literature, and paralleled in India only by Sanskrit. It is spoken by 39,400,000 people (1981 est.) in the Indian state of Tamil Nadu, by another 2,697,000 in Sri Lanka (Ceylon), by smaller numbers of people in Burma, Malaysia,

Indonesia, and Vietnam (about 1,400,000). The percentage of occurrence of vowels and consonants in Tamil are 48.7% and 51.23% respectively. Among consonants (Table 1), the nasals /m/, /n/ and /ɳ/ (retroflex) have 4.7%, 2.2%, and 0.7% of representation in the language (Rajaram, 1972).

Tamil has the following nasal sounds:

Phoneme	**Place of articulation**	**Tamil script**
/m/	Bilabial	ம
/ŋ/	Velar	ங
/n̪/	Alveolar	ன
/ñ/	Palatal	ஞ
/ɳ/	Retroflex	ண

Table 1: Nasal phonemes in Tamil language.

Based on a review of literature in the area of speaker recognition, it is evident that there is a dearth of research in the area of speaker identification in Tamil. Also, there is no report of establishment of a benchmark for nasal continuants in Tamil language. Therefore, the present study aims at examining speaker identification using nasal continuants in Tamil.

Objectives of the Study

- To provide a benchmark for speaker identification in Tamil nasal continuants using MFCC.
- To compare speaker identification scores obtained using live recording and mobile network recording

REVIEW OF LITERATURE

Forensic speaker identification or forensic speaker verification is the term given to the legal process by which one identifies if two or more recordings of speech are from the same speaker. (Rose, 2002). Primitive efforts at speaker identification date back to the year 1660, when voice identification was offered in the case of 'William Hamulet'. Identification of a perpetrator by means of his voice was accepted as testimony in a court in Florida as early as 1907.

The most common task in forensic speaker identification involves the comparison of one or more samples of an unknown voice (sometimes known as the questioned sample) with one or more samples of a known voice. The unknown voice often belongs to the individual alleged to have committed an offence and the known voice belongs to the suspect. Both prosecution and defence are then concerned with being able to say whether the two samples have come from the same person, and thus being able either to identify the suspect as the offender or to eliminate them from suspicion (Rose, 2002).

A number of important speaker identification related events occurred during World War II. When the world was in a state of confusion as to whether Adolf Hitler was alive or had escaped Germany, Hitler's previously recorded speeches proved extremely useful. A team comprising of several phoneticians and engineers was appointed in order to compare Adolf Hitler's old and new recorded speeches. After a series of analysis, they arrived at the conclusion that Hitler was still alive then (Hollien, 2002).

Paul Prinzivalli, an air freight cargo handler in Los Angeles was tried for having threatened his employer, Pan Am. However, he was acquitted because the forensic-phonetic analysis conducted on the offender's voice samples clearly determined the differences in their dialects (Labov and Harris, 1994).

In 1987, the voice samples of a suspect helped identify the kidnapper of an 11 year old German girl. Forensic voice analysis revealed several similarities between the suspect's and the kidnapper's voice (Künzel, 1987).

In the late 1990s in Australia, the police intercepted 15 incriminating telephone conversations concerning illicit drug trafficking (Duncan-Lam, 1999). Forensic-phonetic analysis was able to assign the voice samples from the conversation to three different speakers. Hence, speaker identification/recognition is widely accepted as evidence as one of the biometrics evidence.

Speaker recognition has been defined as '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' (Atal, 1976). There are two main classes of speaker recognition task, called identification and verification (Furui, 1994; Nolan, 1997). The primary differences between them include

- The type of question asked
- The nature of the decision-making task involved

The Figure 2 shows the classification of speaker recognition.

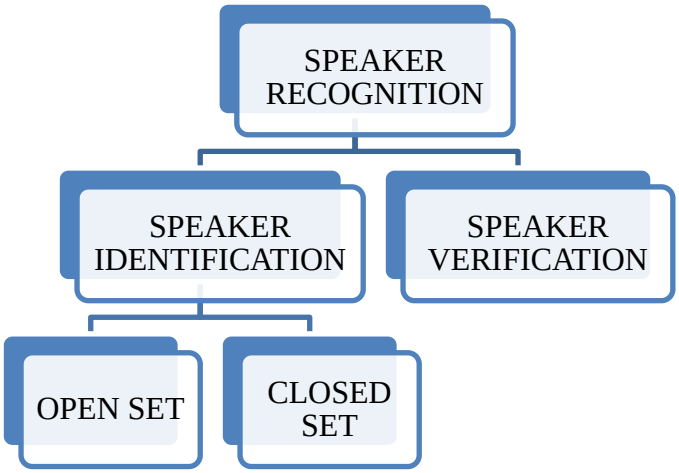

Figure 2: Methods of speaker recognition.

Speaker identification

The aim of speaker identification is 'to identify an unknown voice as one or none of a set of known voices' (Naik, 1994). For example, one has a speech sample from an unknown speaker, and another of speech samples from speakers, whose identity is known. The task of speaker identification is to compare the unknown sample to each of the known samples and determine if it matches with the set of known samples, and if it does, to which one. Figure 3 shows the schematically represents speaker identification.

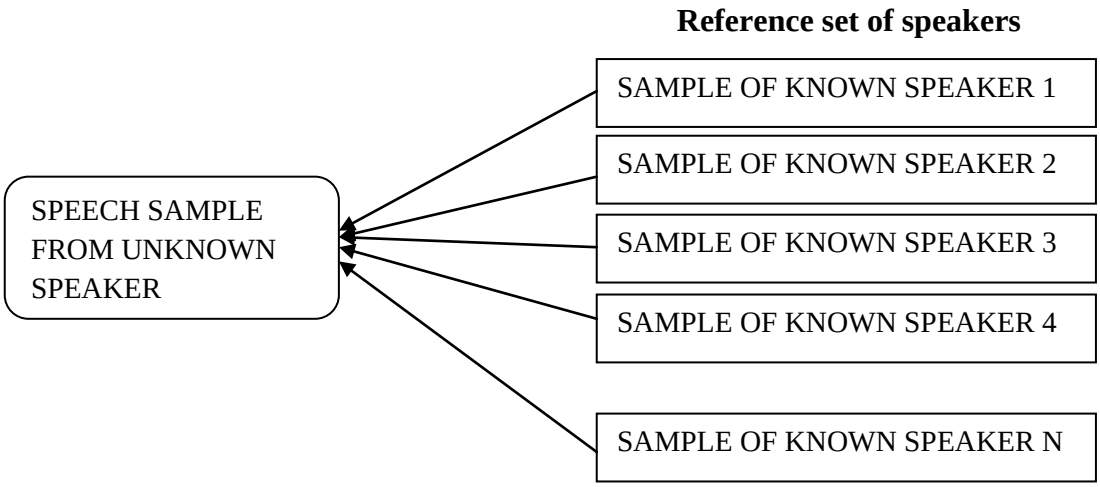

Figure 3: A schematic representation of speaker identification.

In speaker identification, the reference set of speakers can be of two types

- Closed set
- Open set

In the closed set speaker identification task, one knows that the unknown sample definitely matches with one of the known references samples. However, in open set speaker identification, the unknown speaker may or may not belong to the set of know speakers. Closed set speaker identification is a much easier task than open set identification. The closed set identification task lies in

- Estimating the distance between the unknown speaker and each of the known reference speakers
- Picking the known speaker that is separated by the least distance from the unknown speaker.

The pair of samples separated by the smallest distance is then assumed to be from the same speaker (Nolan, 1983).

In an open set identification, one cannot assume that the pair of samples separated by the smallest distance is automatically the same speaker. In order to state that, one needs to have a pre-existing threshold, so that, the distance separating a pair of speakers, when below the threshold can be stated as belonging to the same speaker. In forensic case-work both open and closed sets can occur, however, the former is more common.

Speaker verification

Speaker verification is another common task in speaker recognition. Here, An identity claim from an individual is accepted or rejected by comparing a sample of his speech against a stored reference sample by the individual whose identity he is claiming' (Nolan, 1983). Figure 4 provides a schematic representation of speaker verification.

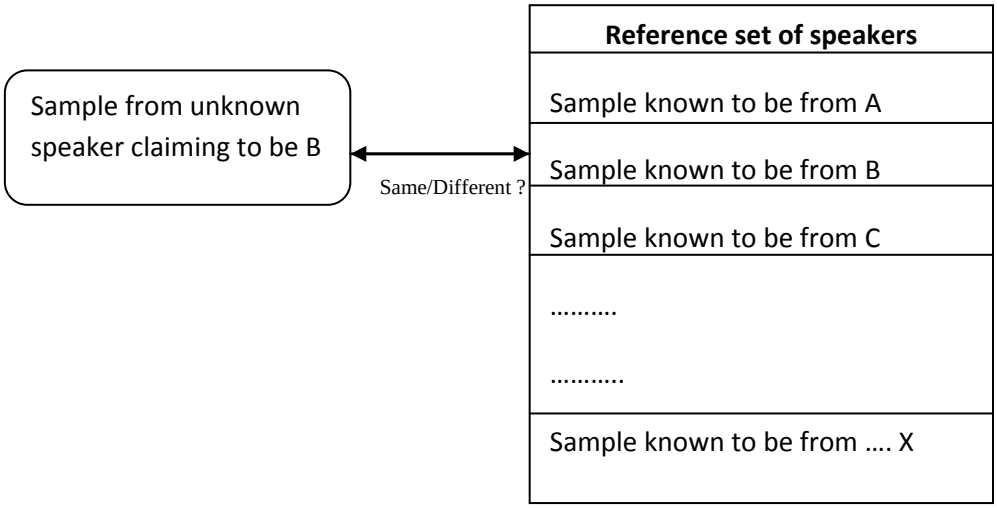

Figure 4: Schematic representation of speaker verification.

Type of Decision

In speaker identification, only two types of decision are possible. They are:

- The unknown test sample is has been identified accurately
- The unknown sample has not been identified.

However, in speaker verification, four types of decision are possible. They are:

- The speaker is correctly identified as who he claims to be
- The speaker is correctly rejected i.e., the speaker is not whom he claims to be
- The speaker is incorrectly accepted (The speaker is not whom he claims to be, in reality, however the speaker recognition system accepts him)

- The speaker is incorrectly rejected (The speaker is who he claims to be, in reality, however, the system incorrectly rejects him).

Figure 5 illustrates the types of errors encountered in speaker recognition.

Classification of errors in speaker recognition

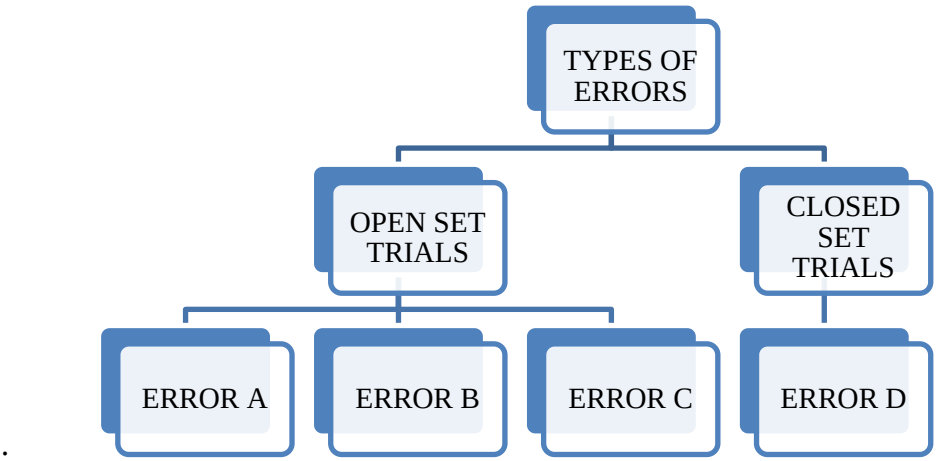

Figure 5: Types of errors encountered in speaker recognition (Tosi et al, 1972).

ERROR A (False identification): A match existed but the examiner selected the wrong one.

ERROR B (False elimination): A match existed although the examiner failed to recognize it.

ERROR C (False identification): No match existed, although the examiner selected one.

ERROR D (False identification): In closed set speaker identification, a match definitely exists; therefore only one type of error is possible. In error type D the examiner selects the wrong one.

Text dependent and text independent speaker recognition

The distinction between text dependent and independent speaker identification is essential in the process of speaker recognition. In the text-dependent condition, the same text / the same words or utterances is used for both training the recognition device and testing it. In the text-independent condition, the recognition device does not require that the lexical content used for training and test remain the same. Generally, text-dependent speaker recognition performs better than text-independent speaker recognition (Nakasone and Beck, 2001).

Between-speaker and within-speaker variation

Speech is a dynamic behaviour, which is subject to constant variation/change. It is known that the pronunciation of a given word/phrase or the way it is produced tends to vary from speaker to speaker. This is known as **inter-speaker or between-speaker** variability.

Speech is the result of co-ordination between various physiological systems such as respiratory, laryngeal, articulatory, resonatory systems. Therefore, a given utterance produced by the same speaker twice is never exactly the same due to slight variations in the performance of each system from utterance to utterance. This is known as **intra-speaker or within speaker** variability. The success of any method of speaker recognition depends on the degree to which the sampled inter-speaker variability is greater than the intra-speaker variability.

Some factors contributing to intra and inter-speaker variability are given in table 2.

INTER-SPEAKER VARIABILITY	INTRA-SPEAKER VARIABILITY
Anatomical & Physiological: Size of vocal tract vocal fold morphology and mass Mass and movement characteristics of articulators **Social & Experiential:** regional accent Native language, dialect, exposure to other languages Socio-economic status	Context Age Emotion Accent Disguise Status of health

Table 2: Factors contributing to intra and inter-speaker variability.

Speaker recognition can be accomplished by the following methods

- Aural-perceptual
- Visual examination of spectrograms
- Machine recognition (semi-automatic and automatic methods)

Speaker identification by Aural-Perceptual method (AP-SPID)

It is one of the oldest methods used in speaker identification. This method typically involves aural presentation of reference samples (unknown samples) and a test sample (known sample) to the examiner. The reference samples consist of a line-up of the suspect's speech (obtained from recorded message, threat call, etc.) along with foil samples. The test sample is the suspect's speech sample, obtained at the time of interrogation (usually of the same text as the

reference). Trained voice experts who participate in the experiment are required to match the test sample with one of the references.

There are several factors that affect Aural Perceptual Speaker Identification. They are

- **Listener related**
    - Familiarity with the suspect's voice
    - Training in the area of voice identification
    - Hearing sensitivity
    - Memory or the ability to remember the voice/speech characteristics of a reference sample and accurately match it with the test sample.

Mc Gehee in 1937 conducted a study in which she selected sets of speakers from a pool of 49 individuals (31 males and 18 females). One of the speakers was to orally read a 56-word passage standing behind an opaque screen. There were fifteen groups of listeners all of which consisted of college students. Initially, each group of listeners heard the speaker read the passage. Subsequently, a voice line-up was arranged in which five foil speakers were also present. All speakers including the foils read the same passage standing behind the opaque screen. The listeners wrote down the number of the speaker they thought they had heard originally. The procedure was repeated after a couple of days, weeks and months. The percentage of correct identification scores was 83% after an interval of 1 day, which was sustained for a week. The scores dropped to 68% after 2 weeks, to 35% after 3 months, and to 13% after 5 months. Bricker and Pruzansky in 1966 reported a similar trend of decline in scores over time.

Hollien, Majewski and Doherty (1982) conducted studies on the effect of familiarity with suspect's voice on speaker identification and reported that participants were able to identify a familiar voice even under difficult conditions. Hollien (1990) stated that "a fairly good rule of

the thumb for establishing the familiarity of a listener with a talker is that they should have good hearing and have heard the target speaker's voice fairly regularly over a period of around 2 years".

- **Speaker-related**
    - Unique speech characteristics: voices that have unique characteristics are easier to identify
    - Disguise: Depending on the type of disguise used by the perpetrator, his voice may or may not be easily identifiable.
    - Stress, emotions
    - Accents, dialects

Reich and Duke (1979) conducted a study on the effect of disguises on speech recognition and determined that free disguise and disguise in the form of a strong nalsalized speech were the most damaging of all disguises.

- **Speech sample related**
    - Length and quality of sample
    - Environment in which samples have been recorded
    - Contemporary vs non-contemporary samples

Rothman (1977) from his study determined that speaker identification scores dropped to 42% when the samples were non-contemporary. Several authors such as Künzel (1995) and Pollack, Pickett & Sumbey (1954) state that atleast 30 seconds of speech sample is required for speaker identification tasks.

System distortion or signal degradation can also contribute to poor speaker identification. Devices such as a telephone may limit the frequency response of the speech signal, thereby eliminating important information from the sample. Other factors such as a noisy environment,

limited frequency response of the microphone used for recording, etc can affect the quality of the speech signal to be analyzed resulting in erroneous results.

The fact that mobile phones can be used almost anywhere means that many types of background noise will be encountered with recordings made from mobiles. Also, aspects of speaker behaviour may differ when mobiles are used. It is clear from casual observation that mobile users have a tendency to speak loudly. Mc Clelland (2000) noted that F0 (Fundamental Frequency) can be as much as 30 Hz higher than F0 in landline calls made by the same speaker. Thirdly, recordings from mobile calls are often affected by GSM radio transmissions. These introduce an interference signal characterized by a fundamental frequency of 217 Hz, plus higher frequency harmonics overlapping the frequency range of speech. It has also been reported by (Künzel, 2001) that the upper frequency cut off for GSM transmission is lower than that for landline transmission at 3,200 Hz.

Speaker identification by visual inspection of spectrograms:

Speaker identification by visual inspection of spectrograms came into use after the invention of Sonograph by Bell Telephone Laboratories, U.S.A. The device first became available in the 1930s, and was initially used to provide the hard of hearing with an alternative way to learn speech. Sonagraph is the forerunner of the current day computerized spectrogram. It is a three dimensional display with time on the X-axis, frequency on the Y-axis and intensity on the Z-axis. In the case of speaker identification using spectrograms, the spectrograms of different utterances of a given word/phrase are presented to a trained observer, who attempts to determine whether some utterances were produced by a common speaker. Kersta (1962) in his paper ‘Voiceprint identification’ claimed that speaker identification using spectrograms was an extremely efficient method, yielding an error rate of less than 1%.

An elaborate study conducted by Tosi et al. (1972) which involved matching spectrograms. The study yielded a correct identification percentage ranging from 86% to 96%. The study also examined related issues such as number of cue words required for speaker recognition, effect of recording conditions, effect of context of cue words on speaker identification, contemporary vs non-contemporary samples, etc.

An attempt at benchmarking using spectrograms was undertaken by Pamela (2002). They examined the reliability of voice prints with the help of several extracted acoustic parameters. Six Hindi speaking male subjects participated in the study, and the target words were 29 bi-syllabic words consisting of 16 plosives, 5 nasals, 4 affricates and 4 fricatives in the word medial position. The acoustic parameters measured were formant transition duration, VOT, Closure duration, duration of phonemes. The results indicated that 67% of the measures varied across speakers and 61% of the measures varied within speakers. The effect of disguise on voiceprints was studied by Ranganathan (2003). The results revealed no significant differences between accuracy of speaker identification in disguised and normal conditions.

Speaker identification by machine

Since the 1970s, speaker identification by machines has become popular. It can be categorized into two types:

- **Semi-automatic** speaker recognition, where the examiner makes the interpretation of the results provided by the system
- **Automatic** speaker recognition, where the involvement of the examiner is minimal, and the system makes use of several algorithms in order to deduce who the speaker is, or whether the speaker is really who he claims to be.

There are two phases in the automatic speaker recognition process, namely the training and the testing phases. During the training phase, a large number of exemplar tokens/samples are collected for each speaker and stored as a database. During the testing phase, an utterance of the speaker is fed to the system and the speaker recognition system compares it with the stored database to determine the identity of speaker or verify speaker's identity. The automatic speaker recognition system goes through the following steps in order to arrive at a decision. They are

- Feature extraction
- Pattern Matching
- Classification

Features are certain acoustic parameters that characterize an individual's speech. A good feature:

- Must be highly discriminable across speakers
- Should vary minimally from session to session
- Must be difficult to impersonate.

One of the earliest approaches in 1972 compared speakers based on 17 parameter sets which included fundamental frequency, vowel and nasal consonant spectra, glottal source spectrum slope and word duration. Over the years, several feature vectors such as formant frequencies, Linear Prediction coefficients (LPC) (Atal, 1974; Imperl, Kacic & Hovert, 1997), Cepstral Coefficients (Jakkar, 2009; Medha, 2010 and Sreevidhya, 2010) and Mel-Frequency Cepstral Coefficients (Plumpe, Quateri & Reynolds, 1999; Hassan, Jamil, Rabbani & Rahman, 2004; chandrika, 2010; Tiwari et al., 2010) have been employed for speaker identification.

A **cepstrum** is the result of taking the Inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal. It was first adopted as a tool for automatic pitch detection by Noll (1964). In its most basic form, the system for producing voiced speech sounds consists of

the vocal source and vocal tract. The source signal s(t) is the periodic puffs of air emitted by the vocal cords. The effect of the vocal tract is completely specified by its impulse response h(t) such that the output speech signal f(t) equals the convolution of s(t) and h(t) (Figure 6). The effects of the vocal cords and vocal tract are therefore convolved with each other. In order to separate the source and filter, the fourier transform of the logarithm of the power spectrum is taken.

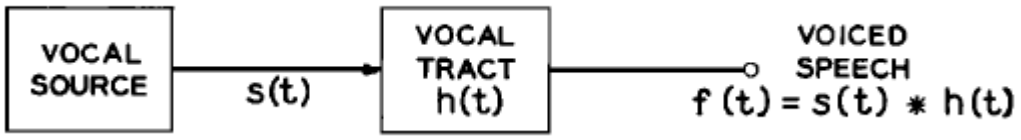

Figure 6: Basic system for production of voiced speech sounds.

The effect of the vocal tract is to produce a "low-frequency" ripple in the logarithm spectrum, while the periodicity of the vocal source manifests itself as a "high-frequency ripple in the logarithm spectrum. Therefore, the spectrum of the logarithm power spectrum has a sharp peak corresponding to the high frequency source ripples in the logarithm spectrum and a broader peak corresponding to the low-frequency formant structure in the logarithm spectrum (Figure 7). The peak corresponding to the source periodicity can be made more pronounced by squaring the second spectrum. This function, the square of the Fourier transform of the logarithm power spectrum, is called the "cepstrum".

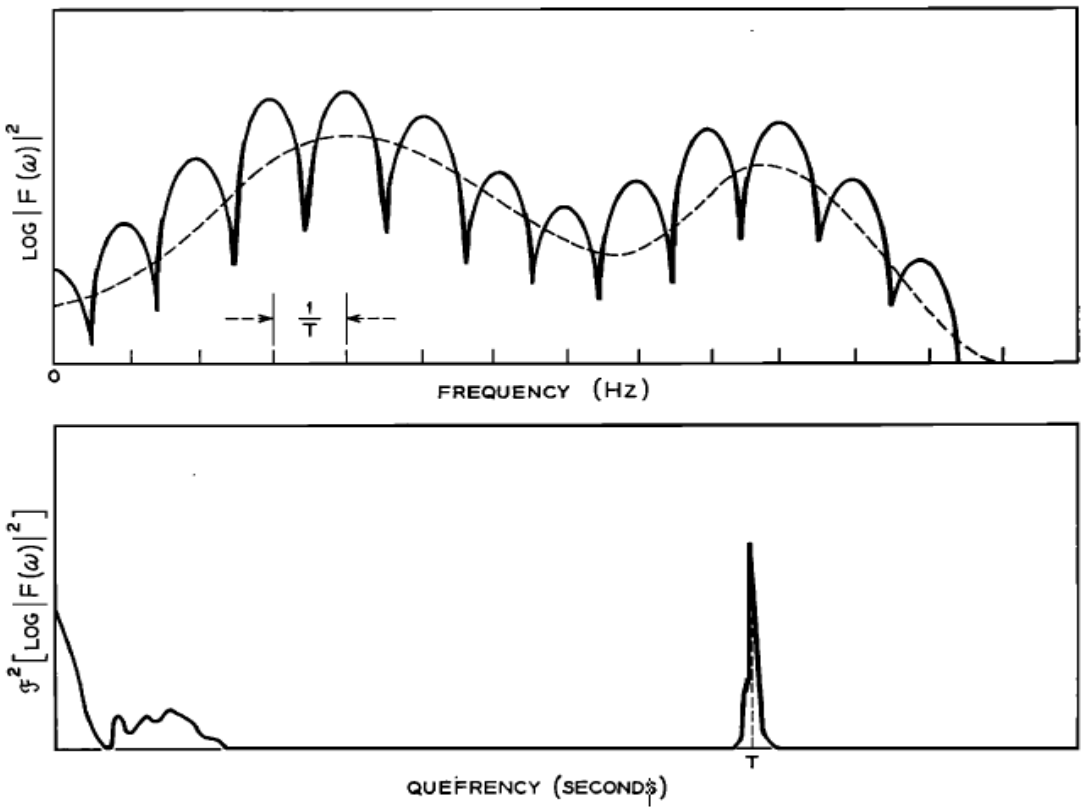

Figure 7: Schematic procedure for extraction of Cepstrum.

Luck (1969) used Cepstral measurements for automatic speaker verification. The standard test phrase 'My code is ____" was chosen for the study, from which several feature vectors were extracted for comparison. The verification decision was treated as two-class problem i.e., either the speaker is an authorized speaker or an impostor. Reference data was used only for the authorized speaker, and the decision was based on the distance between the test and reference samples. Four authorized speakers and 30 impostors were examined, with error rates lying between 6% and 13%.

Wolf (1972) examined the efficacy of several parameters extracted from the speech signal, in order to improve speaker recognition techniques. The choice of these parameters was based on considerations of vocal tract structure and the ways in which speech sounds were produced. A

simulation of speaker recognition system was performed by manually locating speech events within utterances and using parameters measured at these locations to classify the speakers. Useful parameters were found in fundamental frequency features
of vowel and nasal consonant spectra, estimation of glottal source spectrum slope, word duration, and voice onset time. These parameters were tested in speaker recognition paradigms using simple linear classification procedures. When only 17 such parameters were used no errors were made in speaker identification from a set of 21 adult male speakers. Under the same conditions speaker verification errors of the order of 2% were also obtained.

Glenn & Kleiner (1968) conducted an automatic speaker identification experiment using vectors obtained from nasal phonation. They chose nasal phonation over other classes of sounds due to the relatively fixed position of the oral tract, and the steady-state power spectrum generated by the open nasal tract. With an experimental population of 30 speakers, an accuracy of 93% was obtained. With an experimental population of 10 spekaers, an accuracy of 97% was obtained. The results of the study supported the hypothesis that nasal phonation provided a strong clue to speaker identity. The procedure followed in the study provided a basis for automatic speaker identification, in the absence of detailed knowledge of the message spoken. The power spectra of nasal consonants (Glenn and Kleiner, 1968) and co-articulated nasal spectra (Su et al., 1974) provide strong cues for speaker recognition by machines.

Mel-scale Cepstrum: The standard practice is to represent the log spectrum with frequency axis in Hz scale. It is possible to compute the log spectrum with the frequency axis in Mel-scale (or in bark-scale or in logarithmic scale). Mel-Frequency analysis of speech is based on human perception experiments. It is observed that human ear acts as filter, which does not follow a

linear scale. It concentrates on only certain frequency components. These filters are non-uniformly spaced on the frequency axis, such that more filters are present in the low-frequency regions and fewer filters in the high frequency regions (Figure 8).

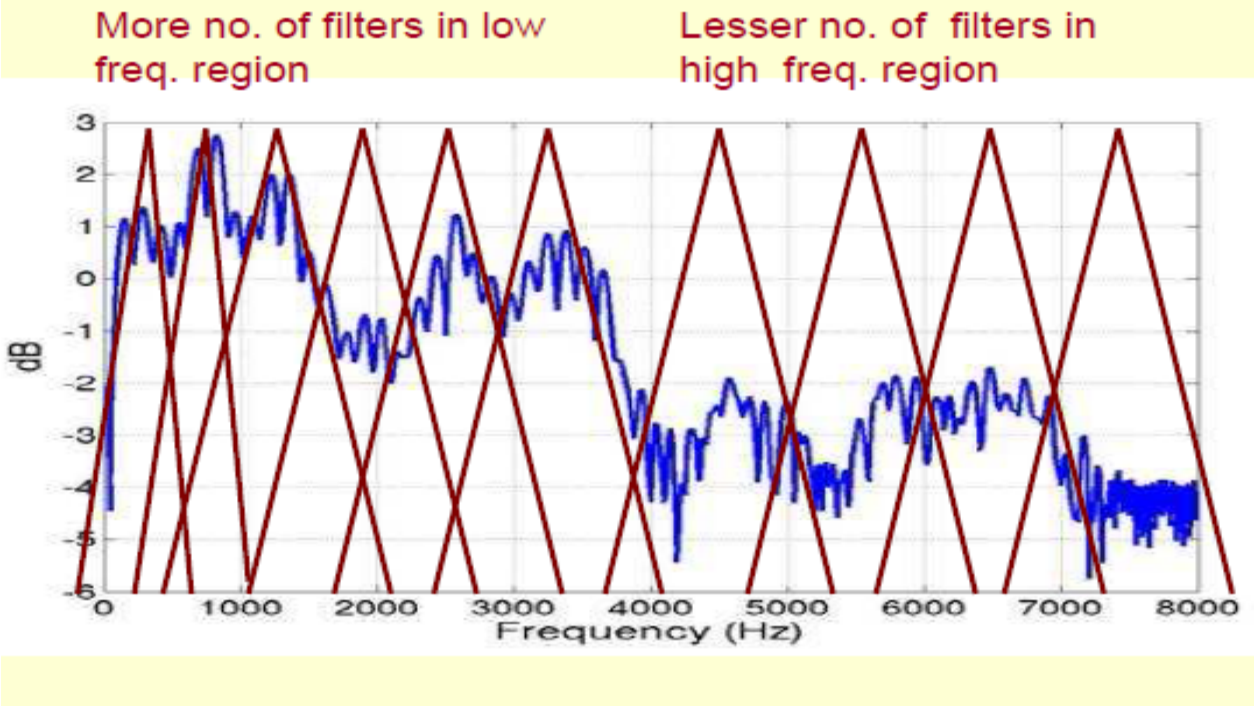

Figure 8: Distribution of frequency across the Mel-Scale.

Thus for each tone with a physical measure of frequency in Hertz, there is a corresponding subjective measure on the Mel scale. Since, Mel-Scale more closely resembles the way the auditory system analyzes sound, it has been used widely in research pertaining to speaker recognition.

Hasan, Jamil, Rabbani and Rahman (2004) conducted a speaker identification experiment using Mel-Frequency Cepstral Coefficients. Vector quantization was used to minimize data of the extracted feature. The study revealed that as the size of the codebook (number of centroids) increases, the accuracy of identification increases. The study concluded that a combination of

Hamming window and Mel-frequency Cepstral Coefficients gave the best results. The results also showed that a linear scale can also have reasonable identification rate if a higher number of centroids were used.

Pruthi and Epsy-Wilson (2007) extracted acoustic parameters from nasalized vowels for automatic detection and reported accuracies of 96.28%, 77.9% and 69.58% using StoryDB, TIMIT and WS 96/97 databases respectively.

Chandrika (2010) studied the efficacy of a speaker verification system using speech recorded over a mobile network and digital recording. 10 subjects participated in the study and the Mel-frequency Cepstral Coefficients obtained from long vowels /a: /, /i:/ and /u:/ were analyzed. Results indicated an overall verification of 80% and that the vowel /i:/ performed better than the other two vowels.

Ramya (2011) using long vowels /a:/, /i:/ and /u:/ , provided a benchmark for speaker identification for electronic vocal disguise for females, using MFCCs. The results showed, correct identification percentage of 96.6%, 93.3% and 93.3% for the vowels /a: /, /i:/ and /u:/ respectively.

Ridha (2014) conducted a study using MFCCs derived from Hindi nasal continuants. The study was carried out with 10 participants using both live recording and mobile network recording. The nasals chosen for the study were bilabial /m/, dental /n/ and velar /ŋ/ embedded in words in all positions. The percentage of correct identification obtained when the live recording was compared with live recording were 100%, 90% and 100% for the nasals /m/, /n/ and /ŋ/. The accuracy when mobile network recordings were compared with mobile network recordings was 50%, 80% and 90% respectively.

Thus, the above studies support the extraction of MFCCs over other parameters for experiments in speaker recognition

METHOD

Participants: Twenty male participants between the ages of 20 and 40 years were participated in the present study. The participants were native speakers of Tamil language and they were living in Coimbatore (a district in the western part of part of Tamil Nadu). The participants selected were either graduates, or had completed schooling, hence were proficient in reading, writing and spoken Tamil. Also, they had

- No history of speech, language and hearing difficulties
- Normal oral structures and
- No other associated psychological and neurological problems.

Material: Ten meaningful Tamil sentences, relating to common messages in a threat call, were chosen for the study. These sentences consisted of the nasal sounds /m/ (bilabial), /n/ (alveolar) and /n̪/ (retroflex) in the initial, medial and final positions of words in the sentences, and the sentences were derived based on the colloquial/informal spoken language. The frequency of occurrence of the nasals /m/, /n/ and /n̪/ in the above sentences are 7, 9 and 7 times respectively. Out of these occurrences, the five best ones were chosen for analysis. The subjects were asked to repeat each sentence thrice at habitual pitch, loudness and rate. They were specifically instructed not to adopt a strict reading style, instead asked to adopt a casual conversational style while reading out the sentences. The current study was considered as kind of text- independent study because the same nasals sounds were selected from different phonetic environment, which means that a nasal continuant from different phonetic environments was compared. For e.g., /m/ in

‘**m**arubadijum fo:n panna ma:tte:n’ could have been compared with the /m/ in ‘pe:sa:ma na:n solra **m**a:dhiri ke:ɻu’.

The sentences were as follows

Sl.No.	Sentences written in English	Sentences in Tamil
1	na:lu latʃam ve:ɳum	நாலு லட்சம் வேணும்
2	marubadijum fo:n panna ma:tte:n	மறுபடியும் ஃபோன் பண்ண மாட்டேன்
3	ka:laila paɳam vandhu se:raɳum	காலைல பணம் வந்து சேரணும்
4	po:li:sukku po:na avlodha:n	போலீசுக்கு போனா அவ்ளோதான்
5	va:ja mu:du	வாய மூடு
6	pe:sa:ma na:n solra ma:dhiri ke:ɻu	பேசாம நான் சொல்ற மாதிரி கேளு
7	ni: mattum paɳatho:da va:	நீ மட்டும் பணத்தோட வா
8	pathu maɳija:chu	பத்து மணி ஆச்சு
9	na:laikku dha:n kadaisi na:ɻ	நாளைக்கு தான் கடைசி நாள்
10	paɳam ke:ttu romba na:ɻa:chu	பணம் கேட்டு ரொம்ப நாளாச்சு

Table 3: Stimuli used for the present study.

Recording procedure: All the recordings were done in the participants’ natural environment (field recording). Two types of recordings live recording & mobile network phone recording were carried out.

The live recording was done using an Olympus LS100 (Olympus America Inc.) recorder. It had sampling frequency of 96 kHz and 24 bit rate resolution. The participants were seated comfortably, and were asked to familiarize him with the sentences. Then they were asked to read out the entire set of 10 sentences five times. This was recorded with the recorder, which was held around 10cm away from the participants. Simultaneously, a mobile network phone call was placed to the participants, from another room using a smart phone (Nokia ASHA, 301). The participants in turn received the phone call through another smart phone (HTC Wildfire S). Therefore, the sentences read out by the subject were being recorded using the Olympus recorder and the smart phone (Nokia ASHA 301) at the same time simultaneously.

During the recording procedure, the participants were required to read the entire set of 10 sentences five times in a habitual speech rate with comfortable pitch and loudness. Out of these five repetitions, the two best repetition sets were selected. Further, out of these two sets, ten of the best occurrences (5 from each set) of each nasal (/m/, /n/ and /ɳ/) were segmented for analysis. The examiner was aware that the “unknown” speaker is one among the “known” samples; hence it was a closed-set speaker identification task.

Analysis: The recorded samples were transferred to computer memory and analyzed using SSL-Workbench for Semi-Automatic vocabulary dependent speaker recognition (Voice and Speech Systems, Bangalore, India) software. The phone network recording samples were down-sampled to 8kHz and live recording samples to 16kHz for analysis. Mel-Frequency Cepstral Coefficients were extracted from the samples and compared. Figure 1 depicts the analysis window of SSL Workbench software, which is followed by an explanation of each components involved in the initial phase of analysis.

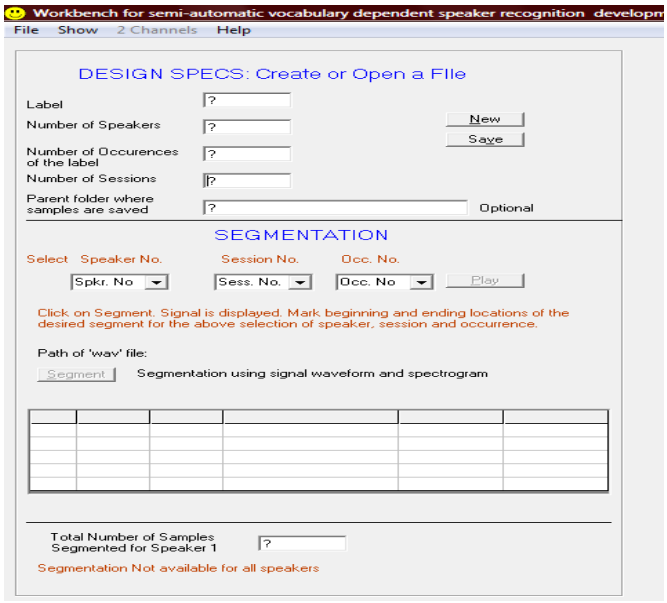

Figure 9: Analysis window for SSL workbench.

The SSL workbench analysis window is required to be filled before starting the segmentation.

The basic terms glossary used are given below

- **Label:** the phoneme or sound being analysed e.g (/m/, /n/ and /n̥/ in the present study)
- **Number of speakers:** the number of participants in the study (20 in the present study)
- **Number of sessions:** number of repetitions of the stimulus (2 in the present study)
- **Number of occurrences of the label:** the frequency of occurrence of a sound in a particular stimulus (In the present study, the five best occurrences of each nasal/m/, /n/ and /n̥/, in both repetitions)

The above information is entered in the system and stored as a text file which in term creates dbs file automatically.

Procedure to compute Euclidean Distance in SSL workbench software

Each of the 20 participants were asked to repeat the set of 10 sentences five times, out which, two of the best sets were chosen. Five occurrences of every nasal (/m/, /n/ and /n̥/), were chosen from both the sets. A portion of the nasal phonation (min 30 msec), with a total of 10 occurrences/segments in both sessions were segmented and stored with help of visual inspection of spectrogram. Therefore, every speaker was represented by 10 segments of nasal phonation for each of the nasal sounds (/m/, /n/ and /n̥/). The details of segmentation for every speaker are stored in the dbs file, as represented in figure 10.

File Edit Format View Help

speaker No	Occ. No.	Sess. No.	FileName	From	To
1	1	1	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_1.wav	2.028	2.085
1	2	1	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_1.wav	3.780	3.819
1	3	1	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_1.wav	4.336	4.378
1	4	1	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_1.wav	6.295	6.350
1	5	1	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_1.wav	6.996	7.067
1	1	2	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_4.wav	2.906	2.965
1	2	2	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_4.wav	3.242	3.292
1	3	2	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_4.wav	5.240	5.282
1	4	2	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_4.wav	7.191	7.260
1	5	2	H:\NITHYA-PROJECT\GAUTHAM\PHONE\REC_4.wav	11.647	11.698
2	1	1	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_2.wav	2.634	2.676
2	2	1	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_2.wav	6.594	6.631
2	3	1	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_2.wav	7.555	7.588
2	4	1	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_2.wav	8.045	8.101
2	5	1	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_2.wav	9.206	9.259
2	1	2	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_3.wav	6.771	6.818
2	2	2	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_3.wav	7.894	7.947
2	3	2	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_3.wav	8.396	8.440
2	4	2	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_3.wav	9.372	9.426
2	5	2	H:\NITHYA-PROJECT\VIJAY ANAND\PHONE\REC_3.wav	12.835	12.881

Figure 10: The dbs file with details of segmentation for every speaker

In Figure 10, the segmentation of the each sound with the starting duration and the ending duration were given. The details were given below

Speaker no: represents the speaker selected (1,2,3, etc…)

Session no: 2 (in the present study)

Occurrence no: 5 occurrences for each session, with a total of 10 occurrences for every speaker for the present study

File Name: Name of the file corresponding to the speaker chosen, with details of the drive in which the samples are stored (e.g., H-Drive)

From & To: The duration of each segmented nasal sound.

SSL–Workbench speaker identification system requires a set of training samples, representing a speaker in order to identify the speaker accurately when provided with a test sample. Therefore, upon completion of segmentation, the number of test and reference samples was designated as 3 and 7 respectively. The software randomly assigns the occurrences as reference and test (For e.g., Reference : 1,2,4,7,9,6,10 and Test: 3,5,8). These combinations can be varied at random by clicking on the 'Randomize Training Samples' button in the analysis window (indicated by an arrow mark in Figure 11).

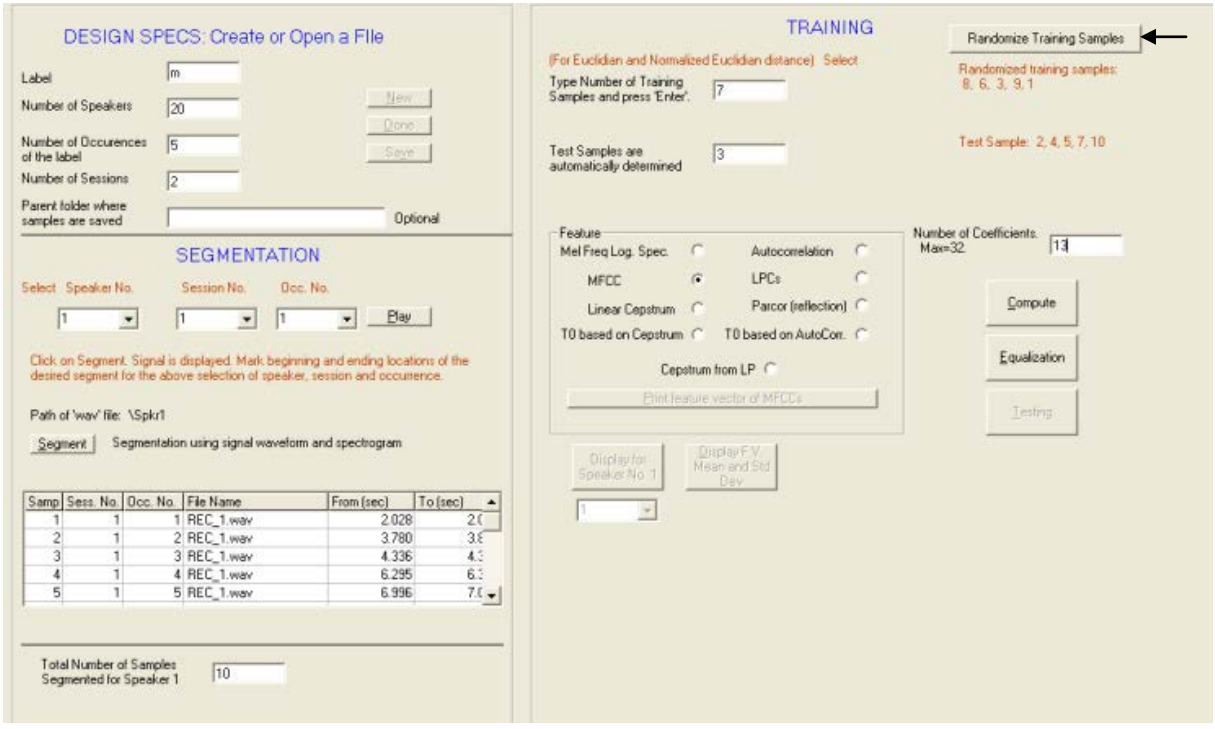

Figure 11: Training window of SSL Workbench depicting the number of reference and test samples selected.

The system has option of selecting several feature vectors (e.g., MFCC, CC, LPC, etc), any of which can be chosen for comparison of the samples. For the present study, the feature vector chosen was MFCC with 13 coefficients. Upon choosing the feature vector, the system computes a measure of distance (Euclidean distance) and displays the summarized distance matrix for the selected test and reference sample. From the distance matrix (Figure 13), the total percentage of correct speaker identification score is displayed (indicated by the arrow mark in figure 12).

Euclidean distance (ED): It is an ordinary distance between two points and is a measure of similarity or dissimilarity. Euclidian distance within and between participants is be noted. If the ED distance between the test sample and corresponding reference sample is least, then the identification was considered as ***correct identification/same speaker.*** Anything above the least distance was considered as a ***different speaker.*** The percent correct identification was calculated using the following formula:

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

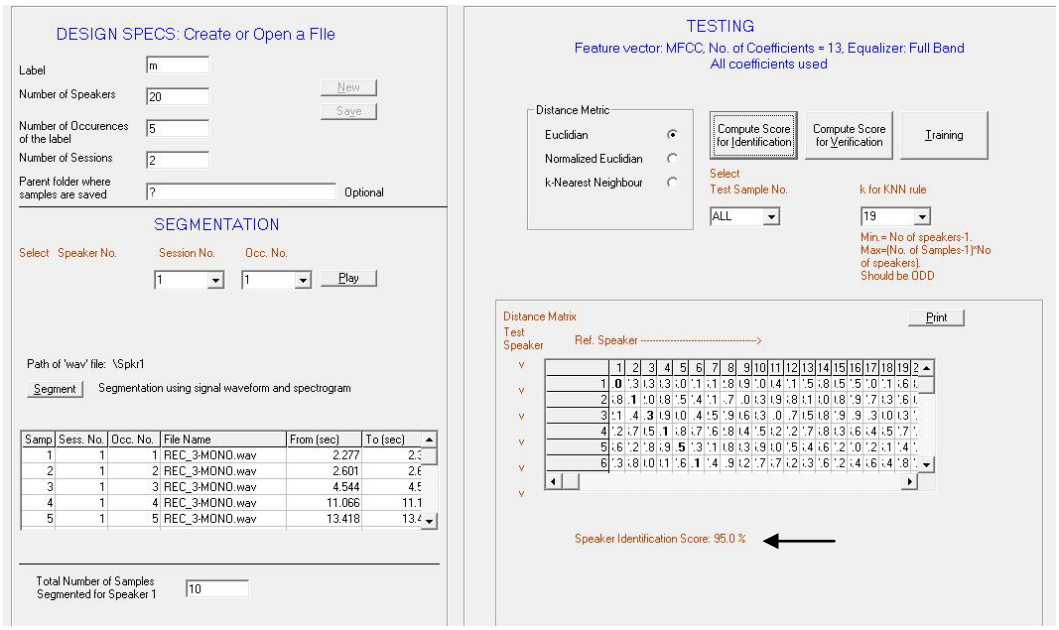

Figure 12: Training window of SSL Workbench depicting the percentage of correct identification.

In the present study the variables considered are the effects of various nasals on speaker identification. Analysis was done on the following ways

- Live recording (test) was compared with other live recording (reference)
- Mobile network recording (test) was compared with other mobile network recording (reference)

Percentage of correct identification was calculated for the nasals /m/, /n/ and /n̪/ in the following conditions

- Live recording Vs Live recording
- Mobile network recording Vs Mobile network recording

The percentage of correct identification for the three nasals /m/, /n/ and /n̪/ were examined using distance matrix.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	3.614	5.618	5.287	6.584	4.836	5.660	4.165	9.884	12.765	8.092	11.644	11.100	8.085	4.267	6.823	5.394	10.798	12.055	7.912	8.692
2	8.530	3.995	8.085	6.614	7.481	6.970	5.081	7.649	8.731	6.383	7.391	7.134	6.202	8.248	9.238	6.216	7.655	8.950	5.605	6.591
3	4.797	8.360	2.802	6.691	3.778	7.033	7.234	8.896	14.248	8.957	13.097	12.676	9.421	4.720	4.719	7.722	12.820	12.914	9.435	9.606
4	7.746	6.704	6.201	2.988	5.559	5.440	6.486	8.236	11.285	8.480	9.218	9.799	8.808	6.585	7.098	5.528	9.737	10.205	8.456	8.164
5	4.983	6.084	5.030	6.452	4.019	5.755	5.483	10.297	14.048	9.225	12.371	11.884	9.974	4.876	7.174	6.529	11.932	12.963	9.016	10.196
6	5.639	6.875	5.529	5.626	5.092	2.635	6.382	9.840	13.430	9.703	11.273	11.269	9.690	5.263	5.766	5.131	11.329	12.503	9.637	9.657
7	5.102	4.647	5.459	6.490	5.125	5.579	3.368	8.966	11.786	7.781	10.584	10.282	7.361	5.399	6.884	5.308	10.361	11.567	7.576	8.359
8	10.176	9.464	8.163	8.315	8.613	10.075	9.524	2.811	9.712	7.211	9.149	9.359	6.794	9.976	8.413	10.079	10.572	10.106	7.890	7.094
9	12.909	9.919	11.837	9.522	12.085	11.476	9.900	6.902	3.498	7.392	5.040	5.953	5.310	12.363	11.299	9.402	7.094	6.435	7.572	5.297
10	9.758	7.282	8.754	8.485	8.416	9.379	8.283	6.890	8.017	2.171	7.765	6.033	5.718	8.375	9.037	7.865	6.331	6.617	3.408	4.283
11	14.152	10.208	12.901	9.646	12.607	11.407	10.944	8.294	5.151	8.712	2.839	4.769	7.651	13.214	12.518	9.881	5.964	5.999	8.184	6.676
12	14.145	10.442	13.087	10.845	12.936	12.109	11.230	8.405	4.193	7.218	4.592	3.629	6.495	13.061	12.379	10.157	4.812	5.107	6.829	5.554
13	9.670	7.838	8.265	7.917	9.050	9.197	7.758	5.913	7.183	5.424	7.349	6.795	2.940	9.120	7.739	7.537	7.712	7.516	5.701	4.096
14	5.455	7.282	5.369	6.476	5.249	6.945	6.519	8.651	11.234	6.022	10.544	9.762	7.864	4.002	6.241	6.007	9.376	9.768	6.876	6.891
15	6.143	9.446	4.304	7.147	6.123	6.932	8.218	8.462	13.048	8.912	12.239	11.830	8.501	5.814	1.946	7.233	12.207	12.001	9.769	8.735
16	6.228	7.877	6.489	5.847	6.309	4.872	6.856	11.119	12.654	8.983	11.401	10.978	9.538	4.835	5.973	3.771	10.537	11.154	9.509	8.855
17	11.771	8.647	11.394	9.153	10.666	9.911	9.428	9.933	7.734	6.613	7.043	6.029	8.007	10.284	11.323	7.914	4.155	5.868	5.803	5.980
18	12.610	10.278	11.756	9.350	11.408	11.150	10.699	9.819	6.699	6.177	6.858	5.915	8.104	10.926	11.245	8.600	4.246	2.820	6.448	5.001
19	10.024	6.714	9.225	7.865	8.849	9.272	7.530	6.894	7.137	3.946	6.885	5.560	4.957	9.001	9.672	7.485	5.371	6.518	3.006	4.016
20	8.635	7.088	7.706	6.749	7.920	8.099	7.067	7.268	7.706	4.508	7.347	6.522	4.891	7.618	7.679	5.962	5.926	6.325	4.361	3.319

Speaker Identification Score = 100.00%

Figure 13: Distance matrix for the nasal /m/ in the Live recording vs Live recording condition.

RESULTS

The aim of the study was to establish a benchmark for speaker identification in Tamil using MFFCs derived from nasal continuants. Results of the study are presented under the following headings

- Speaker identification scores for live recording
- Speaker identification scores for mobile network recording

Speaker Identification Scores for Live Recording

The Euclidean distance of the samples for the reference and test samples of each speaker were averaged separately by the software. This was then tabulated as a distance matrix comparing all the speakers (Figure 13). The one with the minimum distance from the reference was identified as test speaker. A distance matrix was computed by the software, for different combinations of test and reference speakers chosen. In this case, both the reference and test speakers were chosen from the live recordings. These are tabulated in Tables 4, 5 and 6. In this study, 30 combinations of 7 references and 3 tests (10 occurrences of each nasal for each speaker) were chosen. They were divided into 3 trials of 10 each. An average percentage correct identification was obtained for each trial, which were finally pooled to obtain the grand average. Results showed an average correct identification score of 97.6%, 85.6% and 76.5% for /m/, /n/ and /n̥/ respectively. Table 4, 5 and 6 depict the speaker identification scores obtained for all three trials for the nasals /m/, /n/ and /n̥/ respectively, along with the test sample combinations. A sample of a distance matrix for the combination highlighted (underlined), in every trial, is attached to the Appendix section. The red colour in the matrix table depicts wrong identification, and the green in the matrix table depicts correct identification. Table 4 shows the percentage of correct identification using the nasal /m/, Table 5 shows the percentage of correct identification using the nasal /n/ and Table 6 shows the percentage of correct identification using the nasal /n̥/.

Live Recording Vs Live Recording					
TRIAL 1		TRIAL 2		TRIAL 3	
Test samples	Percentage	Test samples	Percentage	Test samples	Percentage
2,3,7	**100**	3,8,10	**100**	1,3,9	**95**
2,4,10	**100**	3,7,9	**100**	3,6,7	**95**
4,5,9	**95**	7,8,9	**95**	3,7,9	**100**

5,7,8	**95**	2,5,6	**100**	2,6,9	**95**
3,9,10	**100**	7,8,9	**95**	2,6,7	**100**
2,6,8	**95**	2,3,9	**100**	3,8,9	**95**
2,3,4	**100**	3,8,9	**95**	1,8,9	**100**
7,8,9	**95**	1,2,3	**85**	3,7,10	**100**
1,8,9	**100**	1,4,10	**100**	2,3,6	**100**
3,6,10	**100**	1,8,9	**100**	3,5,7	**100**
Average = 98%		**Average =97%**		**Average = 98%**	

Table 4: Percentage of speaker identification score for nasal /m/ along with the test samples for live recording.

Live Recording Vs Live Recording					
TRIAL 1		TRIAL 2		TRIAL 3	
Test samples	Percentage	Test samples	Percentage	Test samples	Percentage
1,6,9	**85**	3,7,8	**90**	3,6,10	**80**
5,7,8	**90**	4,6,7	**85**	2,4,8	**85**
6,7,10	**95**	2,7,10	**90**	2,3,10	**85**
2,4,8	**85**	1,3,9	**95**	4,6,7	**85**
5,7,9	**80**	2,5,9	**75**	1,5,6	**90**
2,4,6	**90**	2,4,10	**95**	3,8,10	**80**
3,6,10	**80**	3,8,9	**80**	1,3,9	**95**
4,7,9	**80**	7,8,10	**80**	2,3,9	**85**
3,5,7	**100**	2,5,9	**75**	5,6,8	**95**
7,8,10	**80**	4,7,9	**80**	6,8,10	**80**
Average = 86.5%		**Average =84.5%**		**Average =86%**	

Table 5: Percentage of speaker identification score for nasal /n/ along with the test samples for live recording.

Live Recording Vs Live Recording		
TRIAL 1	TRIAL 2	TRIAL 3

Test samples	Percentage	Test samples	Percentage	Test samples	Percentage
3,5,6	**70**	7,8,10	**70**	3,6,9	**75**
1,4,6	**65**	2,4,7	**60**	4,6,7	**80**
2,5,10	**80**	1,6,10	**70**	2,3,4	**80**
2,3,10	**75**	3,4,5	**80**	5,6,8	**80**
2,4,5	**80**	2,5,6	**90**	2,5,6	**90**
1,2,9	**70**	2,4,8	**80**	4,6,9	**70**
2,8,10	**85**	2,4,5	**80**	4,7,8	**80**
3,5,7	**90**	2,3,5	**85**	2,3,7	**75**
2,8,10	**85**	2,4,6	**70**	1,2,6	**65**
1,3,8	**70**	1,3,10	**70**	1,3,7	**75**
Average =77%		**Average =75.5%**		**Average =77%**	

Table 6: Percentage of speaker identification score for the nasal /n̪/ along with the test samples for three trials for live recording.

Speaker Identification Scores for Mobile Network Recording

The Euclidean distance for the reference and test samples of each speaker were averaged separately by the software and tabulated as a distance matrix comparing all the speakers. The one with the minimum distance from the reference was identified as test speaker. Similar to the live recording vs live recording condition 30 combinations of 7 references and 3 tests (10 occurrences of each nasal for each speaker) were chosen. They were divided into 3 trials of 10 each. An average percentage correct identification was obtained for each trial, which were finally pooled to obtain the grand average. Results showed an average correct identification score of 83.5%, 65.8% and 68.3% for /m/, /n/ and /n̪/ respectively. In this case, both the reference and test speakers were chosen from mobile network recordings. Table 7, 8 and 9 depict the speaker identification scores obtained for all three trials for the nasals along with the test sample

combinations chosen. A sample of a distance matrix for the combination highlighted, in every trial, is attached to the Appendix section. The red colour in the matrix table depicts wrong identification, and the correct identifications are depicted in green. Table 7 shows the percentage of correct identification using the nasal /m/, Table 8 shows the percentage of correct identification using the nasal /n/ and Table 9 shows the percentage of correct identification using the nasal /n̥/. Overall for both the conditions (live & mobile network recording), the grand average scores were shown in table 10.

Mobile Network Vs Mobile Network					
TRIAL 1		TRIAL 2		TRIAL 3	
Test samples	Percentage	Test samples	Percentage	Percentage	Test samples
2,3,5	**90**	2,4,6	**80**	2,5,6	**75**
4,8,10	**75**	2,7,9	**95**	3,5,7	**75**
3,4,6	**80**	4,9,10	**85**	3,4,9	**95**
2,6,8	**70**	6,8,10	**85**	6,7,10	**90**
4,5,9	**85**	3,5,8	**85**	3,6,9	**95**
2,6,10	**75**	5,6,9	**70**	1,4,6	**85**
4,7,9	**95**	4,5,6	**90**	2,3,10	**85**
2,6,9	**70**	6,8,9	**85**	2,4,8	**80**
1,2,8	**85**	2,4,9	**90**	2,5,9	**85**
6,8,10	**80**	1,2,9	**85**	1,3,8	**85**
Average =80.5%		**Average =85%**		**Average =85%**	

Table 7: Percentage of speaker identification score for the nasal /m/ along with the test samples for three trials for mobile network recording.

Mobile Network Vs Mobile Network					
TRIAL 1		TRIAL 2		TRIAL 3	
Test samples	Percentage	Test samples	Percentage	Test samples	Percentage

3,5,9	**75**	3,5,9	**75**	2,7,10	**60**
2,3,8	**60**	3,5,6	**80**	1,6,7	**45**
2,8,10	**75**	4,5,6	**75**	5,7,9	**70**
3,6,9	**90**	2,5,7	**55**	4,8,9	**50**
7,8,9	**80**	5,7,8	**65**	3,8,9	**70**
2,3,6	**60**	6,7,8	**65**	2,4,9	**55**
6,7,8	**65**	2,7,9	**45**	3,5,6	**80**
6,7,9	**80**	2,5,9	**60**	4,9,10	**40**
1,6,9	**65**	1,2,9	**65**	4,6,10	**65**
1,2,9	**65**	5,8,10	**55**	2,3,10	**85**
Average =71.5%		**Average =64%**		**Average =62%**	

Table 8: Percentage of speaker identification score for the nasal /n/ along with the test samples for three trials for mobile network recording.

Mobile Network Vs Mobile Network					
TRIAL 1		TRIAL 2		TRIAL 3	
Test samples	Percentage	Test samples	Percentage	Test samples	Percentage
5,7,8	**80**	1,3,8	**80**	2,4,7	**70**
8,9,10	**45**	4,5,8	**85**	1,2,5	**65**
2,3,5	**85**	3,4,8	**55**	4,5,6	**75**
2,3,9	**75**	2,8,9	**65**	1,6,9	**70**
1,6,9	**70**	2,5,9	**70**	2,6,9	**75**
5,7,8	**80**	3,5,6	**65**	3,6,7	**75**
1,5,7	**65**	1,5,9	**60**	2,3,7	**60**
4,7,9	**50**	2,4,8	**65**	2,4,10	**70**
1,5,9	**70**	2,3,7	**60**	2,6,10	**65**
4,7,8	**60**	3,4,7	**70**	1,6,9	**70**
Average =68%		**Average =67.5%**		**Average =69.5%**	

Table 9: Percentage of speaker identification score for the nasal /n̥/ along with the test samples for three trials for mobile network recording.

Grand Average Percentage of Speaker Identification

Percentage of Speaker Identification						
	/m/		**/n/**		**/n̥/**	
	MEAN	*STD*	*MEAN*	*STD*	*MEAN*	*STD*
Live vs Live recording	97.6	3.40	85.6	6.66	76.5	7.78
Mobile network vs mobile network	83.5	7.44	65.8	12.32	68.3	9.41

Table 10: Grand average and standard deviation of the percentage of speaker identification for all three nasals across both conditions.

DISCUSSION

The present study aimed at establishing a benchmark for speaker identification using MFCCs extracted from nasal continuants of Tamil language in both live and mobile phone network recordings. Speaker identification scores ranged from 97.6% to 76.5% for live recordings and 83.5% to 68.3% for mobile network recordings. Nasal continuants perform better than vowels in speaker identification tasks. Chandrika (2010) reported that the overall accuracy using MFCCs extracted from long vowels /a:/, /i:/ and /u:/ was about 80% and the performance accuracy using vowel /i/ was 90% to 95%. Ramya (2011), in her study reported an accuracy of 93.3%, 93.3% and 96.6% for the vowels /a:/, /i:/ and /u:/ respectively. The higher percentage of speaker identification using certain vowels in the above studies, might be attributed to the fact

that the study was conducted in a controlled, laboratory environment, and the stimuli used were read out in a formal manner. However, the current study was carried out in a natural environment with some amount of ambient noise. Also, the stimuli were not formally read out, but spoken using a conversational style. On the other hand, Amino et al. (2006) compared the performance of nasal and oral sounds in speaker identification, using perceptual and acoustic analysis methods, reported greater inter-speaker distances while using nasals. Also, studies based on cepstral coefficients conducted by Amino and Osanai (2013), concluded that on an average, vowels were more efficient at identifying a speaker when compared to nasals.

Speaker identification using live recording

For the purpose of speaker identification, live recording was carried out using a digital voice recorder. The reference and the test samples in this condition were derived from the live recordings. The results indicate that the percentage of correct identification for the nasal /m/ was 97.6% and the Standard deviation was 3.4. The performance of using /m/ was better than that of /n/ and /ɳ/, whose average accuracy was 85.6% and 76.5% with a standard deviation of 6.6 and 7.7 respectively.

A higher percentage of correct identification with the nasal /m/ in the present study could be attributed to the fact that the duration of the nasal continuant /m/ was longer in the speech sample compared to the other nasals in the stimuli used for the study. This, in turn enabled selection of a more representative segment of the nasal /m/ for every speaker.

The study conducted by Ridha (2014) using MFCCs extracted from nasal continuants in Hindi, reported an identification accuracy of 100% and 90% respectively for the nasals /m/ and /n/.

Although, a similar pattern is observed in the present study, the difference between accuracy for /m/ and /n/ are larger in the current study.

In contrast, Amino and Arai (2008) concluded from their study that the coronal nasals /n/ were more useful in identifying a speaker, when compared to a bilabial nasal /m/, in Japanese. They explained that this could be due to larger intra-speaker variability encountered in a bilabial nasal. Lakshmiprasanna (2009) conducted a study on Telugu nasal continuants using formant and bandwidth measures, which showed that nasals /n/ and /n̥/were better for speaker identification compared to other nasals.

In the present study, the identification scores for the nasals /n/ and /n̥/were 85.6% and 76.5% respectively. This is significantly lower compared to the identification accuracy of the nasal /m/. Ridha (2014) reported scores of 90% and 100% for the nasals /n/ and

/ŋ/.

Also, perceptual studies and studies based on cepstral measures conducted by Amino and Arai (2009) state that coronal nasals were more reliable in identifying a speaker. A speaker identification experiment conducted on the SCOTUS corpus, using GMM models, by Yuan and Liberman (2008), reported that the velar nasal showed more inter-speaker variability compared to /m/ and /n/.

In the current study, the poorer scores on identification with nasals /n/ and /n̥/ could be due to the reduced duration of those nasal continuants available in the stimuli used. Also, the stimuli were to be read out in a casual, conversational manner, and not a strict, reading style. This might have contributed to the reduced duration of these nasal continuants, which in turn precludes good representation of the speaker using that nasal.

Also, the standard deviation is the least with the nasal /m/ and increases with /n/ and /n̪/. This could be explained by the fact that there was more variability in the speaker identification scores for the nasals /n/ and /n̪/.

Mobile network recording vs mobile network recording

Mobile network recording was done over Vodafone network using two smart phones at either end. Here, the reference and the test samples were both extracted from the mobile network recordings.

The results showed that the percentage of speaker identification for mobile network recording was significantly lower compared to live recording. The percentage of speaker identification for the nasal /m/ was 83.5% with a standard deviation of 7.4. The accuracy scores for the nasals /n/ and /n̪/ were 65.8% and 68.3% with a standard deviation of 12.3 and 9.4. The accuracy scores dropped drastically in the mobile network condition when compared to the live recording condition. The scores dropped by around 14% for /m/, 20% for /n/ and 8% for /n̪/. The scores for /n̪/were poor in both conditions, which explain the reduced difference between both conditions. The difference was most evident for the nasal /n/.

GSM (Global System for Mobile Communications) is the pan-European cellular mobile standard. Speech coding algorithms that are part of GSM compress speech signal before transmission, reducing the number of bits in digital representation but at the same time, maintain acceptable quality. Since this process modifies the speech signal, it can have an influence on speaker recognition performance along with perturbations introduced by the mobile cellular network (channel errors, background noise) (Barinov, Koval, Ignatov and Stolbov, 2010). During transmission of voice signals through communication channels, the signals are reproduced with

errors caused by distortions from the microphone and channel, and acoustical, electromagnetic interferences and noises affecting the transmitting signal.

These distortions change the formant's energy and position which are crucial for speaker identification. Barinov, Koval, Ignatov and Stolbov conducted a study in 2010 to examine the characteristics of speech transmitted over a mobile network. They concluded that the non-linearity of the GSM channel's frequency response in the range 750-2000 Hz might cause a change in the energy distribution and affect 2nd and 3rd formants (F2 and F3). They also reported a fall-off in the channel's frequency response at 3500 Hz which led to the shifting of the fourth formant (F4). Nasal murmur is typically present below 400 Hz. This information might have been lost due to the transmission characteristics of the mobile network. This could have led to poorer scores in the mobile network condition in comparison with live recording.

Ridha (2014) reported similar results when mobile network recording was compared with mobile network recording i.e., the scores dropped drastically by about 50% for /m/, 10% for /n/ and 10% for /ŋ/. Zakia Ridha (2014) reported scores of 50%, 80% and 90% for the nasals /m/, /n/ and /ŋ/. This could be due to the loss of information over the network frequency bandwidth (900/1800 in Vodafone). This limitation might have masked the characteristics of nasals useful in identifying a speaker.

The percentage of correct identification for the nasals /m/, /n/ and /n̥/were 83.5%, 65.8% and 68.3% respectively. As observed in the live recording, the nasal /m/ shows the highest accuracy followed by /n/ and finally /n̥/. The poorer scores obtained for /n/ and /n̥/could be explained by the reduced duration of the nasal continuants, further limited by the recording characteristics of a mobile network.

Overall, the speaker identification scores obtained in the live vs live condition was better than the scores obtained for the network recording vs network recording condition. Both the mobile network recordings and live recordings were done in a natural environment, without controlling parameters such as background noise. This may be the reason not achieving 100% in either of the conditions, for any of the nasals. Also, the current study was kind of text independent procedure. This could have contributed to reduction in scores for all nasals in both conditions. Typically the performance of a text- independent speaker verification system is poorer than a text-dependent system (Doddington,1998; Boves and den Oves, 1998).

Limitations

The study was conducted with a limited number of speakers, also did not considered the female participants. Only commonly used nasal continuants were chosen for this study, namely /m/, /n/ and /n̪/. Other nasals in Tamil can be experimented for speaker identification.

SUMMARY AND CONCLUSION

Biometrics refers to the identification of a person's identity based on his/her traits. Such traits may vary from simple factors such as height, weight, build, facial complexion, colour of the eyes, etc. to the more sophisticated factors such as finger prints, DNA etc. Identification of a person's through his/her speech is called speaker recognition. Speech as a biometric has gained popularity due to the extensive use of speech in man-machine communication.

Speaker recognition/identification is defined as any decision making process that uses the speaker dependent features of the speech signal (Hecker, 1971). Speaker identification by machines has become popular since the invention of telephone and computers. It can be classified into semi-automatic method, where human interference is required in the decision making process, and automatic method, where the entire procedure of speaker identification/verification is carried out by the computer program. Typically, a speaker verification system extracts feature vectors from the speech sample, does a pattern matching with the available set of database or references and finally classifies the speaker as the true speaker or impostor.

Several authors in the past have used feature vectors such as Linear Prediction Coeffients (Atal, 1974), Cepstral coefficients (Jakkar, 2009), Mel-Frequency Cepstral Coefficients (Plumpe, Quateri and Reynolds, 1999) etc. MFCCs have found to be the most efficient feature vectors in classifying a speaker. Glenn and Kleiner (1968) conducted an experiment on nasal continuants using automatic speaker verification methods, which yielded them a result of 93% accuracy in identifying speakers. Ridha (2014) conducted a study using MFCCs derived from Hindi nasal

continuants and achieved scores of 100%, 90% and 100% for the nasals /m/, /n/ and / ŋ/ respectively on speaker identification. The current study aimed at establishing a benchmark for speaker identification using Tamil nasal continuants in both live and mobile network recording conditions. A dearth of research in the area of speaker identification in Tamil language using nasal continuants has validated the need to conduct this study.

Twenty male participants between the age of 20 and 40 years were chosen for this study. They were native speakers of Tamil and had no history of speech, language or hearing difficulties. Ten meaningful Tamil sentences, relating to common messages in a threat call, were chosen for the study. These sentences consisted of the nasal sounds /m/ (bilabial), /n/ (alveolar) and /ɳ/ (retroflex) in the initial, medial and final positions of words in the sentences, and the sentences were derived based on the colloquial/informal spoken language. The frequency of occurrence of the nasals /m/, /n/ and /ɳ/ in the above sentences are 7, 9 and 7 times respectively. Out of these occurrences, the five best ones were selected for analysis. The subjects were asked to repeat each sentence thrice at habitual pitch, loudness and rate. They were specifically instructed not to adopt a strict reading style, instead asked to adopt a casual conversational style while reading out the sentences.

Live recording was done using an OLYMPUS LS100 digital voice recorder. Mobile network recording was done using two smartphones (NOKIA ASHA 301 and HTC Wildfire S). A call was placed to one of the participants from a smartphone (NOKIA ASHA 301). The participant received the call using another smartphone (HTC Wildfire S). As the participant read out the stimuli, it was recorded using NOKIA ASHA 301.

The recorded samples were transferred to computer memory and analyzed using SSL-Workbench for Semi-Automatic vocabulary dependent speaker recognition (Voice and Speech Systems, Bangalore, India) software. From the stimuli, the nasal portions (>30 msec) were segmented and stored. Every speaker was represented by a total of ten occurrences of each nasal /m/, /n/ and /n̪/, for each condition (live and mobile network). The analysis was performed separately for Live vs live condition and Mobile network vs mobile network. In the live vs live condition, the reference and the test sample were obtained from the live recording. For the mobile network vs mobile network condition, the reference and test samples were obtained from the mobile network recordings.

MFCCs derived from the nasal continuants were used to compute the Euclidean distance between the test and reference samples. For the present study, the feature vector chosen was MFCC with 13 coefficients. Upon choosing the feature vector, the system computes a measure of distance (Euclidean distance) and displays the summarized distance matrix for the selected test and reference sample. From the distance matrix, the total percentage of correct speaker identification score is displayed.

The results for the live recording condition showed that the percentage of correct identification for the nasal /m/ was 97.6% and the Standard deviation was 3.4. The performance of using /m/ was better than that of /n/ and /n̪/, whose average accuracy was 85.6% and 76.5% with a standard deviation of 6.6 and 7.7 respectively. A higher percentage of correct identification with the nasal /m/ in the present study could be attributed to the duration of the nasal continuant /m/ was longer compared to the other nasals in the stimuli used for the study. Hence, it would facilitate for the better identification.

In the mobile network recording condition, the scores obtained /m/ was 83.5% with a standard deviation of 7.4. The accuracy scores for the nasals /n/ and /n̪/were 65.8% and 68.3% with a standard deviation of 12.3 and 9.4. The poor scores in the mobile recording condition compared to the live recording condition, could be attributed to the transmission characteristics of the network. The current study was a text-independent study conducted in a natural environment with some amount of background noise. These factors could have contributed to further reduction in accuracy of speaker identification.

The current study shows that the nasal /m/ was reliable for speaker identification compared to /n/ and /n̪/. The benchmark for speaker identification using MFCCs in the live recording condition and mobile network recording condition are as follows

	/m/		**/n/**		**/n̪/**	
	MEAN	**STD**	**MEAN**	**STD**	**MEAN**	**STD**
Live vs Live recording	97.6%	3.40	85.6%	6.66	76.5%	7.78
Mobile network vs mobile network	83.5%	7.44	65.8%	12.32	68.3%	9.41

Table 11: Benchmark for speaker identification using Tamil nasal continuants.

REFERENCES

Amino, K., (2004). Properties of the Japanese phonemes in aural speaker identification. *Tech.Rep.IEICE,* 37, 49-54.

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.

Amino, K., & Arai, T. (2009). Speaker dependent characteristics of the nasals. *Forensic Science International*, 158(1), 21-28.

Amino, K., & Arai, T. (2009). Speaker dependent characteristics of the nasals. *Forensic Science International*, Vol.185, Issues 1-3, 21-28.

Amino, K., & Arai, T. (2009). Effect of linguistic contents on perceptual speaker identification: Comparison of familiar and unknown speaker identifications. *The Acoustical Society of Japan*, 30, 2-4.

Amino, K., & Osanai, T. (2013). Speaker Identification Using Japanese Monosyllables and Contributions of Nasal Consonants and Vowels to Identification Accuracy. Japanese Journal of Forensic Science and Technology, Vol.18, No.1, 13-21.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, Vol.55, No.6, 1304-1312.

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

Barinov, A., Koval, S., Ignatov, P. & Stolbov, M. (2010). Channel compensation for forensic speaker identification using inverse processing. *In Proceedings of Audio Engineering Society 39th International Conference.53-58.*

Boves, L. and den Os, E. (1998). Speaker recognition in telecom applications. In *Proceedings IEEE IVTTA-98,* Torino, 203-208.

Bricker, P. D., & Pruzansky, S. (1966). Speaker recognition. In N. J. Lass (Ed.), Contemporary issues in experimental phonetics. New York: Academic press, 295–326

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

Dang, J., & Honda, K. (1996). Acoustic characteristics of the human paranasal sinuses derived from transmission characteristics measurement and morphological observation. *Journal of the Acoustical Society of America*, 100, 3374-3383.

Doddington, G. (1998). Speaker recognition evaluation methodology- an overview and perspective. In *Proceedings for RLA2C Workshop on Speaker Recognition and its Commercial and Forensic Applications, Avignon, France,* 60-66.

Duncan lam Q.V. (1999). District Court of South Wales 99-11-0711. In P. Rose, 2002, (ed), Forensic Speaker Identification. Taylor and Francis, London.

Eatock, J., & Mason, J. (1994). A quantitative assesment of the relative speaker discriminating properties of phonemes. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP),* 133–136.

Engwall, O., Delvaux, V., & Metens, T. (2006). Interspeaker variation in the articulation of nasal vowels. *Proceedings of International Seminar on Speech Production,* 3-10.

Furui, S. (1994). An overview of speaker recognition technology. *Proc. ESCA Workshop on Automatic Speaker Recognition,* 1-8.

Glen, J.W., & Kleiner, N. (1968). Speaker Identification based on nasal phonation. *Journal of the 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 International Conference on Electrical and Computer Engineering, 565-568.*

Hecker, M. H. L. (1971). Speaker recognition: An interpretive survey of the literature. *No.16, USA*, ASHA Monographs.

Higgins, A., Bahler, L., & Porter, J. (1991). Speaker verification using randomized phrase prompting. *Digital Signal Process*, 1, 89–106.

Hollien, H., Majewski, W. & Doherty, E. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics,* 10: 139-148.

Hollien, H. (1990). *The Acoustics of Crime*. New York: Plenum

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

Imperl, B., Kacic, Z., & Horvat, B. (1997). "A study of harmonic features for speaker recognition," *Speech Communication*, vol. 22, no. 4, pp. 385– 402.

Jakkar, S. S. (2009). Benchmark for speaker identification using Cepstrum. Unpublished project of Post Graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysuru.

Jyotsna, (2011). Speaker Identification using Cepstral Coefficients and Mel-Frequency Cepstral Coefficients in Malayalam Nasal Co-articulation. Unpublished project of Post Graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysuru.

Kersta, L. G. (1962). Voice Identification, Nature, 196, 1253-1257. In Nolan, 1983(ed). The Phonetic Bases of Speaker Identification. Cambridge: Cambridge University Press.

Künzel, H. J. (1987). Spechererkennung. Grundzüge forensicher sprachverarbeitung. Heidelberg: Kriminalistik Verlag.

Künzel, H. J. (1995). On the problem of speaker identification by victims and witnesses, *Jounal of Forensic Linguistics*, 1(1), 45-57.

Künzel, H. J. (2001) ‘Beware of the ‘telephone effect’: the influence of telephone transmission on the measurement of formant frequencies’, *Forensic Linguistics*, 8(1): 80–99.

Labov, w., & Harris, W.A. (1994). Addressing social issues through linguistic evidence, in Gibbons (ed.) (1994), 287-302.

Lakshmi, P. & Savithri, S.R. (2009). Benchmark for speaker identification using vector F1 and F2. *Proceedings of the International Symposium, Frontiers of Research on Speech & Music, FRSM-2009,* 38-41.

Lekshmi Devi. (2012). Benchmark for speaker identification using Cepstral Coefficients and Mel-Frequency Cepstral Coefficients in Malayalam nasal continuants. Unpublished project of Post Graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysuru.

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

Matsui, T., Pollack, I., & Furui, S. (1993). Perception of voice individuality using syllables in continuous speech, Proceedings of Autumn meet. *The Acoustical Society of Japan*, 379-380.

McClelland, E. (2000). ‘Familial similarity in voices’, paper presented at the BAAP Colloquium, University of Glasgow.

Mc.Gehee, F. (1937). The reliability of the identification of the human voice. *Journal of General Psychology,* 31, 53-65.

Medha, S. (2010). Benchmark for speaker identification by cepstral measurement using text-independent data. Unpublished project of Post Graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysuru.

Miyajima, C. (2001). A new approach to designing a feature extractor in speaker identification on discrimination feature extraction. *Journal of Speech Communication. 35, (3-4), 203-218.*

Naik, J. M., & Doddington, G. R. (1987). Speaker verification over long distance telephone lines. *ICASSP*, 5243-527.

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

Nakagawa, S., & Sakai, T. (1979). "Feature analysis of Japanese phonetic spectra and considerations on speech recognition and speaker identification". *The Acoustic Society of Japan*, 35, 111-117.

Nakasone, H., & Beck, S. D. (2001). Forensic automatic speaker recognition, *Proc. 2001 Speaker Odyssey Speaker Recognition Workshop*: 1-6.

Nolan, F. (1983). *The phonetic bases of speaker recognition*, Cambridge, Cambridge University press.

Nolan, F. (1997). Speaker recognition and forensic phonetics, in Hardcastle and laver (eds) A Handbook of Phonetic Science. Oxford: Blackwell.: 744-67.

Noll, A. M. (1964). Short-time spectrum and cepstrum techniques for voiced pitch detection. *The Journal of the Acoustical Society of America*, 36, 296-302.

Pamela, S. (2002). Reliability of voice prints. Unpublished dissertation of All India Institute of Speech and Hearing submitted to University of Mysore, Mysuru.

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

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

Pruthi, T. & Espy-Wilson, C.Y. (2007). Acoustic parameters for the automatic detection of vowel nasalization, *in Proc. of Interspeech*, 1925-1928.

Rajaram, S. (1972). *Tamil Phonetic reader*. Central Institute of Indian Linguistics, Mysore.

Ranganathan, M. (2003). Speaker identification in disguised speech. Unpublished dissertation of All India Institute of Speech and Hearing submitted to University of Mysore, Mysuru.

Ramya, B. M. (2011). Benchmark 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, Mysuru.

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

Ridha, Z. A. (2014). Benchmark for speaker identification using nasal continuants in Hindi in direct and mobile network recording. Unpublished dissertation of Master of Science in Speech Language Pathology submitted to University of Mysore, Mysuru.

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

Rothman, H. B. (1977). A Perceptual (Aural) and Spectrographic Identification of Talkers with Similar Sounding Voices, *Proc. Intern. Conf. Crime Countermeasures*, Oxford, 37-42.

Sambur, M. R. (1975). Selection of acoustic features for speaker identification. *Proc. IEEE, Trans. Acoustic Speech Signal Process*., 23, 176-182.

Sreevidya, M.S (2010). Speaker Identification using Cepstrum in Kannada Language. Unpublished project of Post Graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysuru.

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

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

Tiwari, R., Mehra, A., Kumanat, M., Ranjan, R., Pandey, B., Ranjan, S., and Shukla, A. (2010). Expert system for speaker identification using lip feature with PCA. *Intelligent Systems and Applications (ISA), 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, 2030-2043.

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

https://en.wikipedia.org/wiki/Tamil_language

Yuan, J., & Liberman, M. (2008). "Speaker identification on the SCOTUS corpus," *Proceedings of Acoustics 2008*, 5687-5690.