EFFECT OF NOISE REDUCTION TECHNIQUE ON SPEAKER IDENTIFICATION

Project under AIISH Research Fund [ARF]
2015-16

Sanction No.: SH/SLS/ARF-34/2015-16

Total grants: Rs. 4,33,000-00

Project duration: 12 months

Project Investigators

Dr. Hema.N
Principal Investigator, ARF – 2
Lecturer in Speech Sciences
Department of Speech-Language Sciences

All India Institute of Speech and Hearing, Mysore.

Acknowledgements

This project report is the outcome of the study conducted with financial support from AIISH Research fund. The Principal Investigator and the Research Officers extend their gratitude to the Director, All India Institute of Speech and Hearing, Mysore for extending the necessary support for the conduct of the project. Our thanks are also due to all participants for their timely cooperation.

Dr. Hema.N
Principal Investigator, ARF – 34
Lecturer in Speech Sciences
Department of Speech-Language Sciences
All India Institute of Speech and Hearing
Manasagangothri, Mysore-570006

TABLE OF CONTENTS

Sl No.	CHAPTERS	Page No.
1.	Introduction	1-6
2.	Review of literature	7-25
3.	Method	26-45
4.	Results	46-65
5.	Discussion	66-79
6.	Summary and Conclusions	80-86
	References	87-93
	Appendix	

LIST OF TABLES

Table No	*TITLE*	*Page no.*
3.1	Demographic details of the participants	26-27
4.1	Speaker identification of vowels in lab condition	48
4.2	Speaker identification of vowels in traffic condition (BNR)	51
4.3	Speaker identification of vowels in Traffic condition (ANR)	54
4.4	Speaker identification of vowels in Lab condition v/s Traffic (BNR) condition	57
4.5	Speaker identification of vowels in lab condition v/s Traffic (ANR) condition	60
4.6	Speaker identification of vowels in Traffic (BNR) condition v/s Traffic (ANR) condition	63
4.7	Mean and standard deviation (SD) of the percent correct speaker identification for condition I, II, III, IV, V & VI	64
4.8	Difference value between the lower and upper bound calculated for 95% confidence interval of mean	65

LIST OF FIGURES

Figure No	*TITLE*	*Page no.*
2.1	A recorded conversation between two people in a noisy street	**9**
2.2	Power-line buzz masking the conversation between two people	**10**
2.3	A Tapped phone conversation interfered by another line's beeping	**10**
2.4	A basic system for the production of voiced speech sounds.	**15**
2.5	Schematic representation of extraction of Cepstrum	**16**
2.6	Block diagram of the MFCC processor	**17**
2.7	Example of Mel-spaced filterbank	**18**
2.8	Depicting 2-dimensional Vector Quantization	**20**
2.9	Illustration of the recognition process	**20**
3.1	The main window in Sound Cleaner software to load typical schemes	**30**
3.2	Selection of 'Street Noise Scheme' amongst the choice of other built in schemes	**31**
3.3	Window opened to load sound file & apply sound reduction technique	**31**
3.4	Sound file selected from the existing destination before the application of sound reduction technique	**32**
3.5	Output file created before the initiation of sound reduction technique.	**32**
3.6	Signal during the sound reduction processes	**33**
3.7	Window after completion of sound reduction processes	**33**
3.8	Segmentation of samples from the vowel /a:/, /i:/ and /u:/	**35**
3.9	Notepad file created for a pilot study.	**36**
3.10	SSL Workbench window for analysis.	**37**
3.11	Illustration of speaker number being selected for segmentation	**37**
3.12	Illustration of selecting the session number and occurrence number	**38**
3.13	Depiction of segmentation window showing one occurrence of /a/ for a speaker	**38**
3.14	Segmentation window using spectrogram for one occurrence of /a/ for a speaker	**39**

Figure No	*TITLE*	*Page no.*
3.15	Illustrating saves the segmentation window and .dbs file.	**40**
3.16	Testing window of SSL Workbench	**41**
3.17	Mel frequency filter bank without normalization.	**42**
3.18	Mel frequency filter bank with normalization	**42**
3.19 (a) (b)	Analysis window of SSL Workbench showing diagonal matrix and speaker identification score.	**43**
3.20	Results for pilot study depicted in .text file.	**44**
4.1	95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for lab condition	**49**
4.2	95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for traffic condition (BNR)	**52**
4.3	Percent correct speaker identification score for vowels of lab verse traffic condition (BNR)	**52**
4.4	95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for traffic condition (ANR)	**55**
4.5	95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for lab verses traffic condition (BNR)	**58**
4.6	95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for lab verses traffic condition (ANR)	**61**
4.7	Percent correct speaker identification score for vowels of lab verse traffic condition	**61**
4.8	95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for traffic (BNR) verses traffic condition (ANR)	**64**
5.1	Average spectra of the original 100 Hz sine wave (Black line) and after clipping (blue line)	**73**
5.2	Instantaneous LPC spectra of initial “o”-like sound (Black line) and after clipping (blue line)	**73**
5.3 (a)(b)	High quality speech signal without any noises and with SNR 55-60	**73**
5.4	SNR of noisy recording very low and even negative	**74**
5.5	Dynamic LPC spectrograms of a clean recording (on top) and a noisy one (at the bottom)	**74**
5.6	The reverberation time measurement	**75**
5.	Dynamic FFT spectrograms of a quiet recording (on top) and the reverberated one (at the bottom)	**75**
5.8	Average FFT spectra of the original sound signal and the same signal recorded through the devices with non-linear amplitude frequency response.	**76**
5.9	Average FFT spectrum of a speech signal sampled with 8000Hz.	**76**

Abstract

Speaking environment is always associated with one or more types of noise. For example, the forensic speech sample considered for analysis may be accompanied with some noise. Thus, for the listeners/forensic investigator the speech will not be heard clearly. Therefore background noise plays a major role in forensic speaker identification. Most of the speech recognition instrument will have difficulty in identifying speech signal when it is accompanied by background noise. Therefore to improve the intelligibility of speech signal, noise should be reduced. Hence the process of noise reduction technique plays a major role in current scenario and is available in different forms of software. From the existing software the aim of the present study was to examine the effect of noise and noise reduction technique on speaker identification using Mel-Frequency Cepstral Co-Efficients (MFCCs) on the long vowels in Kannada language. A total of 60 Kannada speaking neuro-typical adults in the age range of 20-40 years (30 males and 30 females) participated in the study. Commonly occurring Kannada meaningful sentences with long vowels /a: /, /i: /, /u: / was used for reading task. The same was recorded in two different conditions: Laboratory condition and Traffic Field condition. These recorded samples were analyzed under two phases: Before noise reduction and after noise reduction, using Sound Cleaner- Universal Noise Cancellation Software. Speech Science Lab Work bench, a Semi-Automatic vocabulary dependent speaker recognition software was used to extract Mel-Frequency Cepstral Coefficients for the truncated (PRAAT software) vowels. Results of the study revealed that in *Lab condition*, *Traffic condition*, *Traffic condition compared across traffic* condition, *Lab condition compared across traffic* and in *Lab condition compared across traffic condition*, the vowel /a:/ is found to be better followed by /i:/ and /u:/ in the average percentage of correct speaker identification of the vowels. Overall results revealed that vowel /a:/ is better for speaker identification. With reference to 95% Confidence Interval for Mean, vowel /a:/ followed by vowel /i:/ indicated the percent correct speaker identification score to be more consistent compared to vowel /u:/. The contributing factors would be the considered speech segments being vowels and their exceptional acoustical characteristics when compared to consonants, parameter MFCC, difference between any recording conditions, individual variability of speaker in relation to the speech being complex, use of noise reduction technique and the possible parameters like *overloading, signal-to-noise ratio, reverberation, nonlinearity of frequency response, sampling frequency and bit rate* which might influence instrumental identification analysis. However, the above mentioned factors and the use of 'sound cleaner' has a significant effect on percent speaker identification by reducing the influence of noise without majorly affecting the acoustical parameters of certain vowels considered for the present study.

Key Words: Sound cleaner, Semi-automatic, distortion, truncate

CHAPTER 1

INTRODUCTION

The telephone conversation has increased in recent years. Due to the increased usage of mobile phones for conversational purposes, the crime rate is increasing drastically by misusing the same for many crime-related activities like bomb threats, ransom demand, sexual abuse, and hoax emergency call. The different criminal offenses, such as making genuine or hoax emergency service calls to the police, fire brigade, or ambulance, harassing telephone calls or making threatening, extortion demands or blackmail, taking part in criminal conspiracies such as those involved in conspiring to traffic in people or trafficking or manufacture of illegal drugs or importation or, arms, cultural artifacts, and currency speaker identification may be supportive. In civil cases or for the media speaker identification may also be required. These cases include calls to local or other government authorities, radio stations, rallies or meetings, insurance companies, or recorded conversations. Among the biometric identifiers such as speech or handwriting, verification of individuals' identity based on the voice has significant advantages and practical utilizations because speech is a product of an underlying anatomical source, namely, the vocal tract and a result of natural production. Thus, comprising inherent constrained biometric feature where it does not require a specialized input device, therefore the user acceptance of the system would be high.

Voice is one of the mediums through which humans communicate with the outside world. The human voice is a carrier of personality and identity. In history, all over the globe, great personalities were identified and dominated through their invisible strength called the voice. We can also recognize our family members, media personalities, friends, and enemies through voice. As, how no two faces are similar, neither two voices are. Every mature voice has unquestionably a unique character dependent upon the structure of the head, neck, and face of the individual. The speech signal conveys several types of information. For example, speech signal conveys linguistic information (language and message) and speaker information (physiological characteristics, emotional and regional status). With reference to speaker information, different individuals sound different concerning their voice, which is a known fact. This can be illustrated with an example of how an individual is identified through his voice in any telephone conversation. This is due to the property of individuals' speech being speaker-specific. The same principle is considered in one type of speaker identification method. The method in which a person is recognized exclusively (perceptually) from his voice and is known as speaker recognition is known for long period (Atal, 1972).

To improve flexibility in the performance of speaker recognition, recent advancements have developed new tools in speech technologies. While using iris or fingerprint identification techniques there are alternative methods and only some degrees of freedom. However, in speaker recognition speech offers much more flexibility and also different levels to perform. For example, users speaking in a particular manner by force and

enter each attempt differently by using the system. Along with this, the user can also use codes or semantical/dialectical traits which are complex to counterfeit. Thus, apart from speaker identification, these methods can also be employed in forensic scenarios.

The most natural and common way used to communicate information by humans is through speech. As mentioned earlier the speech signal conveys several types of information like linguistic and speaker-dependent information. The voices of different individuals do not sound alike and this is the fact which is known to many individuals. Like how a friend over a telephone is recognized, it is due to the important property of speech of being speaker-dependent. The ability to recognize a person exclusively from his voice (perceptually) is known as speaker recognition.

Speaker recognition is the process of automatically recognizing the speaker based on the information included in the speakers' voice. "The voice is the very emblem of the speaker, indelibly woven into the fabric of speech. In this sense, each of our utterances of spoken languages carries not only its message, but through accent, tone of voice, and habitual voice quality it is at the same time an audible declaration of our membership of particular social regional groups, of our individual physical and psychological identity, and our momentary mood" (Lavner, 1994). Thus, the above-mentioned cues of spoken utterance can be used in any forensic speaker identification task. Here the main goal is to identify the speaker by characterization, extraction, and recognition of the speaker-specific information included in the speech signal according to Hecker (1971).

Apart from the text-independent and text-dependent speaker recognition system (Hollien, 2002), a key problem in attempting to characterize a speaker is that each individual's voice can vary greatly. Our voice will be changed according to with whom we are talking to, the emotion we wish to express, how formal or informal the situation is and whether there is poor quality recordings, background noise, vocal disguise, different text, various language, non-contemporary recording and also electronic scrambling like Text to Speech Converter, Voice synthesizers and the Voice Over Internet Protocol (VOIP) with nearly unlimited potential applications of speech processing in modern communication systems and networking. Speaker's voice also changes if they are, drunk, tired, or have a cold or sore throat and speakers can disguise their voices. Hence a voice is very complex to capture than a fingerprint, which is an unchanging, fixed feature of a person.

Therefore several factors affect speaker identification task as follows: The uniqueness involves an open set of trails in the identification task. From within a large to very large population of 'possibilities' the unknown must be detected. But this could be overcome to some extent so that we can reduce the number of possibilities by taking into consideration, the gender, dialect, language, some common phrases used, and style of speaking by the speaker. Conversely, it is difficult to identify the speaker by his/her voice, in particular when there is channel distortions (individuals speaking in an environment which masks or distorts their utterances) or speech distortions.

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

In certain conditions speaker themselves be the source of many types of distortions which are termed as Speaker distortion. When the perpetrator is speaking during the commission of crime fear, anxiety or stress can occur. They often will degrade identification as the speech shifts triggered by these emotions can markedly change one or more of the parameters within the speech signal. The effects of ingested drugs or alcohol; and even a temporary health state such as a cold can affect the speech. The suspect may sometimes attempt to disguise their voice. All these affect the speaker identification process dreadfully. Voice variations can be due to background noise, extreme emotions, different transmission channels, illnesses, etc and this will degrade to identify the normal voices correctly. Along with these aspects if the voice is disguised intentionally then identification would become harder and sometimes impossible. Hence there is a need to study the effect of these voice variations on forensic speaker identification.

The speaker-specific information is generally a result of the excitation source of the human vocal system. The excitation is produced by the airflow from the lungs, which thereby passes through the trachea and then through the vocal folds. The excitation is categorized as phonation, frication, whispering, vibration, compression, or a combination of these. The acoustic features pertaining to frequency and intensity are studied in different voice recognition methods on consideration of vowels, nasals, and fricatives (in decreasing order) sound in common. This is because they are comparatively easy to identify in speech signals and their spectra contain features that reliably distinguish speakers. The present study is focused on the category of long vowels (/a:/, /i:/, /u:/) of the Kannada script. Kannada is a Dravidian language spoken mainly by people in South India in the state of Karnataka (40 million native speakers). According to the study conducted by Sreedevi (2012), the mean percentage and standard deviation of frequency of long vowels /a:/, /i:/, /u:/ are 5.7 % (0.44), 1.9 %, (0.21) and 0.55 % (0.08) respectively in Mysuru dialect of conversational Kannada. Vowels are the speech sounds produced by the open vocal tract and all vowels are voiced in nature. During the production of a vowel, the vocal tract generally sustains a relatively steady shape and provides minimal obstruction of the airflow. The energy created can be emitted

through the nasal or mouth cavity without stoppage or audible friction. Vowels are described in terms of the tongue in the oral cavity (front, central, and back), the relative height of the tongue (high, mid, and low), the relative position of the lips (spread, rounded and unrounded), the position of the soft palate (closed and open), the phonemic length of the vowel (short and long), the tenseness of the articulator (lax and tense), and the relative pitch of the vowel (high, mid and low). Acoustically vowels are differentiated by fundamental frequency, duration, spectrum, and the important, formant pattern.

However, based on these acoustic features of vowels identifying a speaker from his speech signal consisting of vowels is difficult since speech is a complex and confounding one that includes many aspects, levels, and parameters to be considered during analysis (Bolt et al, 1979; Nolan, 1997). There have been various studies on the choice of acoustic features in speech recognition tasks. The present study is concerned with semi-automatic or fully automatic manner (objective) of speaker identification where machines can be used (Hecker, 1971). In Semi-automatic Speaker Identification (SAUSI), the known and the unknown samples from the speaker are selected by the examiner and are processed by the computer program for exact parameters such as first and second formants (Stevens, 1971; Atal, 1972; Nolan, 1983; Hollien, 1990; Kuwabara & Sagisaka, 1995; Lakshmi & Savithri, 2009), higher formants (Wolf, 1972), fundamental frequency (Atkinson, 1976), fundamental frequency contours (Atal, 1972), Linear prediction coefficients (Markel & Davis, 1979; Soong, Rosenberg, Rabiner & Juang, 1985), Cepstral coefficients and Mel Frequency Cepstral coefficients (Atal, 1974; Fakotakis, Anastasios & Kokkinakis, 1993; Reyond & Rose, 1995; Rabiner & Juang, 1993), Long term average spectrum (Kiukaanniemi, Siponen & Matilla, 1982) and interpretations are made by the examiner.

In the fully automatic method of speaker identification, the majority of the work is done by the computer, and the examiners' role is minimal. For automatic identification, specially designed algorithms are used which differ based on phonetic context. This method is used very often in forensic science and can be easily affected by factors such as noise and distortions, the present study is also planned to study these factors affecting speaker identification. The above-mentioned methods have their advantages and disadvantages and studies have shown varying efficiencies (Thompson, 1985). However, the Cepstral Coefficients and the Mel Frequency Cepstral Coefficients are more effective in speaker identification compared to other features. Hence, the present study is focused on the usefulness of Mel frequency cepstral coefficients (MFCC) on speaker recognition.

With reference to the different methods of speaker identification, the variables affecting speaker recognition in the different contexts of the conversational speech sample would be the background noise. Since the speaking environment is always associated with one or more types of noise, the considered speech sample may be accompanied by some noise. Thus, for the listeners, the speech will not be heard clearly. Thus, background noise plays a major role in forensic speaker identification. Most of the speech recognition instrument will have difficulty in identifying speech signal when it is accompanied by

background noise. To overcome this problem, the noise has to be filtered so that the required speech signals will be free from noise and the same will be used for further analysis.

When the speaker is talking in the environment, most of the time, speech is not heard clearly by the listener due to the surrounded noise. Background noise also plays a major role in forensic speaker identification. Most of the software will have difficulty in identifying speech signal when it is accompanied by background noise. To overcome this problem, the noise has to be filtered so that the wanted speech signals will be heard clearly. Hence there are various researches conducted to reduce the background noise during forensic speaker identification. In the current scenario various software and hardware products consist of noise reduction technology which reduces the noises and compensates for distortions, thus facilitates improving the intelligibility of speech signal. Over 20 years, SpeechPro Inc. as a global leader in speech technologies has been advancing specialized tools for text transcription and efficient noise reduction of low-quality recordings. Numerous research on the perception of poor audio recordings and noisy speech signals performed by SpeechPro have resulted in the creation of the unique sound filtering algorithms that are now presented in the software and hardware products like Sound Cleaner, ANF II, and The Denoiser Box. In the present study Sound, Cleaner Signal Enhancement Program Model 5142 (Noise Cancellation Software) is used in reducing the background noise and also to see its effect after the noise reduction method.

Among the different software and hardware products, the sound cleaner is a distinctive software solution for filtering noise-corrupted sound recordings and enhancement of speech intelligibility. The signal is processed by a series of modules. There are 14 typical schemes of processing existing which can also be tailored to filter noise. Each module is characterized by a separate window with several basic and professional controls. While listening to the output of the signal; the options in the modules can be adjustable based on the user's requirements. It consists of 19 different processing modules including equalizer, adaptive frequency compensation filter, adaptive broadband noise filter, adaptive inverse filter, adaptive stereo filters in time and frequency domain, dynamic processing automatic gain control, and so on. For tonal noise, the signal-to-noise ratio is improved by 50dB and for broadband noise up to 20dB. It is efficient for both restoration and enhancement of poor quality recordings, real-time sound loaded from external sources, and also recordings made in noisy environments.

Several studies have been conducted where various traditional techniques have been used for noise removal or noise compensation. It is evident from the review that MFCCs are, perhaps, the best parameter for speaker identification and less susceptible to variation of the speaker's voice and surrounding environment (noise). Also, the vowels may be the most suitable, among speech sounds, for speaker identification. However, to date, there are limited studies on vowels as strong phonemes for speaker identification using semi-automatic methods in the presence and absence of noisy situations (conditions). Scientific testimony impresses any court of law in whichever country that might be. For any result to be called

scientific, it has to be measured, quantified, and reproducible if and when the need arises. Therefore, a method to carry out these analyses becomes a must. In this context, the present study was aimed to investigate the effect of noise and noise reduction techniques on speaker identification using MFCCs on the long vowels in the Kannada language. The objectives of the study were to 1) evaluate the percent correct Speaker Identification using MFCCs on the long vowels in the Kannada language for lab recording separately and field recording separately (with noise) before the application of noise reduction technique conditions. 2) To evaluate the percent correct Speaker Identification using MFCCs on the long vowels in the Kannada language for field recording separately after the application of noise reduction technique (without noise). 3) To compare speaker identification using MFCCs on long vowels in the Kannada language in lab recording conditions versus field recording (with noise) before the application of noise reduction technique. 4) To compare the percent correct Speaker Identification using MFCCs on the long vowels in the Kannada language for lab recording versus field recording after the application of noise reduction technique (without noise). 5) To compare the percent correct Speaker Identification using MFCCs on the vowels in the Kannada language for field recording (with noise) before the application of noise reduction technique versus field recording after the application of noise reduction technique (without noise).

CHAPTER II

REVIEW OF LITERATURE

2.2 Factors contributing to speaker recognition

The important concern in speaker recognition is to carry out a speaker identification task with the variable called (A) *Uniqueness* and *distortion*. In *uniqueness*, the speaker identification task might involve an open set of trails, where the unknown must be detected from a large to a very large population of 'possibilities'. But this can be overcome to some extent that we can reduce the number of possibilities by taking into consideration such as gender, dialect, language, some common phrases used, and style of speaking by the speaker as unique concern. The other is *distortion*, speaker identification task becomes difficult when the speaker is talking in such an environment where there is more distortion or masking present (channel distortion) or when they are excited or stressed (speech distortions). Thus, distortions are broadly classified into two types: (i) Speaker Distortion and (ii) System Distortion.

In Speaker Distortion, here the distortions are due to the speaker himself. During the commission of a crime, the perpetrator can exhibit fear, anxiety, or stress like emotions. These will degrade identification, as the speech is directly triggered by these emotions which consequently alter one or more parameters of the speech signal. Factors such as temporary health conditions like the common cold; or intake of alcohol or drugs; or disguise of voice can affect speaker identification in a troublesome way.

On the other hand, System Distortion is due to several kinds of signal degradation. Some of the limitations in the system related issues are; (a) reduced frequency response through a telephone line or mobile phone, (b) poor quality tape recorders, and (c) reduced dynamic range and/or frequency response of microphones. In such cases, important information about the speaker is lost and these elements are not usually retrievable. Due to this, essential speaker-specific acoustic parameters can be reduced. In addition to this, noise can cause a particularly debilitating type of system distortion as it tends to make the talker's voice and, therefore can obscure elements needed for identification. Examples of possible noise are wind, motors, fans, automobile movement, and clothing friction. The noise itself may be intermittent or steady-state saw tooth or thermal and so on. Another type of noise is the frequency or harmonic distortions which make the task of speaker identification more complex. Examples include intermittent short circuits, variable frequency response and harmonic distortion, and so on.

The issues related to speaker recognition are prevailed by the use of (B). *Noise reduction techniques with the software/hardware-based technology*. When there is clean speech, the speech recognizers gives acceptable recognition accuracy. Whereas in actual circumstances mainly in noisy environments, the performance level of speech recognizers degrades because mismatch takes place between training (reference) and operating (recognition) environments (Gong, 1995). Das et al., in 1993 found a 1% error rate when the

system trained under quiet conditions and the error rate increased to more than 50% in a cafeteria environment. While processing the speech signal, due to various sources of interference or distortions, the quality of speech will be at risk. Hence it is necessary to adapt techniques of noise cancellation and speech quality improvement while designing the system for speech signal processing. Sounds produced in the vocal tract have an efficient frequency range of 300 Hz to 3000 Hz which is sufficient for understanding speech, though speech has a wider frequency range. For example, all noises below 300 Hz and above 3400 Hz may be suppressed by filtering out the speech signal through the filter band. When filtering takes place for narrowband signals, the understandability of speech won't be affected. But in most of the situations, for example, any engine noise, music, environmental noise, etc where the noise occurs in a wide-frequency band with random distribution. In these situations, the noise is much difficult to segregate and suppress from the signal, because it falls in a similar frequency range as speech. Hence environmental noise has turned as one of the foremost obstacles to commercial use of speech recognition techniques.

The intelligibility of a speech signal will be improved by reducing the noise component. Various software and hardware products have incorporated noise reduction technology where noises are reduced and compensated for distortions. Meanwhile, however, the global leader in Speech Technologies Center is a leading developer of voice and multimodal biometric systems, as well as the solutions for audio and video recording, processing, and analysis. For over 20 years, the SpeechPro under STC has been developing specialized tools for efficient noise reduction and text transcription of low-quality recordings. Various studies on the perception of poor audio recordings and noisy speech signals carried out by SpeechPro have resulted in the formation of the unique sound filtering algorithms that are now presented in the software and hardware products like Sound Cleaner, ANF II, and The Denoiser Box. In the present study, the Sound Cleaner Signal Enhancement Program Model 5142 (Noise Cancellation Software) was used to reduce the background noise and an attempt has been made to see its effect on speaker identification score for the samples which was subjected to noise reduction.

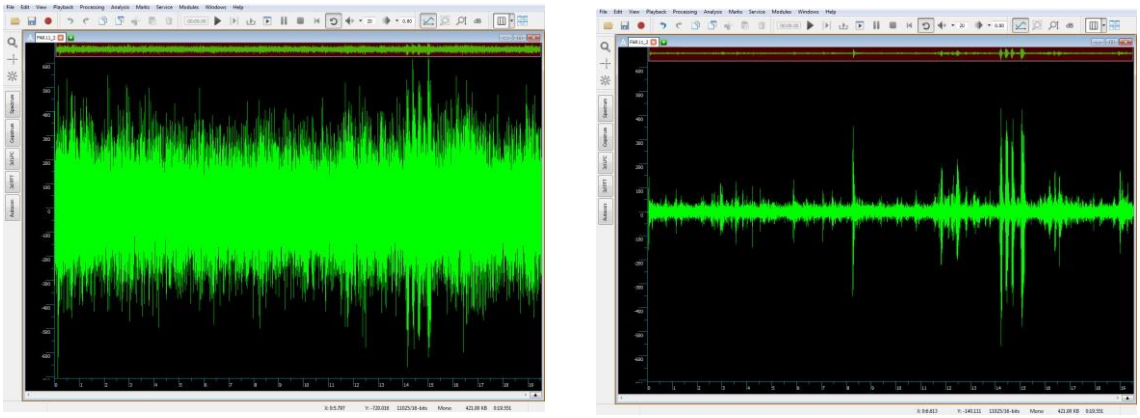
Audio hindrances consist of two main categories: noises and distortions. In the recordings of original human speech, if the recording is considered as a useful signal then the entire extra information which reduces the quality of this functional signal is considered as **noises.** The entire thing which alters the original useful signal is considered as **distortions.** The echo effects and reverberation are the typical distortions at the acoustical level. When the speech signal in the form of an acoustic signal is converted into an electrical signal and if it undergoes numerous technical limitations, the distortions will also emerge.

The noises are characterized based on *time* and *frequency* domains. With reference to the time domain, the noises are characterized as continuous, discontinuous, and pulse-like. *Continuous* is the gradually changing noises resembling the sound of street noise, the sound of the wind, office, industrial equipment, a bad phone line or the hiss of an old record and traffic noise. The *discontinuous* is the repetitive especially tonal noises like beeps, bells, or honks. The *pulse-like* is the sudden especially unharmonious and occasionally loud noises

like bangs, taps of the steps, thumps, clicks, and gunshots. With reference to frequency domains, the noises are characterized as broadband noise and narrow-band noise. The *broadband noises* appear at numerous frequencies like fizzing sounds or background hiss. The *narrowband noises* stand for a set of certain frequencies, comparatively stable tonal sine waves (sinusoid) like in sirens, power-supply, drones, whistle, equipment hindrances (chainsaws & drills), machinery engine noises, and hums.

Therefore, over time, the noise characteristics generally vary. It is essential to employ a special processing scheme that adjusts automatically to noise characteristics. To name one is the adaptive filtration algorithms. Here the digital filtration algorithms will alter to a definite type of audio hindrance. The various types of adaptive filtration algorithms listed by Andrey (2010) are:

1. **Adaptive broadband filtration**: It works on the principle of an adaptive frequency algorithm. It is developed to suppress periodic and broadband noises due to mechanic vibrations or electric pick-ups, communication channel or recording equipment interferences, room and street noise. An example is shown in Figure 2.1. a recorded conversation between two people in a noisy street.



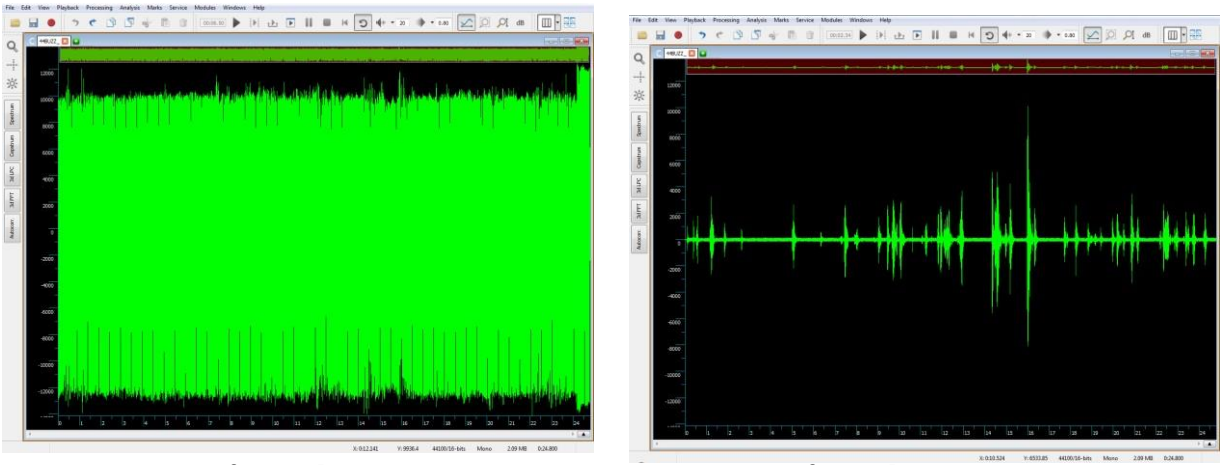
Before reduction After reduction

Figure 2.1. A recorded conversation between two people in a noisy street

2. **Adaptive inverse filtration:** In this process, the adaptive spectral correction algorithm is used. Adaptive inverse filtration efficiently suppresses strong periodic noises from mechanical vibrations or electrical pick-ups thus improve speech and balance the signal. It suppresses the stronger ones and boosts poor signal components at the same time.

3. **Frequency compensation:** The Widrow–Hoff adaptive filtering algorithm of one-channel adaptive compensation is used in this process. For narrow-band stationary interferences, this is the most successful one. The filter alters itself effortlessly preserving the high-quality of the speech. It eliminates narrowband stationary interferences as well as regular ones (vibrations, power-line pickups, electrical device noises, steady music, room,

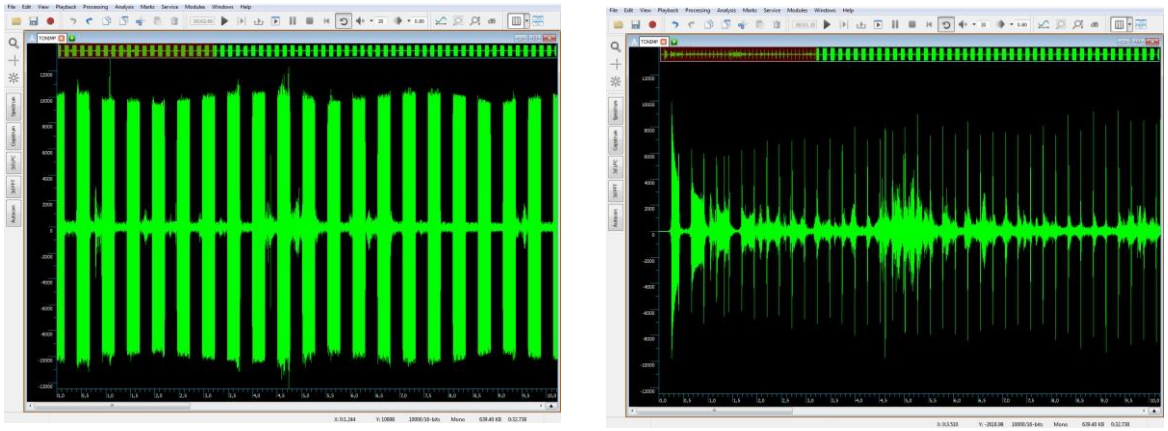
reverberation, traffic, water noises, etc.) together. It maintains the speech signal much better than other filters. An example is shown in Figure 2.2 of a Power-line buzz masking the conversation between two people.



Before reduction After reduction

Figure 2.2. Power-line buzz masking the conversation between two people

4. Adaptive impulse filter: Pulse interferences such as radio noises, clicks, gunshots, knocks, etc distorts and mask speech or musical fragments which can be automatically restored by an adaptive impulse filtering. This filtering algorithm improves the quality of the signal by suppressing dominant signal impulses and thus unmasking the necessary audio signal and increases its intelligibility. An example of a Tapped Phone conversation interfered by another line's beeping is shown in Figure 2.3.



Before reduction After reduction

Figure 2.3. A Tapped phone conversation interfered by another line's beeping

5. Dynamic signal processing: It increases the intelligibility of the speech if the signal fragments widely vary in level, in the case of resonant knocks (i.e. long impulses) and room

noises. This progress and unmask the audio signal by suppressing the dominant clicks and impulses and diminishing the listener's fatigue in case of lengthy audio recordings.

6. Stereo filtration: This is the most recent advancement in the field of noise reduction technologies. The difficulty in eliminating the noises could be determined with the assistance of dual-channel audio information monitoring and further dual-channel adaptive filtration. Here the process successfully diminishes crowd noises and background music by increasing the valuable speech signal and is ideal for recordings in big-sized rooms like restaurants, halls, theaters, etc. In general, it is important to be aware of these adaptive filtration algorithms in forensic speaker verification and there is a need to study the influence of these in speaker identification processes.

2.3 Review related to background noise

Due to the presence of background noise, speech recognition becomes very difficult. This is because the noise influence on speaker's acoustic features of their recorded signal and makes them different from those seen during testing versus training. Various approaches have been implemented to improve the noise robustness of speaker recognition. The study was done by Berouti, Schwartz, and Makhoul (1979) to enhance the speech signals. For the same, the spectral noise subtraction method was used to enhance speech which was corrupted by broadband noise. Results showed that there was no loss of intelligibility corresponding with the enhancement technique.

Techniques like Kalman filtering (Fingscheidt, Suhadi, & Stan, 2003) or spectral subtraction (Garcia & Rodriguez, 1996) can also be used to filter noise from speech, based on the prior knowledge of the noise characteristics. It is also possible to extract noise-robust features, e.g. relative spectral (RASTA) features (Hermansky & Morgan, 1994) from speech signals instead of removing the background noise.

Kalman filtering is done with reference to estimation of the time delay of arrival (TDOA) of sound signals through a pair of spatially separated microphones. Following this, the estimated TDOAs of different microphone pairs will be used in combination with the microphone array geometry to localize the sound source. But, due to the one-sample-precision of the TDOA estimation algorithm and due to noise and reverberation influences, the TDOA estimates only the real TDOA values, which are not identical and leads to relatively high variances in consecutive position estimates. This is the method to smoothen the speaker trajectory and assure the robustness of the signal (Bechler, Grimm & Kroschel, 2003).

It is also possible to ignore some parts of speech which are corrupted by background noise using the missing feature theory (Bonastre, Besacier & Fredouille, 2000). For example, consider a spectrum that has been passed through a high-pass filter. If we assume that the first eight spectral magnitude features are below the threshold and are labeled as "missing." Once each spectral magnitude feature in a frame is labeled as present or missing, a

computationally simple modification of probability models discards missing features and forms densities that would have been obtained by training without missing features.

Based on the same principle of Missing Feature theory, in some instances, the relative spectral features (Hermansky & Morgan, 1994) from speech signal might be removed instead of removing the background noise. It is also possible to ignore the parts of speech corrupted by background noise. Few approaches are used in statistical speakers' models (e.g. Gaussian Mixture Models (GMMs). Gaussian Mixture Model (GMM) is defined as a function of the parametric probability density, a weighted sum of Gaussian component densities is its representation. This is commonly used in biometric systems, like vocal-tract related spectral features in speaker recognition systems thus it has the competency of symbolizing a large class of sample distributions.

Researchers found that noise separation properties would become much easier in vowels because frequency properties of vowels are known, whereas, in the case of consonants, they have a wide frequency range which is difficult to separate them from noise using filtering techniques (Davis, 2002).

Barinov, Koval, and Ignatov (2010) conducted a study to check the effect of channel compensation for forensic speaker identification using inverse filtering. The speaker was made to call from a cell phone to a landline phone, and the speech was recorded in two different manners simultaneously. The first recording (original) was recorded using a high-quality digital recorder, and the second recording (signal from GSM channel) was recorded from a landline phone using a high-quality recording station. Inverse processing was used to compensate for the influence of the transmission channel which improves the formants representation accuracy. The results turned up positively, the signal which was corrupted by the transmission through the low-quality communication channels (GSM lines) using the inverse process was able to restore the original formants structure. Hence the study concluded that channel compensation is more transparent, convenient, and effective than cepstral mean subtraction, relative spectral (RASTA), etc.

Md Imdad, Akhtar, & Md Imran (2012) conducted a study to investigate the difficulty of speaker verification and identification in noisy conditions. Here they described a method that merges missing-feature theory and multi-condition model training to model noise with unidentified temporal-spectral characteristics. The introduction of such a technique is very useful since it removes noise and avoids the problem of recognizing the voice.

2.4 Review related to speaker identification

There are several methods of speaker identification and it is classified as i). Speaker Identification by listening, ii). Speaker Identification by visual method and iii). Speaker identification by a machine which has two subtypes a). Semi-automatic speaker identification and b). Automatic speaker identification according to Hecker (1971) and Bricker and Pruzansky (1976).

2.4.1 Speaker Identification by Listening (Subjective method)

The listening method is also known as Aural-Perceptual Speaker Identification (AP-SPID). It is one of the oldest methods used in speaker identification. In this method, the examiner will be given reference samples (unknown samples) and a test sample (known sample) aurally. The reference samples consist of a line-up of the suspect's speech (obtained from the recorded message, threat call, etc.) along with the 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 will be asked to match the test sample with one of the references.

However, several factors are affecting Aural-perceptual speaker identification with reference to (1). Listener- the 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.

In a study conducted by Mc Gehee (1937), listeners were asked to match the target voice from a set of five male voices. The procedure was repeated after a couple of days, 2 weeks, 3 months, and 5 months. Results revealed 83% of correct speaker identification after 1 day, which was sustained for a week. The percentage dropped to 68% after 2 weeks, 35% after 3 months, and 13% after 5 months. A similar trend of decline in percentage was found by Bricker and Pruzansky (1996).

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

(2). The speaker related- Here the *unique speech characteristics* that are the voices that have unique characteristics are easier to identify. Apart from this, *the disguise* that is depending on the type of disguise used by the perpetrator; his voice may or may not be easily identifiable. The last is *the stress* resulting in different emotions and *accents* which is a resultant of dialects.

According to the study done by Reich and Duke (1979), the effect of disguise on speech recognition was studied where they concluded that the most damaging disguises of various disguises are free disguise and disguise in the form of a strong nasalized speech.

(3). Speech sample related- This includes the length and quality of the sample, the environment in which samples have been recorded, and contemporary versus non-contemporary samples. Speaker identification scores dropped to 42% when the samples are non-contemporary (Rothman, 1977). Various authors like Kunzel (1995) and Pollack, Pickett, and Sumbey (1954) states that for speaker identification tasks, speech samples should be present for a minimum of 30 seconds.

2.4.2 Speaker Identification by Visual Examination of Spectrograms

The second method of speaker identification is based upon visual examination by comparison of spectrograms. During the mid-1940s, the first sound spectrograph (Sonagraph) was invented by the scientists of Bell Telephone Laboratories, U.S.A. A spectrograph is a three-dimensional representation of speech sounds where the X-axis represents time, Y-axis represents frequency and Z-axis represents intensity. In the case of speaker identification, trained experts were given the spectrograms of different utterances of word/phrase and determined whether the utterances are from the same speaker or not. Kersta (1962) published a paper on 'Voiceprint identification' in which he claimed that speaker identification using spectrograms yields an error rate less than 1% hence conclude speaker identification using spectrogram is an efficient method in speaker identification.

A large-scale study done by Tosi et al. (1972) used spectrogram matching. Results revealed 86% to 96% of correct speaker identification. They also focused on issues such as the number of cue words required for speaker recognition, the effect of recording conditions, the effect of context of cue words on speaker identification, contemporary v/s non-contemporary samples, and so on.

Reich et al., (1976) investigated the effect of vocal disguise upon spectrographic speaker identification. The speakers were made to produce two sets of the sentence in normal speaking mode, hoarse disguise, old-age disguise, hypernasal disguise, free disguise, and slow rate disguise. Results revealed certain vocal disguises markedly interfere with spectrographic speaker identification. Performance in speaker identification ranged from 14.17% (slow-rate) to 35% (free-disguise) and 56.67% when there was no disguise in the utterance.

A study done by Pamela (2002) attempted at benchmarking using spectrograms. The reliability of voiceprints was investigated by extracting the acoustic parameters in the speech samples using wideband spectrograms. A total of six Hindi-speaking males participated in the study, and the target words were 29 bi-syllabic words which consisted of 16 plosives, 5 nasals, 4 affricates, and 4 fricatives in the word medial position. Acoustic parameters such as formant transition duration, VOT, closure duration, duration of phonemes were measured. The results indicated that 67% of the measures varied across speakers and 61% of the measures varied within speakers.

Ranganathan (2003) investigated speaker identification using spectrograph in disguise speech. Results revealed no significant difference between accuracy speaker identification in normal and disguised conditions.

Arjun & Hema (2014) conducted a preliminary study on 'Speaker identification using spectrographic analysis on fricatives in Kannada speaking individuals. Acoustic parameters such as fricative duration, fricative amplitude, and center frequency of frication were measured. Results showed relatively positive results on a few specific combinations of acoustic parameters.

2.4.3 Speaker Identification by Machine Method (Objective Method)

Speaker identification by machines became popular in the 1970s. In the semi-automatic method, the examiner interprets the results provided by the system. In the automatic method, the contribution of the examiner is minimal and the system makes use of several algorithms to assume who the speaker is, or whether the speaker is actually who he claims to be. The automatic speaker recognition process involves two phases namely, the training phase and the testing phase. In the training phases, each speaker's samples are collected and stored as a database. While in the testing phase, the speaker's utterance is fed into the system. The speaker recognition system compares it with the stored database to determine the identity of the speaker or verify the speaker's identity. To arrive at a decision, the automatic speaker identification goes through the following steps. They are: i) Feature extraction; ii) Pattern matching and iii) Classification.

Features are certain acoustic parameters that characterize an individual's speech. Few desirable characteristics for the features are: i) They must be highly discriminable across speakers ii) Should vary minimally from session to session and iii) Must be difficult to impersonate.

During the 1970s speakers were compared based on certain parameter sets such as fundamental frequency, vowel, and nasal consonant spectra, global source spectrum slope, and word duration. Over the years, several feature vectors such as formant frequencies, Linear Prediction Coefficient (LPC) (Atal, 1974), Cepstral Coefficients (Jakkar, 2009; Medha, 2010 & Sreevidhya, 2010) and Mel-Frequency Cepstral Coefficients (Chandrika, 2010; Hassan, Jamil, Rabbani & Rahman, 2004; Plumpe, Quateri & Reynolds, 1999; Tiwari et al., 2010) have been employed for speaker identification.

Noll (1964) was the first to implement cepstrum (an anagram of frequency) as a tool for automatic pitch detection. That is taking the Inverse Fourier Transform (IFT) of the logarithm of the estimated spectrum of a signal result in Cepstrum. The voiced speech sounds are produced from the vocal source and vocal tract. The periodic puffs of air emitted by the vocal cords constitute the source signal s(t). The effect of the vocal tract is entirely precised by its impulse response h(t) such that the output speech signal f(t) equals the convolution of s(t) and h(t) (Figure 2.4). The output of the vocal source and vocal tract are almost independent or simply identifiable and separable. The effects of vocal source and vocal tract are segregated by the Fourier transform (decomposing the signal into sine or cosine component) of the logarithm of the power spectrum.

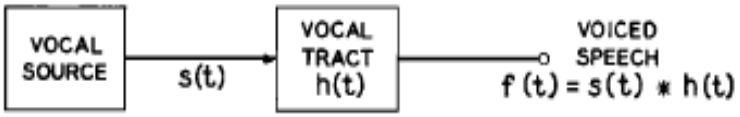


Figure 2.4: A basic system for the production of voiced speech sounds.

The effect of the vocal tract is to produce a “low-frequency” ripple in the logarithm spectrum, though the periodicity of the vocal source exhibits itself as a “high-frequency” ripple in the logarithm spectrum. Hence, the spectrum of the logarithm power spectrum has a sharp peak equivalent to the high-frequency source ripples in the logarithm spectrum and a broader peak corresponding to the low-frequency format structure in the logarithm spectrum (Figure 2.5). The peak corresponding to the source periodicity can be made more distinct by squaring the second spectrum. This function, the square of the Fourier transform of the logarithm power spectrum, is called the "cepstrum" (Noll, 1967).

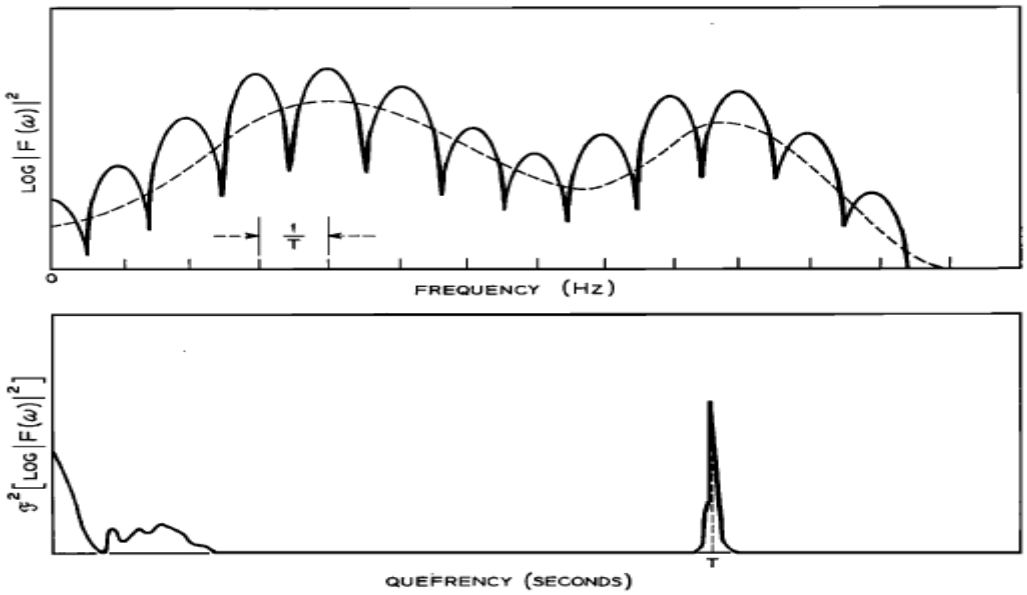


Figure 2.5: Schematic representation of the extraction of Cepstrum

Various studies were done on automatic speaker identification using cepstral measurements. Luck (1969) used cepstral measurements for speaker identification. The standard test phrase used was “My code is_____”, where several feature vectors were extracted for comparison. The verification decision was treated as a two-class problem were, the speaker being either the authorized speaker or an imposter. Authorized speaker’s samples were considered as reference data. The distance between the test samples and the reference sample was checked. Based on the nearest reference distance with that of the test sample’s distance, judgment is made. Results revealed there were 6% to 13% error rates when 4 authorized speakers and 30 imposters were examined.

Atal (1974) investigated the effectiveness of automatic speaker recognition by several parameters using a linear prediction model. A total of 10 speakers participated in the study where, the speech data considered were a total of 60 utterances, where there were 6 repetitions of the same sentence. For every 50 msec from the speech sample at 10 kHz, twelve predictor coefficients were determined. Impulse response function, the autocorrelation function, the area function, and the cepstrum function were the predictor coefficients and other speech parameters that were derived and were used as input to an automatic speaker recognition system. The identification decision was done based on the distance between the

test sample vector and the reference sample vector. The speaker matching with the smallest distance of the reference vector was identified as an unknown speaker. Whereas in verification, the speaker was verified if the distance between the test sample vector and the reference vector for the claimed speaker is less than a set threshold. Among all the parameters examined, cepstrum was seemed to be the most successful in providing identification accuracy of 70% for speech having a duration of 50 msec. it was also found that when the duration of the sample was increased to 0.5 sec, identification accuracy was 98%. Meanwhile, verification accuracy was calculated where for 50 msec duration of speech sample it was 83% and when the duration increased to 1-sec verification accuracy increased to 98%.

2.4.3.1 Mel-Frequency Cepstral Coefficients (MFCCs)

Numerous techniques are available for parametrically exhibiting the speech signal for speaker recognition tasks, such as Mel-frequency Cepstral Coefficients (MFCCs), Linear Prediction Coding (LPC), and so on. The MFCCs are provoked by studies of the human peripheral auditory system. Among them, MFCCs are most accepted and best recognized. MFCCs are derived from the known variation of the human ear's critical bandwidths with frequency (Hansen & Proakis, 2000). The two main filters used in MFCCs have linearly spaced filters and logarithmically spaced filters. To incorporate the phonetically essential characteristics of speech, MFCCs will be used in the speech signal. A series of calculations will take place which uses cepstrum with a nonlinear frequency axis following mel scale. To get mel cepstrum, the speech signal will be windowed first using the analysis window and then Discrete Fourier Transform will be computed. The main rationale behind MFCC is to mimic human ears behavior (Figure 2.6).

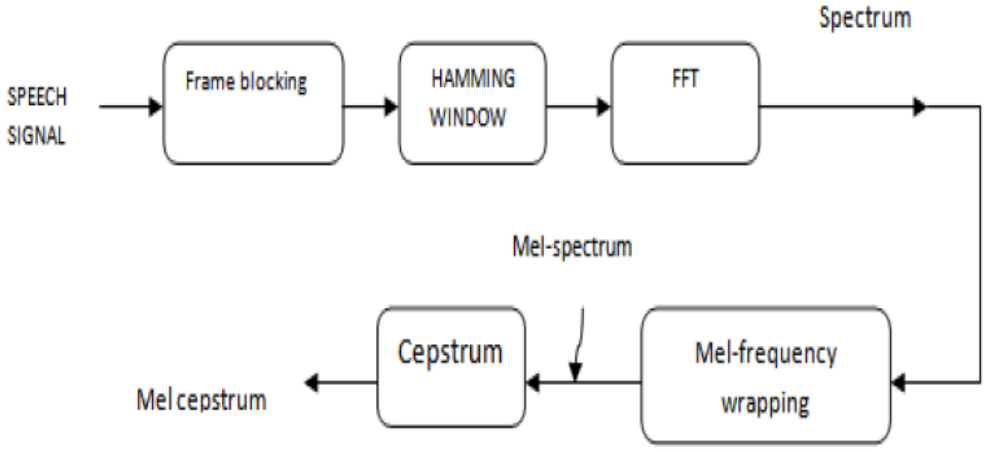


Figure 2.6: Block diagram of the MFCC processor

When a continuous speech is passed into frame blocking, it blocks the continuous speech signal into frames of N samples. The adjacent frames will be segregated by M (M<N). Therefore the first frame compresses of first N samples; meanwhile, the succeeding frame begins with M samples following the first frame and overlaps it with N-M samples and so on. To minimize the signal discontinuity at the beginning and end of each frame, windowing will

take place. This is done to minimize the spectral distortion. Then the signal will be passed into Fast Fourier Transform, which alters each frame of N samples from the time domain into frequency domain which implements Discrete Fourier Transform (Linde, Buzo & Gray, 1980). As we know that the speech signal consists of tones with different frequencies. Using the 'Mel' scale a subjective pitch is calculated, for every tone with an actual frequency. The mel-frequency scale is linear frequency spacing below 1000Hz and logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels (Seddik, Rahmouni, & Sayadi, 2004). Hence Mels for a given frequency f in Hz can be calculated using the formula "Mel(f) = 2595*log10(1+f/700)".

The filter bank is used to stimulate the subjective spectrum which is spaced evenly on the mel-scale (Figure 2.7). Filter bank has a triangular bandpass frequency response and the spacing as well as the bandwidth is decided by a constant mel frequency interval.

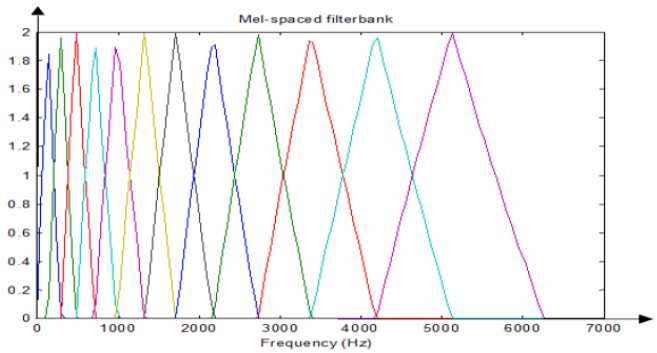


Figure 2.7: Example of Mel-spaced filterbank

In the final step, the log mel spectrum will be converted back to the time domain using the Discrete Cosine Transform (DCT) because the mel spectrum coefficients are real numbers (and so are their logarithms). Therefore the result is called the Mel frequency cepstrum coefficients (MFCCs). The MFCCs may be calculated using the following equation.

$$\tilde{C}n = \sum_{k=1}^{K} \left(\log \tilde{S}_k\right)\left[n\left(k-\tfrac{1}{2}\right)\tfrac{\pi}{K}\right]$$

where $n = 1, 2 \ldots\ldots K$

For the calculation *'K'* the coefficient length is typically chosen as 20. $\tilde{S}_k$ constitute cepstrum. A set of mel-frequency cepstrum coefficients is calculated by applying this for every speech frame. This set of coefficients is called an *acoustic vecto*r. These acoustic vectors are used to symbolize and recognize the voice characteristic of the speaker. As a result, every input utterance is altered into a sequence of acoustic vectors.

Numerous studies have been conducted on speaker identification using MFCCs. Hasan, Jamil, Rabbani, and Rahman (2004) have used Mel-Frequency Cepstral Coefficients for feature extraction and vector quantization in a security system based on speaker

identification. A total of 21 speakers participated in the study. During framing in linear frequency scale different types of windows were used such as triangular, rectangular, and hamming windows. The hamming window yielded a better result when compared to the triangular and rectangular windows. Hamming window is the sum of rectangle and hanning window and it is amplitude weighting of the time signal which is used with gated continuous signals which give a slow onset and cut-off in turn to decrease the ability to generate side lobes in their frequency spectrum. This window has similar properties to the Hanning window with the supplementary feature which suppresses the first sidelobe, which gives the best results for a large signal. The study revealed that when the codebook size is 1 speaker identification score was 57.14% as the codebook size increased to 16, the speaker identification increased to 100%. Hence it was concluded that the combination of Mel-Frequency and Hamming windows gives the best results.

Mao, Cao, Murat, and Tong (2006) used linear predictive coding (LPC) parameter and Mel Frequency Cepstrum Coefficient (MFCC) for speaker identification. The text-dependent recognition rate of 50 speakers improved from 42% to 80% and the text-independent recognition rate of 50 speakers improved from 60% to 72%.

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 Story databases.

According to the study conducted by Wang, Ohtsuka, and Nakagawa (2009), integrated new feature called phase information in MFCCs on speaker identification task. The speech database consists of normal, fast, and slow speaking modes. 35 Japanese speakers participated in the study NTT database was used in the study. NTT database consists of sentences uttered by the speakers (on five sessions over ten months). Results revealed that in all speaking there was robustly seen in phase information than the original information. Using the phase information, the speaker identification error rate was reduced by 78% for clean speech. Also, the error rate reduced remarkably when new phase information was integrated with MFCCs by 20%~70% in comparison with using only MFCC in a noisy environment. The study also evaluated speaker verification experiments using phase information and found very effective results.

A study was conducted by Singh and Rajan (2011) to evaluate the accuracy affecting factors of MFCCs and Vector Quantization based speaker recognition system. The results revealed background noise was the most dominating factor which degrades the accuracy of the speaker recognition system whereas; speech-related factors and sample length were less critical. However, Gill, Kaur, and Kaur (2010) has used Vector Quantization (VQ) successfully in speaker identification task. This process involves the extraction of a small number of representative feature vectors that characterize the speaker. Following this, a specific speaker codebook is formed based on the cluster (small representative feature vectors). This is based on the principle of block coding under the lossy data compression

method fixed to the fixed-length algorithm (Md Rashidul Hasan, 2004). The following Figure 2.8 shows the 2-dimensional representation of Vector Quantization.

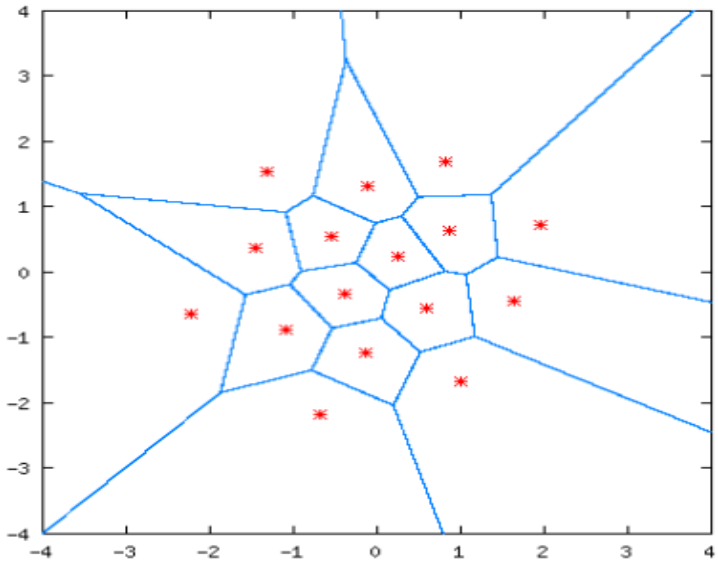


Figure 2.8: Depicting 2-dimensional Vector Quantization.

Each pair of numbers falling in a particular region is approximated by a star associated with the region. The stars are called as codevectors and the region shown by the borders are called encoding regions. The codebook is a set of all codevectors and the set of all encoding regions is called the partition of space. The following Figure 2.9 illustrates the recognition processes.

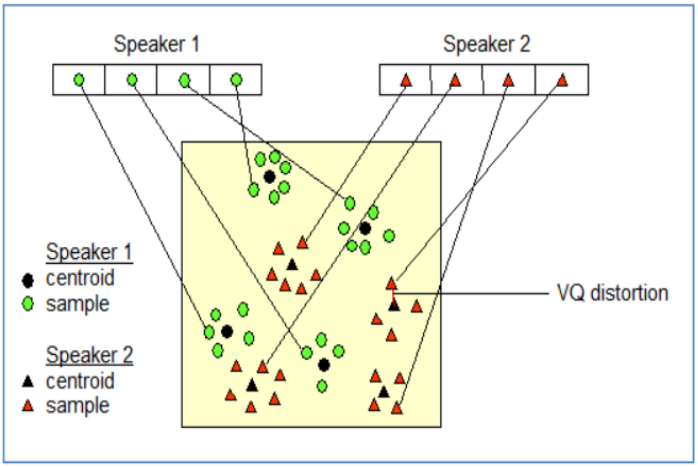


Figure 2.9. Illustration of the recognition process

Apart from the above review specifically related to the parameter MFCC, Tiwari (2010) used MFCCs to extract, characterize, and recognize the information about speaker identity. During Mel-frequency wrapping the subjective spectrum was stimulated using a filter bank. The author used a different number of filter settings (12, 22, 32, and 42) to check

its effectiveness. Out of these, the results showed 85% effectiveness using MFCCs with 32 filters in the speaker recognition task.

MFCCs were also used to study the influence of the nasal co-articulation in Malayalam language samples and an attempt was made to obtain a benchmark for the same. Jyotsna (2011) studied speaker identification using cepstral coefficients and MFCCs in Malayalam nasal coarticulation. Results showed using cepstral coefficients, the benchmark for speaker identification was 80%, and using MFCCs it was 90% for nasal co-articulation in Malayalam.

Sukor and Syafiq (2012) conducted a study on speaker identification using MFCCs procedure and noise reduction method. The study is an implementation of speech recognition as medium security access control to restricted services such as phone banking system, voice mail, or access to database services. The background noise was removed by passing the signal to the pre-treatment process. Then the MFCCs method was used to extract the features from the speech signal. Then features will be matched in the database using vector quantization. The main goal of the study was speaker identification, where a speech signal from an unknown speaker was compared with the database of the known speaker using text-dependent utterances. From the experimental results, this method has explained that it was able to recognize the correct voice pattern.

Ridha (2014) studied the benchmark for speaker identification using nasal continuants in Hindi speakers. Nasals /m/, /n/ and /ŋ/ were chosen which were embedded in words in all positions. Results revealed 100%, 90% and 100% of correct identification obtained for /m/, /n/ and /ŋ/ respectively when live recording was compared with live recording. Meanwhile, when samples were compared within the same recording conditions (mobile network recording was compared with mobile network recording) the percent correct identification was 50%, 80%, and 90% respectively. Among /m/, /n/ and /ŋ/, /ŋ/ had best percent correct speaker identification except under telephone equalized/ not equalized conditions. Under these conditions, /m/ had the best percent correct speaker identification. Similar findings were reported by Ayesha (2016), where the percent correct speaker identification score for /m/, /n/, and /ŋ/ was 70%, 80%, and 100%, respectively when samples from the same recording conditions were compared within the same recording conditions (direct recording were compared with direct recording) using MFCCs. The percent correct speaker identification score for /m/, /n/, and /ŋ/ was 60%, 70%, and 60%, respectively when samples from the same recording conditions were compared within the same recording conditions (network recording were compared with network recording) using MFCC. The percent correct speaker identification scores decreased drastically when network recording was compared with network recording. Overall, the results revealed that the velar nasal continuant /ɳ/ had the best percent correct speaker identification in this study.

Nithya (2015) reported a benchmark for speaker identification using three Tamil nasal continuants in live recording and mobile network recording conditions. Results of the study showed that the percentage of correct identification in live recording condition for /m/, /n/

and /n̥/ was 97.6%, 85.6% and 76.5% and in mobile network conditions the scores were 83.5%, 65.8% and 68.3% respectively.

Chandrika (2015) reported a benchmark for speaker identification using three Kannada nasal continuants in live recording and mobile network recording conditions. The author had also compared the MFCCs across three age groups of 20≤30 years, 30≤40 years, and 40≤50 years. Results of the study revealed that the nasal continuant /n./ had the highest percentage of correct speaker identification score in case of direct recording and /m/ and /n/ had the highest score in case of network recorded samples.

2.4.3.1 Speaker Identification studies on vowels

Generally, in most forensic analysis, the significant phonemic cues of certain phonemes only will be considered. Among these, speech sounds, vowels, nasals, and fricatives (in decreasing order) provide better speaker recognition compared to plosives. This is because they are comparatively easy to be identified in speech signals and their spectra contain features that reliably distinguish speakers (Shaughnessy, 1987; Sigmund, 2008). Vowels have proven to be effective for characterizing individual speakers and have been widely used for speaker recognition and forensic analysis.

To list out other few Indian reviews, for example, Jakhar (2009) studied the benchmark for text-dependent speaker identification in the Hindi language using cepstrum. Live and telephonic recordings were done. For five speakers, the results in terms of highest speaker identification scores were 83.33%, 81.67% and 78.33% for vowel /a:/, /i:/ and /u:/ respectively. For ten speakers, the results in terms of highest speaker identification scores were 81.67%, 68.33% and 68.33% for vowel a:/, /i:/ and /u:/ respectively. Whereas for twenty speakers the results in terms of highest speaker identification scores were 60%, 50%, and 43.33% for vowel a:/, /i:/ and /u:/ respectively for the conditions such as live v/s live, mobile v/s mobile, and live v/s mobile respectively. The results indicated that as the number of speakers increase, the percentage of correct speaker identification decreases, and also scores are better when conditions are similar. Among /a:/, /i:/ and /u:/, /a:/ yielded better results in live recording and vowel /i:/ in mobile recording condition.

With reference to the previous study on speaker identification using cepstrum, Sreevidya (2010) conducted a study to check the benchmark in the Kannada language by text-independent speaker identification method using cepstrum in both direct and mobile recording conditions. The results of the study showed indirect speech and reading, vowel /u:/ had the highest score (70 and 80%), and vowel /i:/ had the highest score (70 and 67%). Also, the study quoted that for both the direct verse mobile recordings, for all vowels, and groups of speakers, the results were below chance level.

Medha (2010) studied the benchmarks for speaker identification of three long vowels /a:/, /i:/, and /u:/ using cepstral coefficients on text-independent data in the Hindi language. Among 20 Hindi speakers who participated in the study, 10 were males and 10 were females.

For females, the percent correct speaker identification scores were 40%, 40% and 20% for /a:/, /i:/ and /u:/ respectively. Whereas for males, it was 80%, 80% and 20% for /a:/, /i:/ and /u:/ respectively. Therefore, the benchmarking for female speakers was below chance level whereas for male speakers it was 80% for the vowels /a:/ and /i:/. Hence the study concluded that in text-independent conditions, the extraction of cepstral coefficient quefrency and amplitude is useful in speaker identification for vowels /a:/ and /i:/ only in males.

Chandrika (2010) compared the efficacy of a speaker verification system using MFCCs in the Kannada language. Ten Speakers participated in the study and the material consisted of long vowels (/a:/, /i:/, and /u:/) in medial position occurring in five targets Kannada words embedded in sentences (text-dependent). Speech recording was carried out in two conditions: mobile network and digital recording. MFCCs values were extracted for all the long vowels and the results indicated an overall verification of 80%. The overall performance of speaker recognition was 90% to 95% for the vowel /i:/ whereas, the accuracy of performance of vowel /i:/ was marginally better than /a:/ and /u:/.

Ramya (2011) used electronic vocal disguise and checked speaker identification using MFCCs. The percent correct identification was beyond chance level for electronic vocal disguise for females. Interestingly vowel /u: / had higher percent identification (96.66%) than vowel /a: / 93.33 %, and /i: / 93.33%.

Suman and Hema (2015) aimed at establishing the benchmark for speaker identification using MFCCs on vowels following nasal continuants in Kannada language. Total of twenty males participated in the study. Sentences consists of /a:/ /i:/ and /u:/ vowels following the nasal continuants /m/ and /n/. Recordings were done in live and mobile network recording. Results revealed in live recording, on comparison among the three vowels following the nasal continuant /m/, /i: / was better followed by /a: / and /u:/. Whereas for the nasal continuant /n/ the vowel /a: / was better followed by /i: / and /u: /. In mobile recording, on comparison among the three vowels following the nasal continuant /m/, /a: / was better followed by /i: / and /u: /. Similarly, for the nasal continuant /n/ the vowel /a: / and /u: / were better followed by /i: /. In live verses mobile recording, on comparison among the three vowels following the nasal continuant /m/, /i: / was better followed by /a: / and /u: /. Whereas, for the nasal continuant /n/ the vowel /i: / and /u: / were better followed by /a: /.

Arjun (2015) studied the benchmark for speaker identification using Kannada vowels preceding nasal continuants. Twenty males participated in the study. The sentences consisted of words with three basic vowels /a:/, /i:/ and /u:/ preceding nasal consonants /m/ and /n/. Recording was done under two conditions live and mobile network recording. Results for vowel /a:/ showed 93% of correct speaker identification than vowel /i:/ and /u:/ across nasals /m/ and /n/, both (/i:/ & /u:/) showed similar results with 84% when live recording was compared with live recording when 20 speakers were considered. When mobile network compared with mobile network vowel /a:/ with 74% of correct speaker identification than vowel /i:/ and /u:/ and across nasals, /m/ with 61% for 20 speakers. Vowel /a:/ with 94% and both the nasals /m/ and /n/ with 88% had the highest correct identification scores when the

live recording was compared with live recording considering 10 speakers. Vowel /a:/ with 83% and the nasal /n/ with 71% had the highest correct identification scores when the mobile network was compared with mobile network recording for 10 speakers. Hence the study concluded that vowel /a:/ preceding both nasals /m/ and /n/ are best for speaker identification compared to other vowels.

Aswathy (2016) studied the effect of native versus non-native languages in speaker identification in the lab recording condition. Ten male participants were taken for the study in the age range of 20 to 25 years with Kannada as their mother tongue. The material consists of 10 hypothetical Malayalam sentences and 15 Kannada sentences containing vowels /a:/, /i:/, and /u:/ in the word medial position. The samples were recorded and analyzed using PRAAT Software and SSL Workbench. Results revealed average percent correct speaker identification for vowels /a:/, /i:/ and /u:/ to be 95% for condition I (Kannada language v/s Kannada language). For condition II (Malayalam language verses Malayalam language) average percent correct speaker identification for vowels /a:/, /i:/ and /u:/ were 94%, 87% and 75% where vowel /a:/ was found to be the better on comparison with /i:/ and /u:/ for speaker identification through MFCCs. For the Condition III (Kannada Verses Malayalam) average percent correct speaker identification for vowels /a:/, /i:/ and /u:/ were 92%, 79% and 73% where vowel /a:/ was found to be the better followed by /i:/ and then /u:/ for speaker identification through MFCC. Therefore the study concluded that vowel /a:/ acts as a better cue for speaker identification irrespective of the language used when compared to /i:/ and /u:/.

To summarize, most of studies reports that vowel /i:/ was better compared to /a:/ and /u:/ for speaker identification. Chandrika (2010) and Jakhar (2009) found vowel /a:/ to be better in live conditions and vowel /i:/ in mobile network conditions. Arjun (2015) and Aswathy (2016) found vowel /a:/ was better compared to /i:/ and /u:/. Medha (2010) found vowel /a:/ and /i:/ were better. Interestingly Sreevidya (2010) and Ramya (2011) found a high percent speaker identification for /u:/. However, Suman (2015) found mixed results across vowels. Thus, the review provides the significance of MFCCs, noise reduction methods, and vowels to be considered or essential for speaker identification.

It is evident from these reviews that MFCCs are perhaps the best parameter for speaker identification and less susceptible to variation of the speaker's voice and surrounding environment (noise). Also, the vowels may be the most suitable among speech sounds for speaker identification. However, to date, there are limited studies on vowels as strong phonemes for speaker identification using semi-automatic methods in the presence and absence of noisy situations and after the application of speech signal to any noise reduction techniques. In the present study, the Sound Cleaner software (speaker recognition instrument) is used to reduce the noise and study the effect of the same on speaker identification. In forensic sciences, the scientific testimony has to be provided to impress any court of law and from whichever country the research would have been executed. However, for any result to be called scientific, it has to be measured, quantified, and reproducible if and when the need

arises. Therefore, a method to carry out these analyses becomes a must. In this context, the present study was conducted. Hence, the present study aimed to investigate the effect of noise and noise reduction techniques on speaker identification with reference to the parameter MFCC on the long vowels in the Kannada language.

Thus, the objectives of the study were *initially* (1), (2), and (3) to evaluate the percent correct Speaker Identification using MFCCs on the long vowels in the Kannada language for lab recording conditions and field recording (embedded with noise)) before and after the application of the noise reduction technique. *Next* (4) to compare speaker identification using MFCCs on long vowels in the Kannada language in lab recording conditions versus field recording (embedded with noise) before the application of noise reduction technique. *Later* (5) to evaluate the percent correct Speaker Identification using MFCCs on the long vowels in the Kannada language for lab condition versus field recording after the application of noise reduction technique (probably embedded without noise). *Finally* (6) to compare the percent correct Speaker Identification using MFCCs on the vowels in the Kannada language for field recording (embedded with noise) before the application of noise reduction technique versus field recording after the application of noise reduction technique (probably embedded without noise).

1. Lab condition.
2. Traffic condition before noise reduction technique
3. Traffic condition after noise reduction technique
4. Lab recording versus traffic recording before noise reduction technique. (2)
5. Lab recording versus traffic recording after noise reduction technique. (5)
6. Traffic condition before noise reduction technique versus Traffic condition after noise reduction technique (4)

CHAPTER III

METHOD

3.1 Participants

Native Kannada language-speaking neuro-typical adult male and female from in and around Mysuru was considered as participants. They had a minimum of ten years of formal education with the Kannada language as one of the subjects and all the participants belonged to the Mysuru dialect of Kannada language and were drawn from the work/residential place in and around Mysuru, Karnataka, India. A total of 60 participants with 30 males and 30 females in the age range of 20-40 years were considered for the study and the demographic details are listed in Table 3.1. The inclusion criteria for the participants were no history of speech, language, hearing problems, no associated psychological or neurological problems, and no reasonable cold or respiratory conditions at the time of recording, and normal oral structure. The hearing was screened using the Ling sound test (Ling, 1978) or (Administration of Screening checklist from POCD, AIISH, Mysuru). Kannada Diagnostic Picture Articulation Test (KDPAT- Appendix A) (Deepa & Savithri, 2010) was administered by a Speech-Language Pathologist to rule out any misarticulations present in the speech. All were native speakers of the Kannada language and used English and Hindi as their second language and very few were aware of the third language and it was reported to be Tamil and Telegu. All the participants had varied professions. The profession, education, and language use were not controlled, however, the major inclusionary criteria were the participants speaking Kannada as their first language and only this was focused to consider any individuals as the participants for the present study.

Table 3.1 Demographic details of the participants

Participant No.	Name	Age/gender	Education
1.	Bhanumathi. S. N.	29years/Female	UG
2.	Shobha	30 years/Female	UG
3.	Bhagya	30 years/Female	PG
4.	Shrilekha. B.	23 years/Female	PG
5.	Padmashree. B	24 years/Female	UG
6.	Bhuvana. S.	24 years/Female	PG
7.	Navya. B. N.	24 years/Female	PG
8.	Anitha. K. B.	28 years/Female	UG
9.	Rajeshwari	28 years/Female	UG
10.	Meenakshi Ghasti	32 years/Female	UG
11.	Shruthi. M. S	32 years/Female	UG
12.	Poornima. N.	32 years/Female	PG
13.	Renu Raju Neelgar	32 years/Female	PG
14.	Nagarathna. M. N.	23 years/Female	UG
15.	Latha	33 years/Female	UG

16.	Shashirekha	36 years/Female	UG
17.	Radhamma	37years/Female	UG
18.	Bhagyalakshmi	37years/ Female	UG
19.	Anitha. S.	38years/Female	PG
20.	Ashwini. A	38years/Female	PG
21.	Meenakshi. K. C.	40 years/Female	PG
22.	Nagalakshmi . S. L.	42 years/Female	PG
23.	Sreenidhi. K.A	34 years/Female	UG
24.	Manasa	28 years/Female	PG
25.	Megha. J	32years/Female	PG
26.	Akshatha	30 years/Female	UG
27.	Bhagya. R	32 years/Female	UG
28.	Madhu	29 years/Female	PG
29.	Jyothi. K	30 years/Female	UG
30.	Preethi. G	25 years/Female	PG
31.	Manjesh	28 years/Male	PG
32.	Veeresh. K. V.	32 years/Male	UG
33.	Ramu. K.	27 years/Male	UG
34.	Sandeep	30 years/Male	UG
35.	Puneet Kumar.T	29 years/Male	UG
36.	Raghavendra. R	28 years/Male	UG
37.	Vinod. P	28 years/Male	UG
38.	Raghavendra Nalatawad	28 years/Male	PG
39.	Shivakumar	28 years/Male	PG
40.	Raghava Kumar	27 years/Male	PG
41.	Umaprasad	27 years/Male	UG
42.	Manjunath. S. R.	27 years/Male	UG
43.	Pradeep Kumar	24 years/Male	PG
44.	Deepak. P	23 years/Male	PG
45.	Naveen Kumar. R	30 years/Male	PG
46.	Raghavendra. G. N	31 years/Male	UG
47.	Manjegowda	32 years/Male	PG
48.	Ranachandra	32 years/Male	UG
49.	C. Chethan	32 years/Male	PG
50.	K. M. Yogananda	34 years/Male	UG
51.	Shridhar. R	34 years/Male	PG
52.	Mallikarjuna	35 years/Male	PG
53.	Jagadeesha	35 years/Male	PG
54.	Vinay Nag	35 years/Male	UG
55.	Mahesh. E	38 years/Male	UG
56.	Devaraje Gowda	38 years/Male	UG
57.	Shreepathi	38 years/Male	PG
58.	Shivappa. S.	39 years/Male	PG
59.	Dharshan Hiremath	20 years/Male	PG
60.	Harisha	23 years/Male	PG

Note: * UG- Undergraduate and PG-Post Graduate

3.2 Procedure

3.2.1 Material

Hypothetical Kannada meaningful sentences (forensic speech sample) with commonly occurring long vowels /a:/, /i:/, /u:/ embedded in the medial position of the twenty-one words of nineteen sentences. Among which fifteen words were only considered for the study and these target words formed the material for the recording task which is listed in Table 1 of Appendix B.

3.2.2 Recording Software

Speech samples of participants were recorded individually. The sentences were presented visually and participants were instructed to read the sentences in a normal modal voice. The written material was provided to the participants and was made familiarized before recording begins. The recording was done for three trails (Trail I, II, and III). Vowels occurring consecutively five times in the sentences of Trial II and III only were selected for analysis out of three Trails. Where Trial I acted as a model setter for the following two trails. Participants were informed about the nature of the study. Written consent (Appendix C) was taken from all the participants. The recordings were done in two different conditions: *Condition I*- Laboratory recording and *Condition II*- Traffic Field recording. The time gap between these two conditions was two weeks. For lab recording condition, Computerized Speech Lab (CSL 4500 model; Kay PENTAX, New Jersey, USA) (St. Petersburg, Russia, Speech Technology Center) was used. A desired 16 Bit (analog-digital) computer memory was used (i.e., sample frequency of 16 kHz) and later for analysis, it was converter at a required sampling frequency of 8 kHz using PRAAT software. The distance between the mouth and the dynamic microphone (Shure) was kept constant at approximately 10 cm. These recordings were stored in ***.wav format***. This would consist of the participant's speech sample (target) recording for 4-5 minutes (19 sentences repeated for 3 trials). After two weeks of a gap, the field recording was carried out.

In this two weeks gap, a pilot study for 10 speakers (5 males and 5 females) in two field conditions were carried out to decide on the better noisy condition. Under field conditions, recordings were done in traffic and canteen conditions. For these field conditions, **Olympus digital voice recorder** (LS100) with attached dynamic microphone (Shure) was used for recording the participants' speech in the situations like traffic and canteen condition having background noise of around 80 dB (A) (Kalaiselvi & Ramachandraiah, 2010). In the two field recordings, the digital voice recorder was turned 'on', and the dynamic microphone (Shure) was kept constant at approximately 10 cm. Before the target speech recordings of the participants, as an initial recording, the ambient noise was recorded for 5-10 seconds which ascertain the reference of ambient noise. This was followed by a participant's speech sample (target) recording for 4-5 minutes (19 sentences repeated for 3 trials). The field recording samples were transferred from a digital voice recorder to a computer using a USB cable. The samples were stored in ***.wav files*** so that the analysis could be carried out efficiently. Among the two field conditions, there was constant background noise for the traffic condition and

very intermittent background noise for the canteen condition. For the present study, the constant background noise was very important, since it is easy to identify and process when the noise is continuous than intermittent and also for any noise reduction technique to be applied. Hence, the traffic condition was preferred as a field recording condition for all the participants. Thus, the recorded samples of *Condition I* and *Condition II* were uploaded to one common computer for further analysis.

3.2.3 Analysis software

3.2.3.1 Sound Cleaner Software

The individually recorded samples were analyzed under three Steps: Step I, the audio files of condition I and condition II was not subjected to any noise reduction algorithm. In Step II all the audio samples were subjected to noise reduction algorithm using Sound Cleaner Signal Enhancement Program (Noise Cancellation Software, Model 5142) (Kay PENTAX- A Division of PENTAX Medical Company, Lincoln Park, New Jersey, USA & Speech Technology Center, St.Petersburg, Russia). Step III was the final analysis using WORKBENCH software. The Sound Cleaner Program consists of a number of inbuilt modules/scheme with a series of inbuilt sub-modules such as, 'Input', 'Waveform-input', 'Frequency Compensation', 'Equalizer', 'Inverse Filter', 'Broadband Filter', 'Dynamic Filter', 'Clipping', 'Amplifying', 'Mu-Transform', 'Waveform-output', 'Slowing', 'Output/file' and 'Speaker'. Each of the windows belongs to a separate processing module and the entire signal processing scheme consists of a number of modules. Before processing the signal it was saved in a .wav file. Then 'play' button from the 'Input' process module was pressed where the data starts flowing from the starting 'Input' process module (.wav file) to the final one (.wav file) through intermediate/sub-modules. A pilot study was carried out to construct the necessary combination of modules for processing, thus supporting the flexibility of the process scheme by adjusting to the concrete noise parameters.

Thus in a **PILOT** study, 5 participants' speech samples were subjected to the different permutations (at least 2-3) of the processing scheme of Sound Cleaner software consisting of a single or a number of the processing modules mentioned above. The output of these processed speech samples (analysis after Step II) of 5 participants was subjected to the perceptual judgment of noise reduction on a 4 point perceptual rating scale. The rating was 0 = 0-25% noise reduction, 1 = 25 to 50% noise reduction, 2 = 50-75% noise reduction, 3=75-100% noise reduction. This rating was done for the speech samples which underwent various individual modules or a combination of modules with schemes in the Sound Cleaner software. To mention a few: broadband filter, **street noise scheme**, street noise scheme with equalizer, street noise scheme with inverse filtering, street noise scheme with frequency compensation, and street noise scheme with impulse filter. All these individual modules or combinations were experimented with and found that the 'Street Noise Scheme' was more appropriate for the present study based on the rating obtained from the perceptual judgment task. Hence the same was selected as the target module for processing and thus the same was used for the main study. After a pilot study, the inbuilt setting of noise reduction technique loaded in Sound Cleaner software called 'Street Noise Scheme' was used for the present

study. ‘Street noise’ scheme consists of modules such as, ‘Input’, ‘Waveform-input’, ‘Broadband Filter’, ‘Dynamic Filter’, ‘Output/file’, and ‘Speaker’. ‘Broadband Filter’ was set at its default settings and for ‘Dynamic Filter’, which has options such as ‘strong signal’ and ‘weak signal’, where ‘strong signal’ was remained as ‘strong’ and the weak signal was ‘weakened’ and the threshold was kept at 4kHz since the speech frequency range till 4 kHz. Data flows from the starting ‘Input’ process module (.wav file) to the final one (.wav file) through intermediate modules such as ‘Broadband Filter’ and ‘Dynamic Filter’ and thus the sample was processed and saved as an output file. Figure 3.1, depicts the Sound Cleaner software windows and the following figures (3.2 to 3.7) represent the series of steps involved in the noise reduction technique.

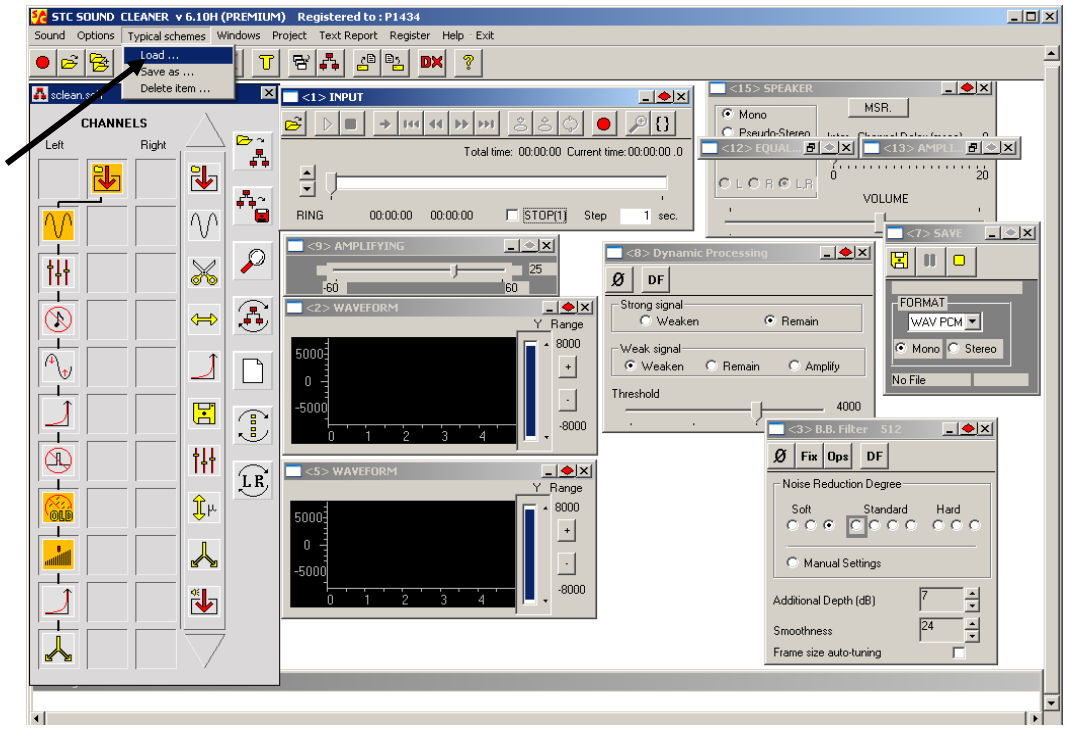


Figure 3.1- The main window in Sound Cleaner software to load typical schemes.

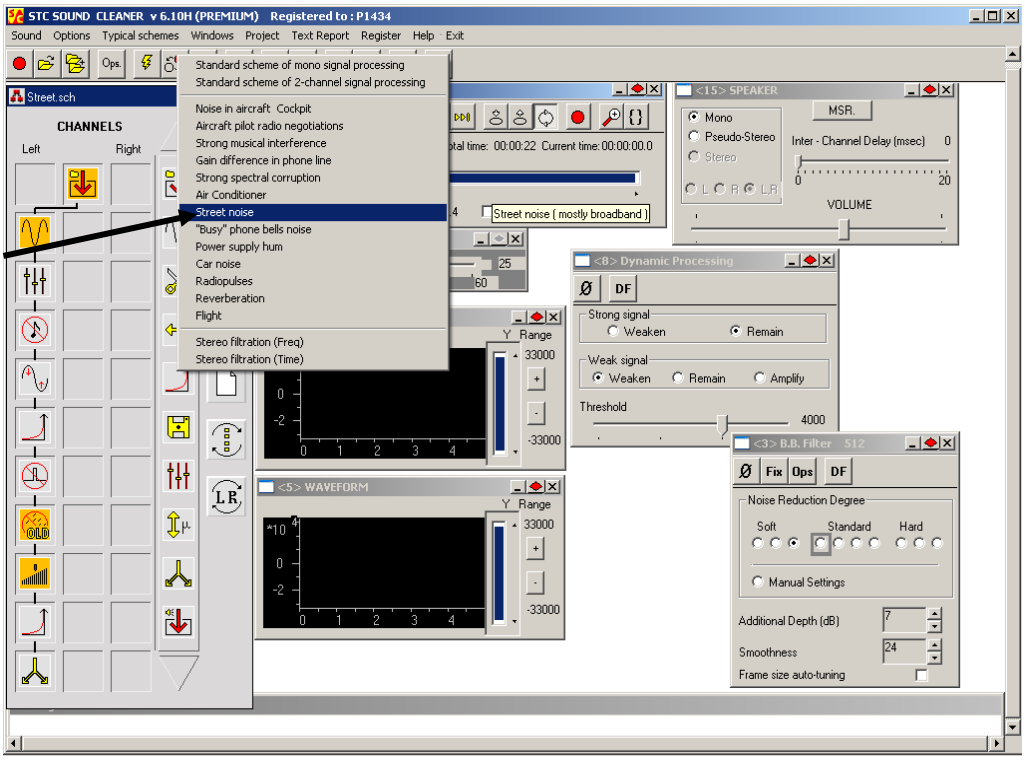


Figure 3.2- Selection of 'Street Noise Scheme' amongst the choice of other built-in schemes

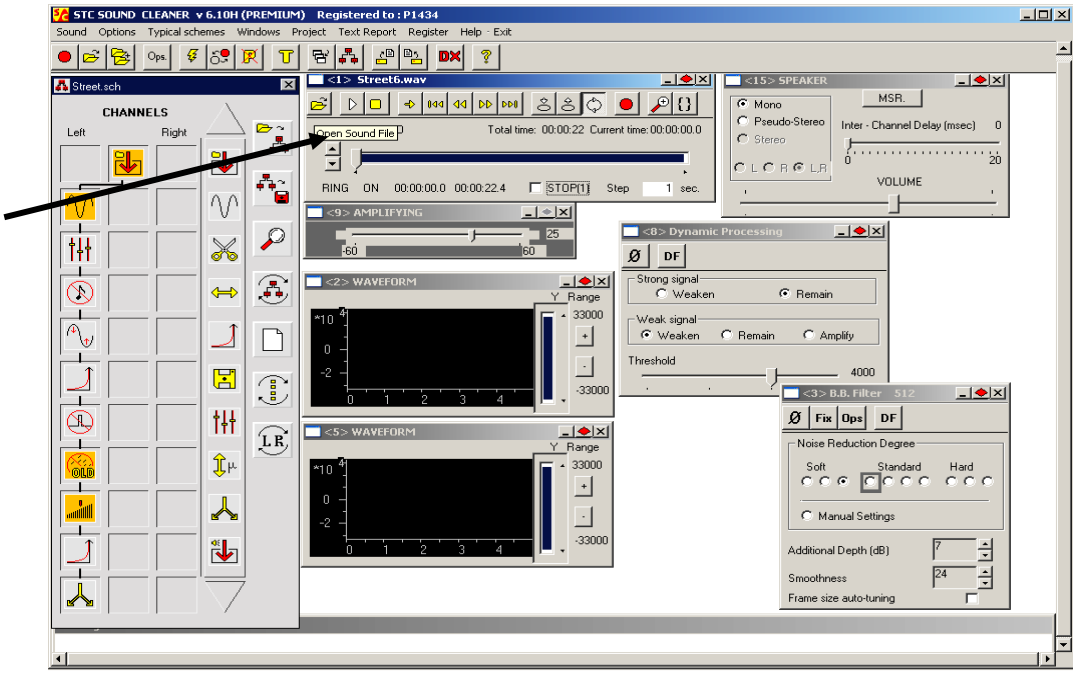


Figure 3.3- Window opened to load sound file & apply sound reduction technique

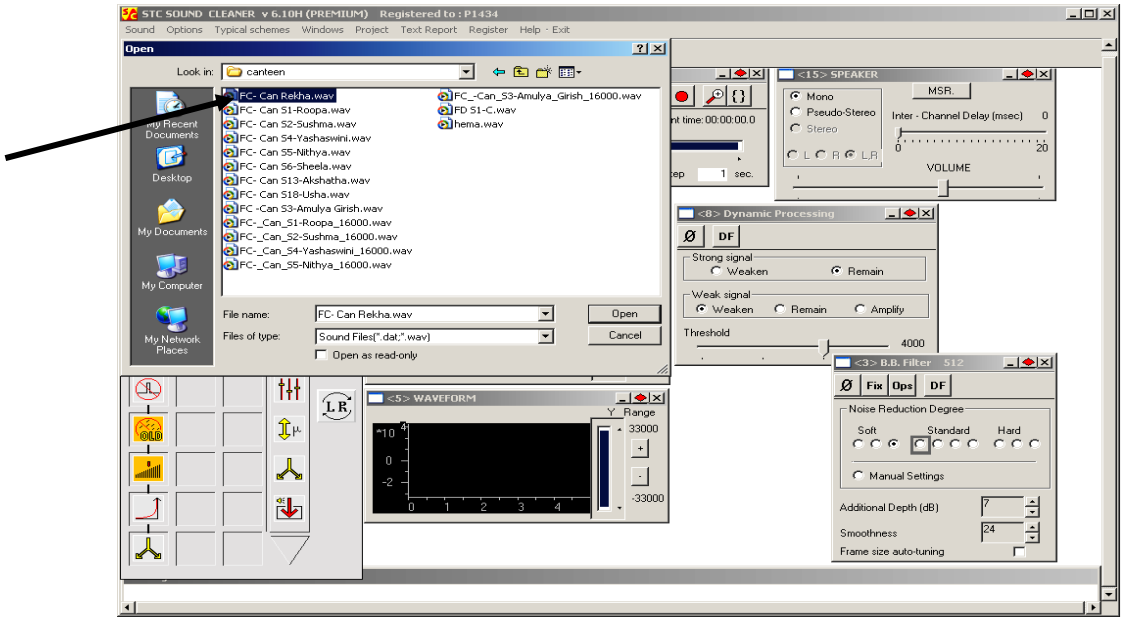


Figure 3.4- Sound file selected from the existing destination before the application of sound reduction technique

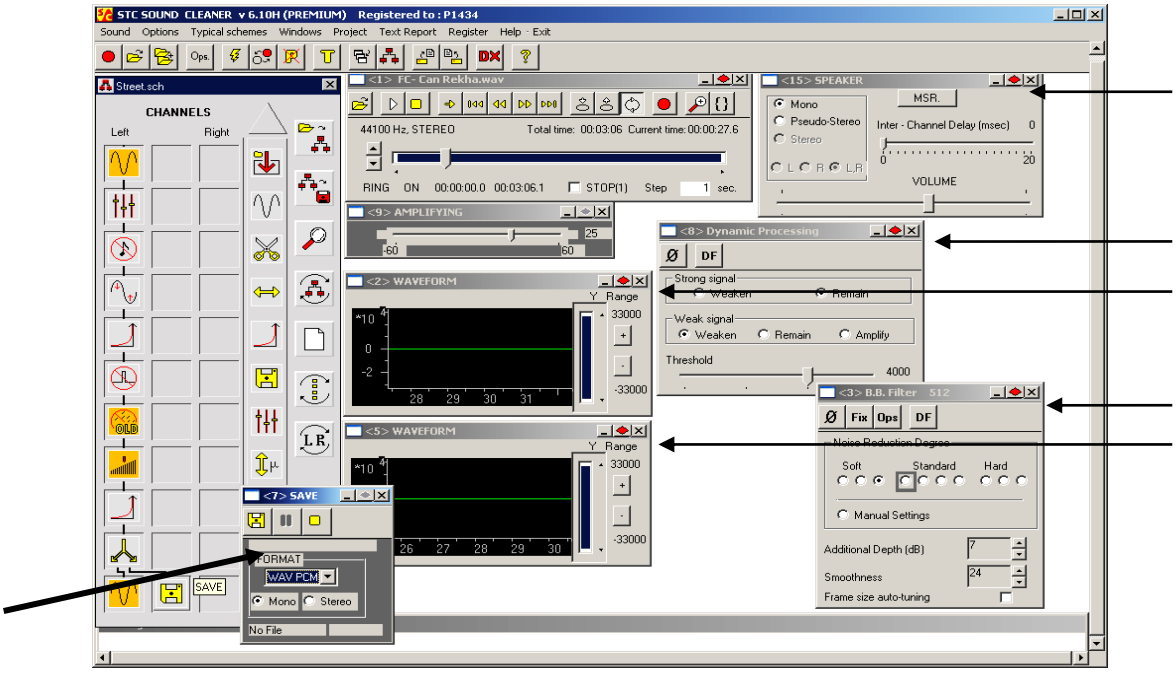


Figure 3.5- Output file created before the initiation of the sound reduction technique. ‘Input’, ‘Waveform-input’, ‘Broadband Filter’, ‘Dynamic Filter’, ‘Output/file’, and ‘Speaker’. ‘Broadband Filter’ set as default settings (Red dot on all the modules)

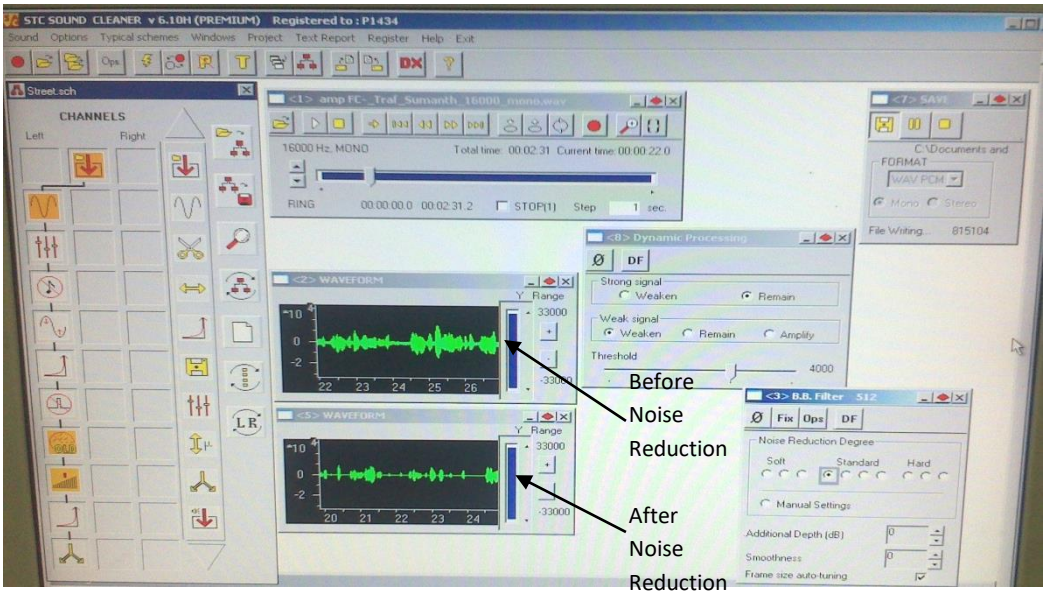


Figure 3.6- Signal during the sound reduction processes

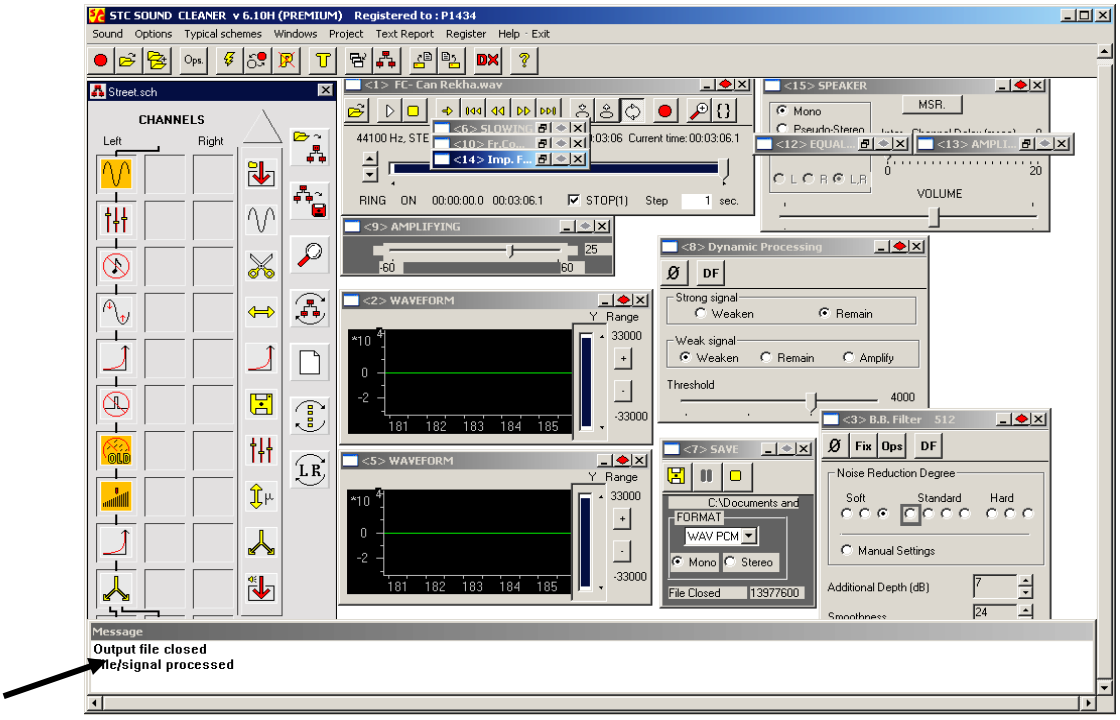


Figure 3.7- Window after completion of sound reduction processes

To explain further, generally, in the dynamic processing module there would be alteration in the dynamic range of the signal. The common process of operations would be compression and expansion. In compression, the dynamic range of the signal will be reduced (minor difference in level among the soft and loud signal parts). Whereas in expansion the

dynamic range of the signal will be enlarged, generally the soft parts of the signal will be enhanced. Thus, it is useful in equalizing the loudness of the sound (compressor), enlarging the dynamic range of the sound (expander), attenuating or enhancing selected frequency ranges (dynamic processing in frequency bands), removal of signal parts which is at the level below the given threshold (noise gate) and limiting the maximum signal level value (limiter).

Thus, the pilot study facilitated in construction of the necessary combination of modules for processing and supports the flexibility of the process scheme by adjusting to the concrete noise parameters of the target speech signal of the participants. The same analysis procedure was carried out for the entire sample of all the participants. The samples of traffic recording condition were only subjected to this noise reduction technique of 'Street Noise Scheme' of sound cleaner software and were saved in a separate folder. This was the sample that had undergone Step II analysis.

3.2.3.2 PRAAT Software

The samples of Step I (before the application of noise reduction technique to the traffic condition of recording) and Step II (after the application of noise reduction technique to the traffic conditions of recording) stored in a separate folder in the CSL 4500 (original sampling frequency of 16 kHz) were opened in PRAAT software (Boersma & Weenink, 2009) and downsampled to 8 kHz. Since further analysis using WORKBENCH software could be done from 4 kHz (frequency distribution of an individual's speech frequency ranges till 4 kHz) up to 8 kHz. Of the three trails of recording, the first recording was not analyzed as the material was novel to the participant and the second and third recordings were only used for analysis and comparison. From the downsampled speech material, the long vowels /a:/, /i:/, and /u:/ in the medial position of the target words were truncated from the wideband bar type of spectrograms using the PRAAT software program and was stored in different folders for each participant for the convenience of further analysis. Three complete cycles (approximately 300 ms) of the long vowels were segmented and pasted onto a particular file name convenient to the investigator. For Ex: Condition (Lab), speaker (No. 1), first occurrence (target word), vowel (target vowel), and first session (Trail II) was given the file name as "***LB_ SPM1_ (thupaaki)_(a)_2.wav***" and saved in a folder with the name ***SPM1*** (Speaker No. 1). Figure 3.8, depicts the segmentation of samples from the vowel /a:/, /i:/ and /u:/.

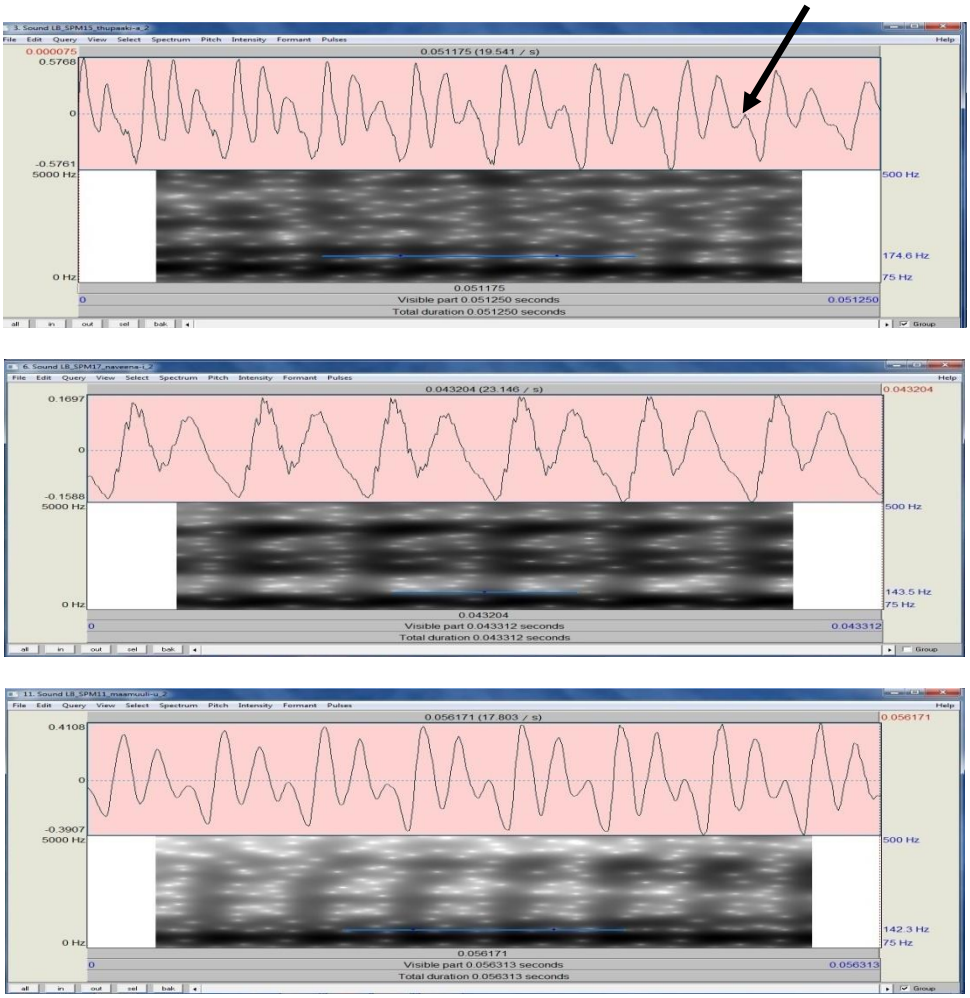


Figure 3.8: Segmentation of samples from the vowel /a:/, /i:/ and /u:/

A total of 15 long vowels consisting of five vowel /a:/, five vowel /i:/ and five vowel /u:/ (from 15 target words) occurring in 19 sentences, in two different conditions (Lab and traffic conditions) and two different phases (phase I lab condition and traffic condition before noise reduction and phase II traffic condition after noise reduction). Thus, the total number of samples for each speaker for phase I was 3 X 5 X 2 X 2 = 60 [3-vowels /a:/. /i:/ and /u:/) X 5 (target words) X 2 (trials-II and III) X 2 (Lab and traffic before noise reduction] and phase II- 3 X 5 X 2 = 30 [3 (vowels /a:/. /i:/ and /u:/) X 5 (target words) X 2 (Trials-II and III)] and the total number of samples for 60 speakers was 5400 (phase I- 60 x 60= 3600 plus phase II- 30x 60 = 1800).

3.2.3.3 Speech Science Lab (SSL) WORKBENCH software

The final analysis which was carried out under Step III was using WORKBENCH software. Speech Science Lab (SSL) WORKBENCH, (Voice and Speech Systems, Bangalore, India) is a Semi-Automatic vocabulary dependent speaker recognition software.

This was used to extract Mel-Frequency Cepstral Coefficients (MFCC) for the truncated (PRAAT software) long vowels.

The foremost thing was to **create a notepad file** and **.dbs file,** the extension of the notepad file was created by specifying the phoneme, speaker, number of sessions, and occurrences and was then segmented. The same can be explained in detail under four headings. 1). *Label*- Here the phoneme or sound being analyzed had to be typed; for example: (/a:/, /i:/, /u:/). 2). *The number of speakers*- This is the number of participants in the study; for example 60. 3). *The number of occurrences of the label*- This is the frequency of occurrence of a sound in a particular stimulus; for example: /a:/ is 5, /i:/ is 5, and /u:/ is 5. 4). *The number of sessions*- This is the number of repetitions of the stimulus; for example: (Trail II and Trail III). Thus, with all these details initially, the file was specified using a notepad and the extension of the same was the .dbs file as shown in Figure 3.9.

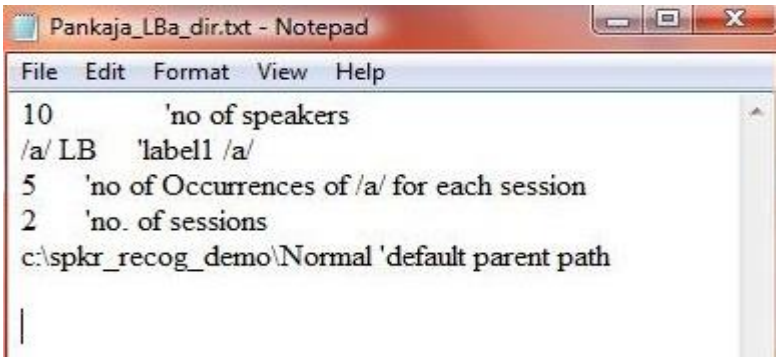


Figure 3.9: Notepad file created for a pilot study.

After creating a notepad file, the next was followed by **segmentation**. Here, the truncated samples were segmented to the workbench software for further analysis. For this, the notepad file was opened in SSL Workbench. After this “label”, “number of occurrences”, and “number of sessions” would appear on the window as they are already fed into the software. This is represented in Figure 3.10.

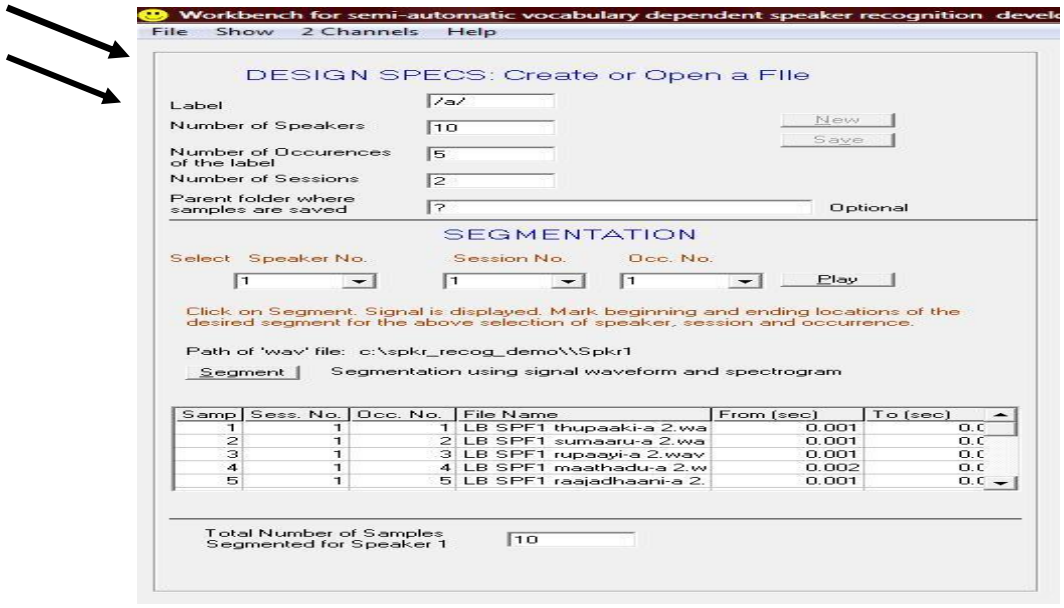


Figure 3.10: SSL Workbench window for analysis.

Following this window of segmentation (Figure 4), the investigator by clicking on the "segment" button which opens the location specified in the parent file path of the notepad file and selects the recording to be analyzed, and marks the segment according to the speaker number, session number, and occurrence number. These specifications facilitate averaging and comparing between the same samples at different sessions. Figure 3.11 represents the specification related to the number of speakers. Similarly, the number of sessions (Trails) and the number of occurrences were also selected. Figure 3.12 illustrates the same.

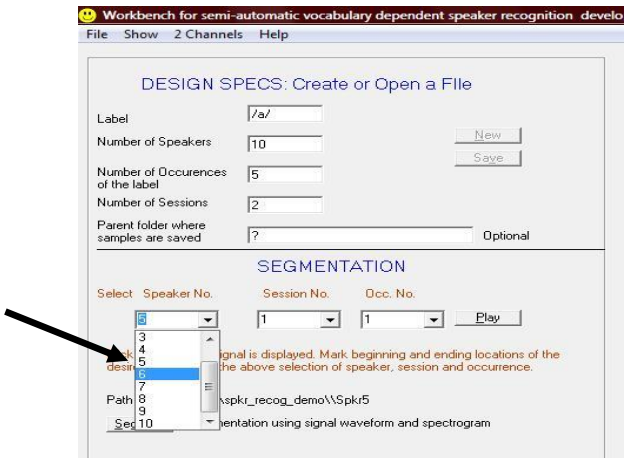


Figure 3.11: Illustration of speaker number being selected for segmentation.

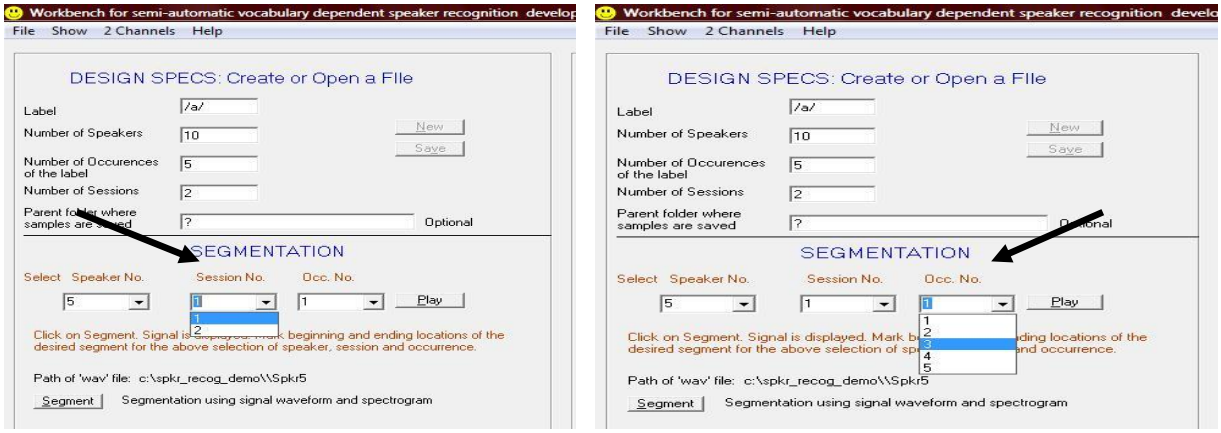


Figure 3.12: Illustration of selecting the session number and occurrence number.

Once these selections were made, the "segment" button was clicked on to open the dialogue box for selecting the file from the parent path specified. Following this, the window would open for segmentation. Figure 3.13 illustrates a segmentation window showing one occurrence of /a/ for a speaker.

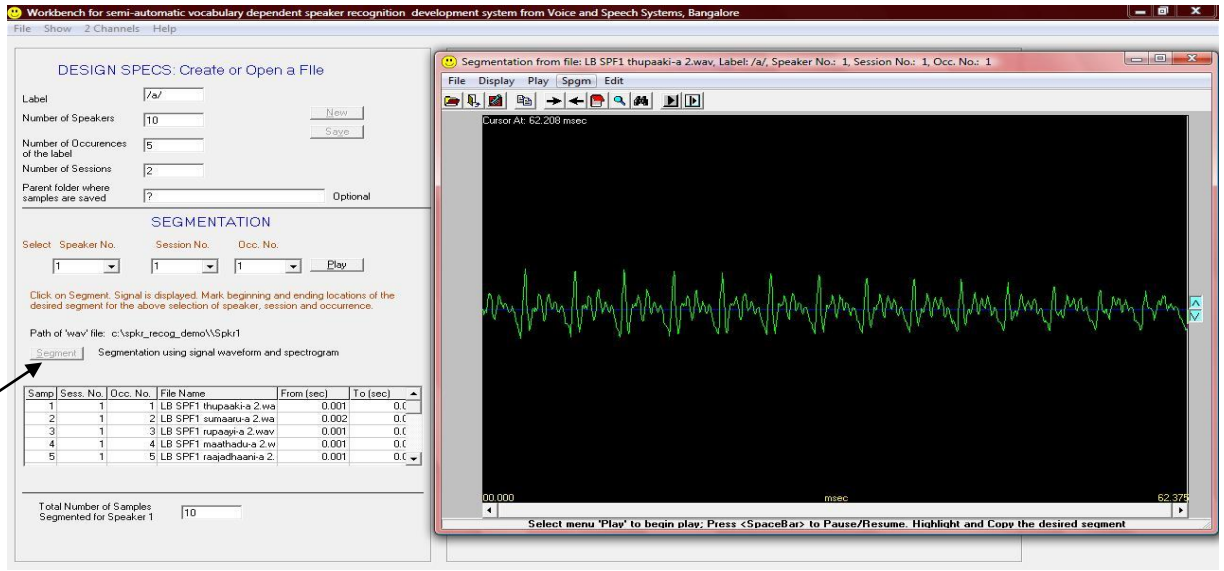


Figure 3.13: Depiction of segmentation window showing one occurrence of /a/ for a speaker.

Then press "Spgm" where the spectrogram window will open. The segment of the file required was selected, and the option of "assign highlighted" was selected from the "Edit" menu. Figure 3.14 illustrates a segmentation window using spectrogram for one occurrence of /a/ for a speaker. After the highlighted segment, a dialogue box will seek confirmation of the assigned segment.

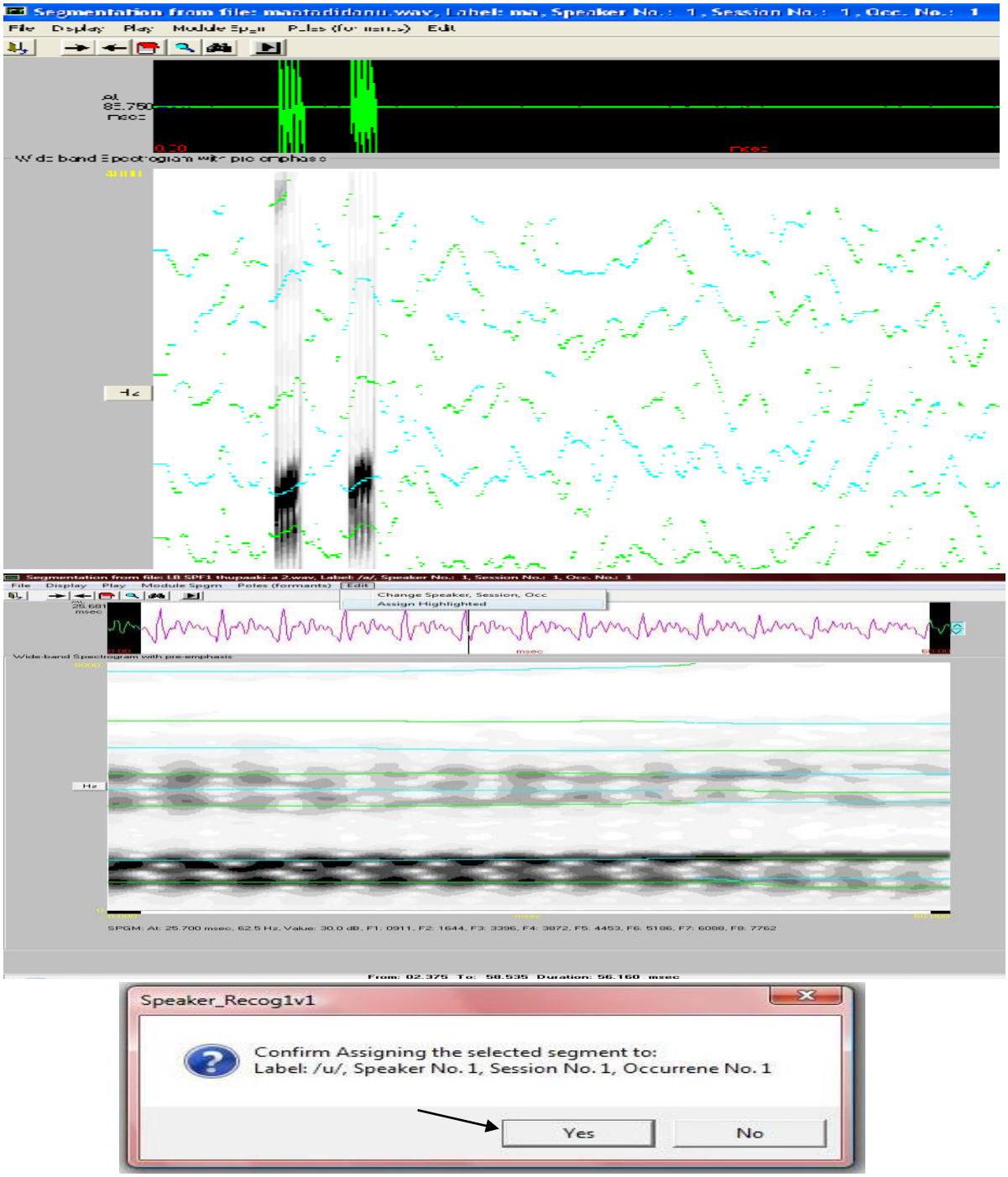


Figure 3.14: Segmentation window using spectrogram for one occurrence of /a/ for a speaker.

Then “save segmentation” option was selected from the “File” menu and the highlighted segment was saved onto the .dbs file created as the extension of the notepad file. Figure 3.15 illustrating the ‘save segmentation’ window and the .dbs file.

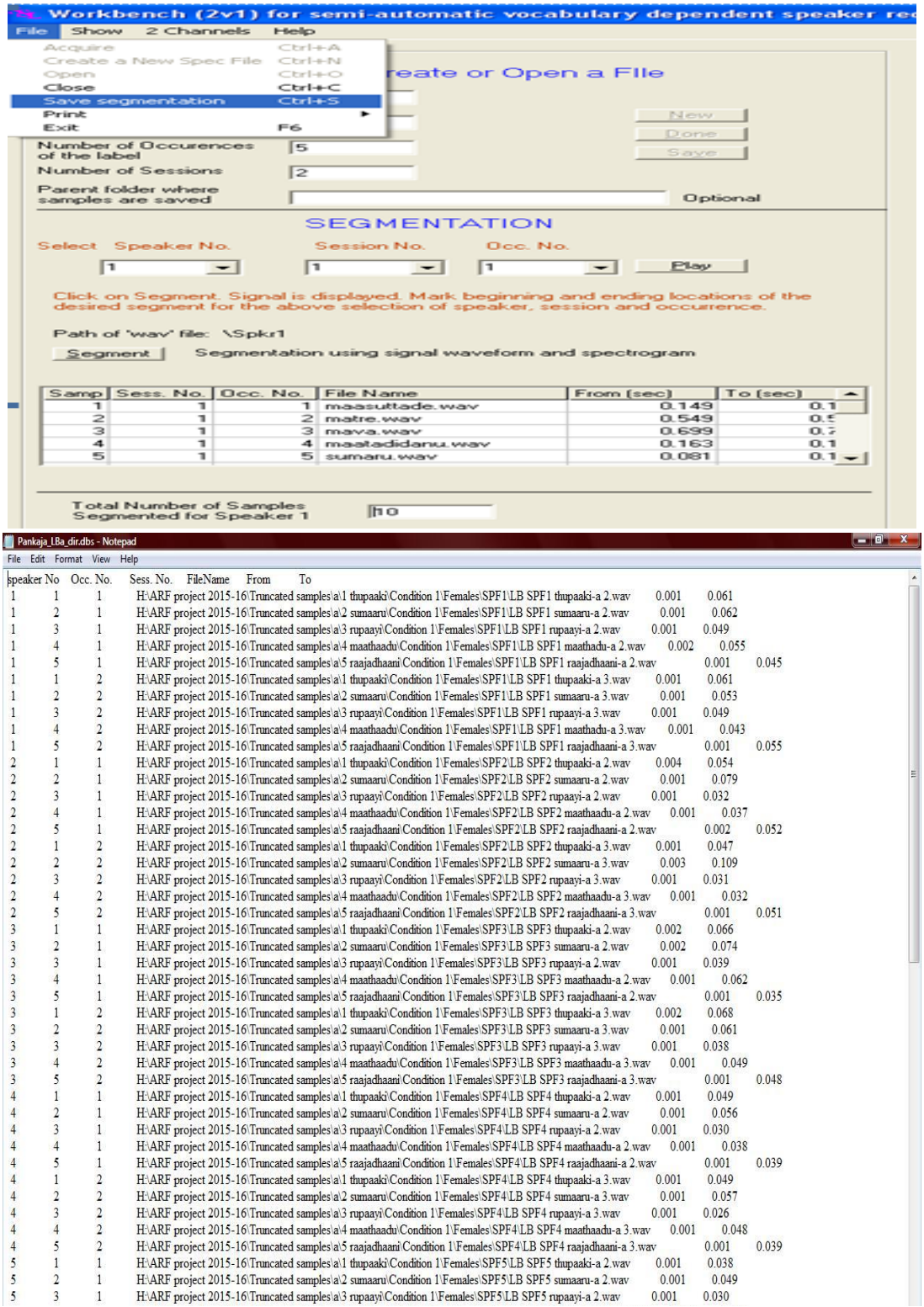


Figure 3.15: Illustrating saves the segmentation window and .dbs file.

The same procedure was carried out for all the speakers, conditions, vowels, and trails, as mentioned earlier in total it was 5400 times of segmentation. Thus, all the files were segmented by selecting the next occurrence number (Example: 2, 3, 4, 5) of the same speaker

(Example: Spk1) and same session (Example: Trail II). When all the samples of all the five occurrences were segmented then the next session (Example: Trail III) of the same speaker was selected and segmented in a similar manner. Once all the samples of all two sessions (Example: Trail II & III) were segmented then speaker two (Example: Spk 2) was selected. Similarly, all the files were segmented in this manner for a total of 60 speakers (Spk 1- Spk 60). Thus, the analysis across the conditions was also done in a similar manner. One example could be lab versus traffic (BNR), hereunder 'session number', the first session was the Trail II of Lab recording and the second session was the Trail III of Traffic recording. Thus, the segmentation was carried out for all the other conditions.

As soon as all the files were segmented the software opens another window to train the samples randomly. The trail/repetitions and utterances of each recording were randomized by the software and were considered as the test set and training set on equal distribution. Thus, the SSL Pro.V4 software was used to test the performance of a distance-based, semiautomatic speaker recognition system, which is vocabulary dependent. After training, 13 MFCC was selected and the sample for identification was tested. Figure 3.16 shows the analysis window of SSL Workbench.

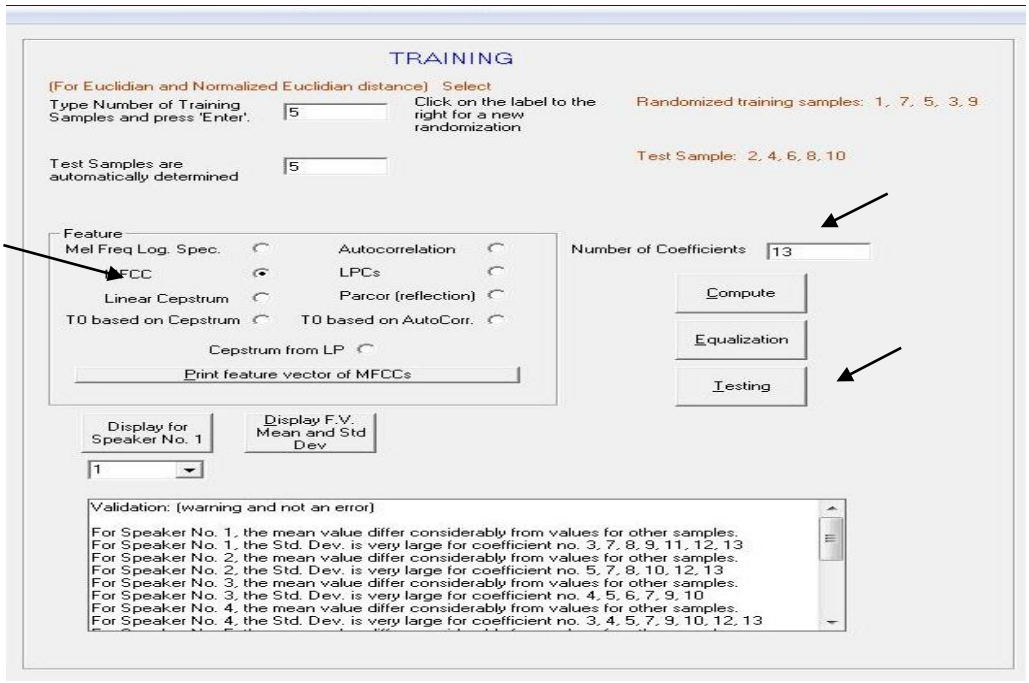


Figure 3.16: Testing window of SSL Workbench.

The segmented material was analyzed to extract 13 MFCCs (In the SSL Workbench, the sampling frequency is 8 kHz and therefore the analysis can be done up to 4 kHz, within 4 kHz only 13 Mel-frequency Cepstral Coefficients (MFCC) can be computed efficiently). The formula for the 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 Figure 3.17 and 3.18, respectively.

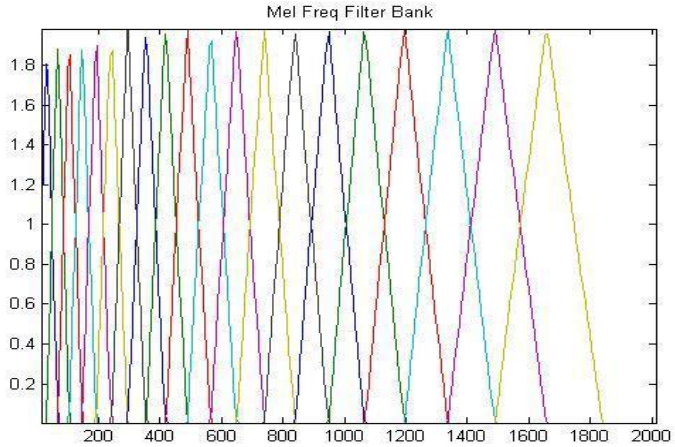


Figure 3.17: Mel frequency filter bank without normalization.

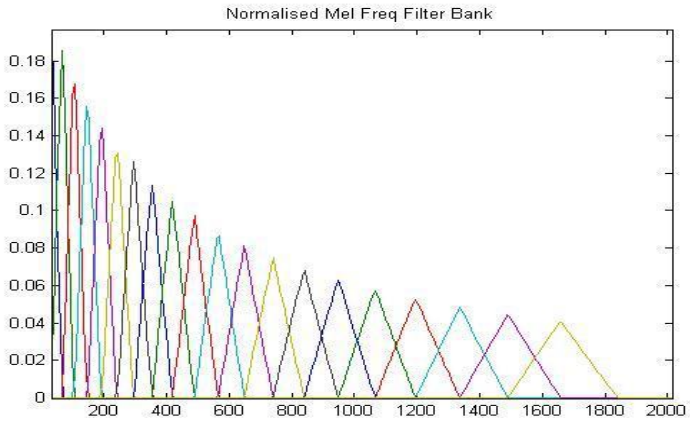
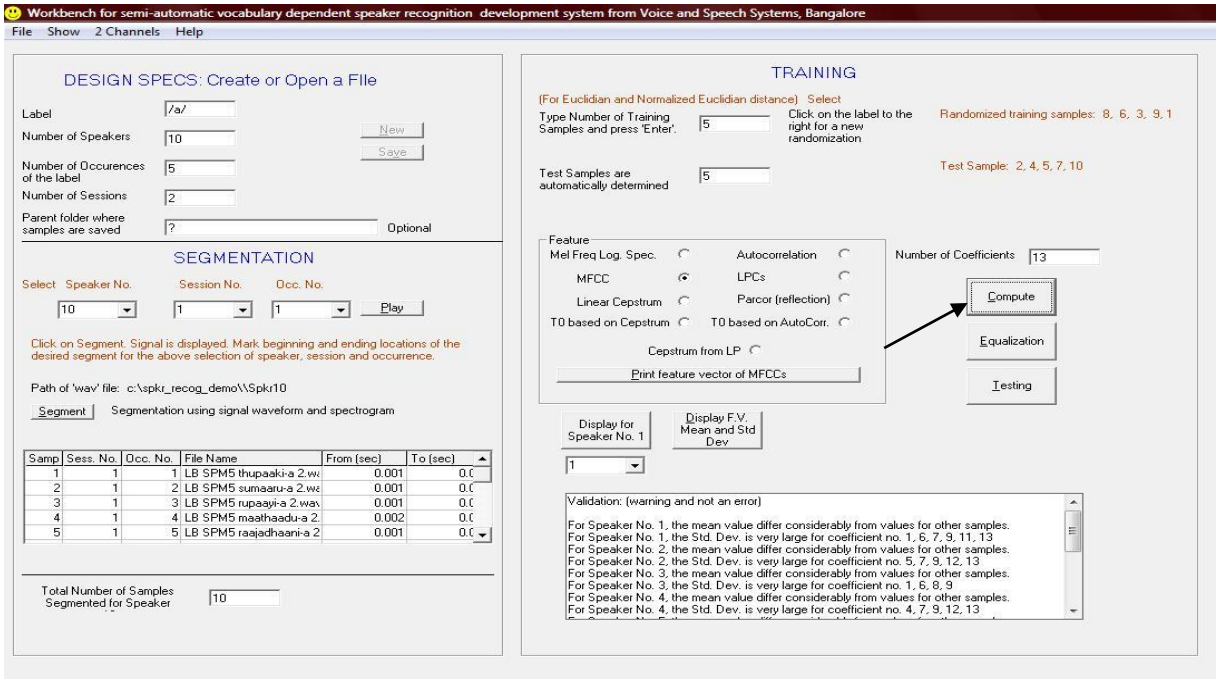
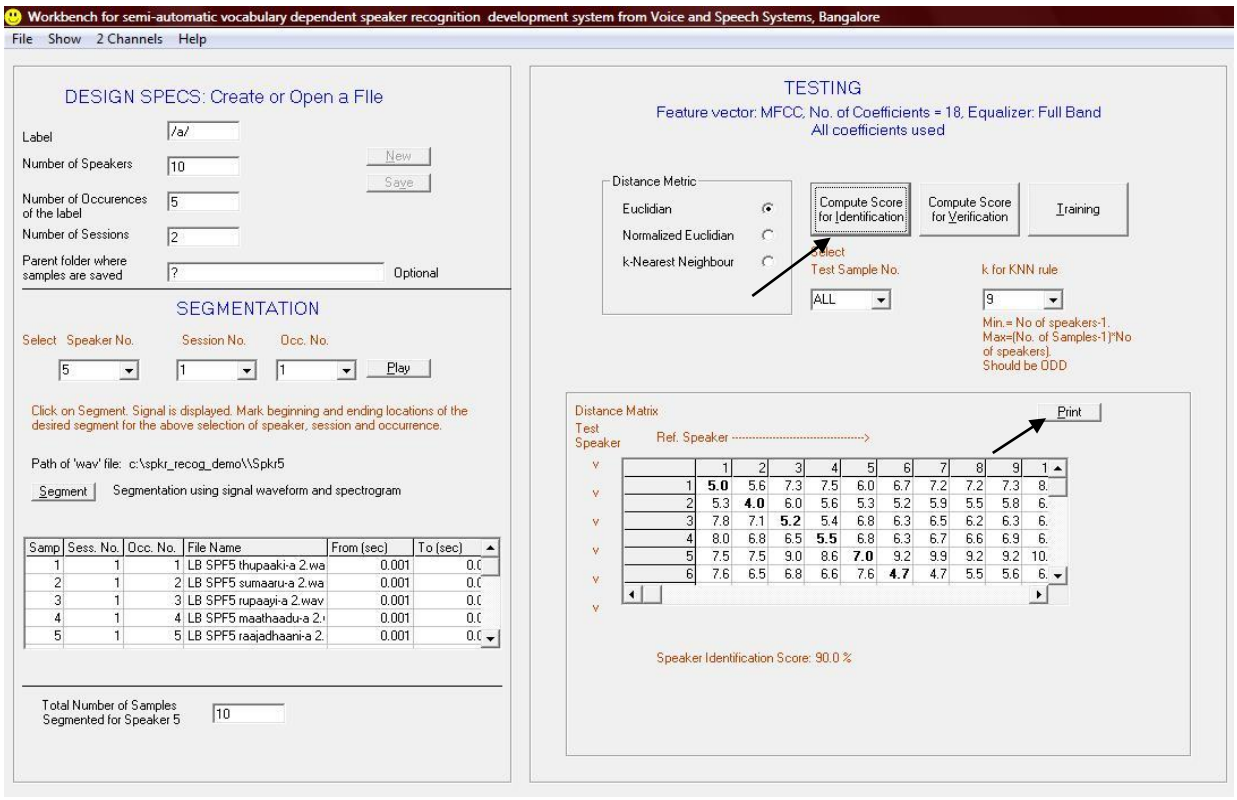


Figure 3.18: Mel frequency filter bank with normalization.

After selecting 13 MFCC, “compute” was clicked as shown in Figure 3.19 (a). On clicking this option the system used to check all the samples and compare them grossly and give a qualitative analysis of each speaker. Following this, the “testing” button was clicked on. This will open a window in which “compute a score for identification” was clicked as shown in Figure 3.19 (b). This gave the diagonal matrix in the lower half of the window and a final percentage for correct speaker identification.



(a)



(b)

Figure 3.19 (a) and (b): Analysis window of SSL Workbench showing diagonal matrix and speaker identification score.

This data was stored by pressing the “print” option and was saved as a .text file as illustrated in Figure 3.20. Thus the data was stored and the same procedure was repeated at least 30 times. Repetitions were done by randomizing the testing and training samples and the speaker identification thresholds were noted for the highest score and the lowest score.

Pankaja results-a LB-4.txt - Notepad

File Edit Format View Help

****

25-01-2016 10:20:57

---------- FEATURE USED MFCC ----------

---------- No. of Coefficients 13 ----------

All coefficients used

---------- Testing Samples Used: 1 3 4 5 7

---------- Euclidian Distance ----------

	1	2	3	4	5	6	7	8	9	10
1	2.705	4.881	8.257	7.676	4.590	5.916	6.236	6.863	7.587	8.256
2	3.790	2.469	5.655	5.796	3.960	3.800	4.876	4.664	5.204	6.673
3	7.439	5.268	2.320	4.135	6.447	4.749	6.236	4.937	5.001	7.428
4	8.170	6.491	4.237	3.421	7.312	5.700	6.723	6.443	7.059	7.692
5	5.835	5.174	6.797	6.601	4.560	6.246	7.267	6.614	7.027	8.496
6	5.977	4.551	5.117	5.282	6.148	3.279	4.286	3.756	5.139	6.679
7	5.869	4.836	6.324	5.921	6.810	3.626	2.941	4.233	5.456	5.925
8	8.085	6.404	6.202	6.646	7.396	5.424	5.910	4.475	5.612	8.428
9	7.429	5.888	5.829	6.752	7.011	5.586	6.097	5.567	3.204	8.230
10	8.056	6.066	6.879	6.094	8.518	5.913	5.762	7.338	8.217	3.799

---------- False Identifications ----------

----------Speaker Identification Score----------

100.00%

----------Speaker Verification Score----------

100.00%

----------Threshsold----------

7.92

****

Figure 3.20: Results for pilot study depicted in .text file.

Euclidian Distance for the lab and traffic conditions derived MFCC was extracted. The Euclidean distance between point’s 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 software uses the above-mentioned equation as one among many other equations to extract the Euclidian distance between 13 MFCCs for within and between participants. Participants having the least Euclidian distance were considered to be the same speakers. If the distance between the unknown and the corresponding known speaker is less, the identification was considered correct. If the distance between the unknown and the

corresponding known speaker is more, then the speaker is considered to be falsely identified as another speaker. The percent correct identification was calculated using the following formula:

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

In this study, the speech samples were both contemporary (lab v/s lab, traffic v/s traffic-BNR and traffic v/s traffic-ANR) and non-contemporary [lab v/s traffic (BNR), lab v/s traffic (ANR) and traffic (BNR) v/s traffic (ANR)]. Closed set speaker identification tasks were performed, in which the examiner was aware that the 'unknown speaker' is one among the 'known' speakers.

CHAPTER IV

RESULTS

The present study aimed to investigate the effect of noise and noise reduction techniques on speaker identification using MFCC on the long vowels in the Kannada language. Percentages of correct speaker identification were calculated for all the five categories [Lab versus Lab, Traffic versus Traffic (before noise reduction), Traffic versus Traffic (after noise reduction), Lab versus Traffic (before noise reduction), Lab versus Traffic (after noise reduction) and Traffic (before noise reduction) versus Traffic (after noise reduction)]. The results of these comparisons are explained under the following conditions:

1. Lab condition.
2. Traffic condition before noise reduction technique
3. Traffic condition after noise reduction technique
4. Lab recording versus traffic recording before noise reduction technique.
5. Lab recording versus traffic recording after noise reduction technique.
6. Traffic condition before noise reduction technique versus Traffic condition after noise reduction technique

The Euclidean distance of the samples for reference and test samples of each speaker were averaged separately by the workbench software. The test samples were taken along the column and the reference average was taken along the row. This was then tabulated as a distance matrix comparing all the speakers. The one with the minimum distance from the reference was identified as a test speaker. A distance matrix was computed by the software, for different combinations of test and reference speakers chosen. The green color in the table (distance matrix) indicates the correct identification of the speaker sample as belonging to the same speaker as the reference sample. The red color in the table (distance matrix) indicates the wrong identification of the test sample as belonging to a different speaker. Following this, the correct percentage of speaker identification scores was obtained and the same was randomized for 30 times to obtain the highest correct percentage of speaker identification (HPI).

Following this, the descriptive statistical analysis was also carried out where mean and standard deviation (SD) was calculated and also 95% Confidence Interval for Mean was calculated which gave lower and upper bound for all thirty randomized trials.

Condition I: Comparison of MFCCs of speakers' lab recording verses lab recording of vowel /a:/, /i:/ and /u:/

In this condition, contemporary speech samples were used where the lab recording (test sample) was compared with lab recording (reference sample). Here the results revealed that the highest percent correct identification (HPI) for vowel /a:/, /i:/ and /u:/ was noted to be 100% respectively. The lowest percent correct identification (LPI) for vowel /a: /, /i: / and /u: / was noted to be 70%, 68.33% and 45% respectively. On an average of 30 times of randomization, the percent correct speaker identification score for the vowel /a: /, /i: / and /u: / was 88.38% (SD: 9.34), 87.61% (SD: 10.7) and 77.11% (SD: 14.89) respectively. This indicates /a: / to be better followed by /i: / and /u: /. Table 4.1 depicts descriptive data of speaker identification scores obtained for all 30 randomized trials for vowels. For example, the trail with the test sample (2, 4, 6, 8, 10) (2, 4, 6, 8, 10) (2, 4, 5, 8, 9) showing the highest percent speaker identification score with reference to distance matrix with Euclidian Distance for the vowel /a:/. /i:/ and /u:/ is shown in Table 1, 2, 3 of Appendix D respectively. The green color in the tables indicates the correct identification of the speaker sample as belonging to the same speaker as the reference sample whereas the red color indicates the wrong identification of the test sample as belonging to a different reference speaker. The 95% Confidence Interval for Mean was also calculated using descriptive statistics. The lower and upper bound for vowels /a:/, /i:/ and /u:/ are 85.11%-91.66%, 83.84%-91.37% and 71.54%-82.67% respectively which is depicted in Figure 4.1. From the figure, it can be observed that for the vowel /a:/ and /i:/ the difference between the lower and upper bound is smaller (6.55 & 7.53) in comparison with the vowel /u:/ which is wider (11.13). Thus, the interpretation for this condition with reference to the percent correct speaker identification score is more consistent when the difference between the lower and upper bound is minimal compared to the wider difference.

Table 4.1: Speaker identification of vowels in lab condition

Lab Condition v/s Lab Condition				
No. of Randomization	**Test samples from randomization**	**Percentage of speaker identification score**		
		/a:/	**/i:/**	**/u:/**
1	2,4,5,7,1	71.67	68.33	45
2	2,3,6,7,10	80	78.33	45
3	2,3,5,7,9	91.67	91.67	81.67
4	**2,4,6,8,10**	**100**	**100**	73.33
5	1,3,4,5,7	96.67	**100**	98.33
6	**2,4,5,8,9**	86.67	85	**100**
7	2,5,7,8,10	70	71.67	71.67
8	1,3,6,9,10	93.33	90	73.33
9	3,4,7,8,9	90	88.33	80
10	1,2,5,6,8	81.67	81.67	78.33
11	2,3,4,9,10	88.33	86.67	76.67
12	2,3,4,7,8	93.33	86.67	76.67
13	1,2,8,9,10	95	70	70
14	2,6,7,8,9	98.33	**100**	68.33
15	1,4,8,9,10	88.33	85	98.33
16	2,3,4,6,10	73.33	76.67	85
17	3,5,7,8,10	80	81.67	45
18	3,4,5,8,10	91.67	**100**	81.67
19	1,3,6,9,10	**100**	83.33	73.33
20	2,3,5,7,9	96.67	98.33	98.33
21	3,7,8,9,10	86.67	90	**100**
22	4,6,7,8,9	70	76.67	71.67
23	2,4,5,6,9	93.33	93.33	73.33
24	2,6,7,8,10	90	**100**	80
25	6,7,8,9,10	81.67	98.33	78.33
26	2,3,5,7,9	88.33	75	76.67
27	2,3,6,8,9	93.33	80	76.67
28	3,6,7,8,9	95	93.33	70
29	1,2,3,5,8	98.33	**100**	68.33
30	1,3,6,7,9	88.33	98.33	98.33
Average		**88.34**	**87.61**	**77.11**
SD		**9.34**	**10.07**	**14.89**

*Note** SD= Standard deviation

Depiction of lower and upper boundary correct percent speaker identification for lab condition

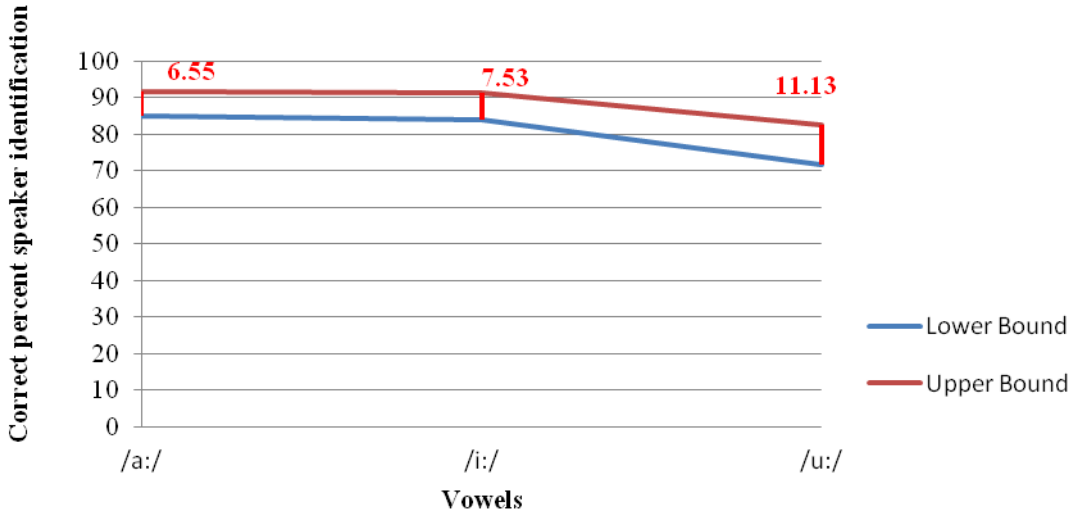


Figure 4.1: 95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for lab condition

Condition II: Comparison of MFCCs of speakers' traffic recording (before noise reduction technique) verses traffic recording (before noise reduction technique) of vowel /a:/, /i:/ and /u:/

For traffic recording, contemporary speech samples were used where traffic (test sample) condition was compared with traffic (reference sample) condition before noise reduction (BNR). Here the results revealed that the highest percent correct identification (HPI) for vowel /a:/, /i:/ and /u:/ was noted to be 100%, 100% and 95% respectively. The lowest percent correct identification (LPI) for vowel /a:/, /i:/ and /u:/ was noted to be 75%, 61.67% and 25% respectively. On an average of 30 times of randomization the percent correct speaker identification score for the vowel /a:/, /i:/ and /u:/ was 87.22% (SD: 8.28), 81.99% (SD: 12.14) and 66.77% (SD: 15.14) respectively. This indicates /a:/ to be better followed by /i:/ and /u:/. Table 4.2 depicts descriptive data of speaker identification scores obtained for all 30 randomized trials for vowels. For example, the trail with the test sample (2, 6, 7, 8, 9) (1, 3, 4, 5, 7) (1, 3, 6, 9, 10) showing the highest percent speaker identification score with reference to distance matrix with Euclidian Distance for the vowel /a:/. /i:/ and /u:/ is shown in Table 4, 5, 6 of Appendix D respectively. The green color in the tables indicates the correct identification of the speaker sample as belonging to the same speaker as the reference sample whereas the red color indicates the wrong identification of the test sample as belonging to a different reference speaker. The 95% Confidence Interval for Mean was also calculated using descriptive statistics. The lower and upper bound for vowels /a:/, /i:/ and /u:/ are 84.12%-90.31%, 77.46%-86.53% and 61.12%-72.43% respectively which is depicted in Figure 4.2. From the figure, it can be observed that for the vowel /a:/ and /i:/ the difference between the lower and upper bound is smaller (6.19 & 9.07) in comparison with the vowel /u:/ which is wider (11.31). Thus, the interpretation for this condition with reference to the percent correct speaker identification score is more consistent when the difference between the lower and upper bound is minimal compared to the wider difference.

Table 4.2: Speaker identification of vowels in traffic condition (BNR)

Traffic (BNR) Condition v/s Traffic (BNR) Condition				
No. of Randomization	**Test samples from randomization**	**Percentage of speaker identification score**		
		/a:/	**/i:/**	**/u:/**
1	2,4,5,7,1	76.67	63.33	25
2	2,3,6,7,10	93.33	70	71.67
3	2,3,5,7,9	88.33	80	65
4	2,4,6,8,10	98.33	98.33	85
5	**1,3,4,5,7**	98.33	**100**	90
6	2,4,5,8,9	76.67	83.33	58.33
7	2,5,7,8,10	85	61.67	51.67
8	1,3,6,9,10	91.67	85	70
9	3,4,7,8,9	93.33	88.33	73.33
10	1,2,5,6,8	83.33	81.67	73.33
11	2,3,4,9,10	80	85	68.33
12	2,3,4,7,8	75	81.67	53.33
13	1,2,8,9,10	85	63.33	56.67
14	**2,6,7,8,9**	**100**	68.33	86.67
15	1,4,8,9,10	88.33	96.67	70
16	2,3,4,6,10	78.33	83.33	68.33
17	3,5,7,8,10	76.67	91.67	60
18	3,4,5,8,10	93.33	**100**	61.67
19	**1,3,6,9,10**	88.33	61.67	**95**
20	2,3,5,7,9	98.33	95	53.33
21	3,7,8,9,10	98.33	83.33	46.67
22	4,6,7,8,9	76.67	88.33	73.33
23	2,4,5,6,9	85	66.67	63.33
24	2,6,7,8,10	91.67	96.67	63.33
25	6,7,8,9,10	93.33	81.67	86.67
26	2,3,5,7,9	83.33	98.33	48.33
27	2,3,6,8,9	80	78.33	68.33
28	3,6,7,8,9	75	70	91.67
29	1,2,3,5,8	85	75	63.33
30	1,3,6,7,9	**100**	83.33	61.67
Average		**87.22**	**81.99**	**66.77**
SD		**8.28**	**12.14**	**15.14**

*Note** SD= Standard deviation, BNR= Before noise reduction

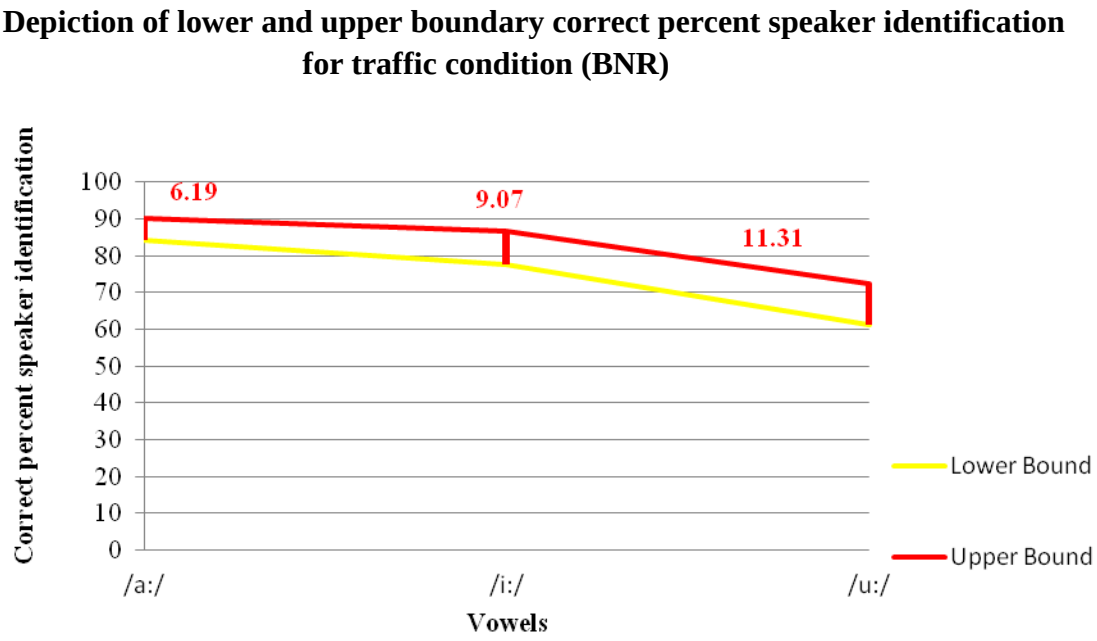


Figure 4.2: 95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for traffic condition (BNR)

Comparison on observation among the average percent correct speaker identification score for lab versus traffic recording condition the differences was seen majorly for the vowel /u:/ when compared to /a:/ and /i:/. The same is represented graphically in Figure (4.3).

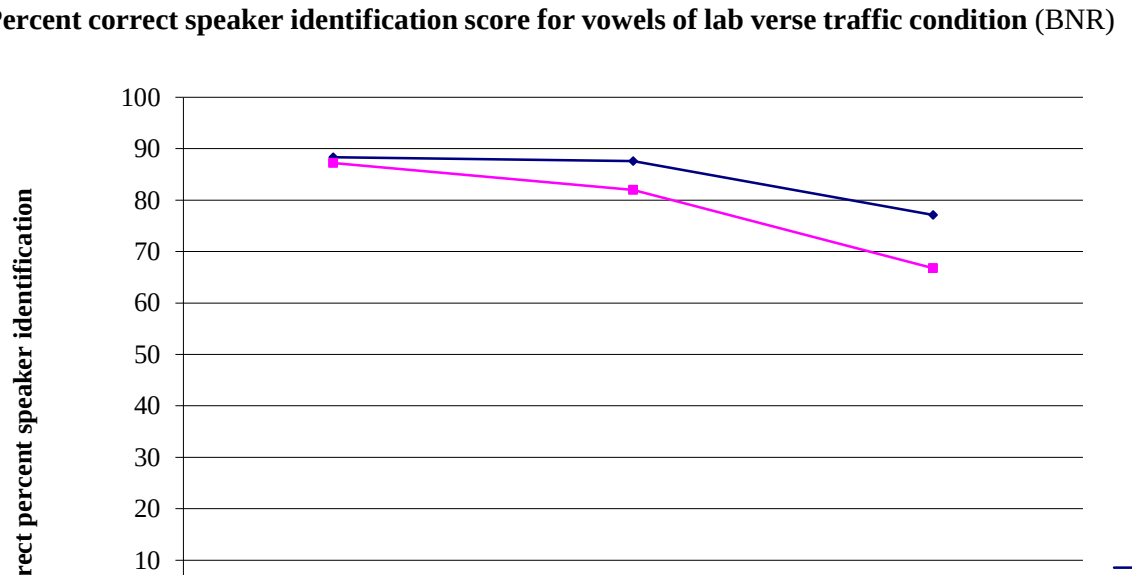


Figure 4.3: Percent correct speaker identification score for vowels of lab verse traffic condition (BNR)

Condition III: Comparison of MFCCs of speakers' traffic recording (ANR) verses traffic recording (ANR) of vowel /a:/, /i:/ and /u:/

In this condition, contemporary speech samples were used where traffic recording (test sample) after the application of noise reduction technique was compared with traffic recording (reference sample) after the application of noise reduction technique. Where these samples had undergone a noise reduction scheme in sound cleaner software. Here the results revealed that the highest percent correct identification (HPI) for vowel /a:/, /i:/ and /u:/ was noted to be 98.33%, 100% and 85% respectively. The lowest percent correct identification (LPI) for vowel /a: /, /i: / and /u: / was noted to be 53.33%, 55% and 20% respectively. On an average of 30 times of randomization, the percent correct speaker identification score for the vowel /a:/, /i:/ and /u:/ was 79.38% (SD: 12.10), 76.72% (13.95) and 53.22% (14.50) respectively. This indicates /a:/ to be better followed by /i:/ and /u:/. Table 4.3 depicts descriptive data for speaker identification scores obtained for all 30 randomized trials for vowels. For example, the trail with the test sample (3, 5, 7, 8, 10) (2, 4, 6, 8, 10) (1, 3, 4, 5, 7) showing the highest percent speaker identification score with reference to distance matrix with Euclidian Distance for the vowel /a:/. /i:/ and /u:/ is shown in Table 7, 8, 9 of Appendix D respectively. The green color in the tables indicates the correct identification of the speaker sample as belonging to the same speaker as the reference sample whereas the red color indicates the wrong identification of the test sample as belonging to a different reference speaker. The 95% Confidence Interval for Mean was also calculated using descriptive statistics. The lower and upper bound for vowels /a:/, /i:/ and /u:/ are 74.86%-83.90%, 71.51%-81.93% and 47.80%-58.63% respectively which is depicted in Figure 4.4. From the figure, it can be observed that for the vowel /a:/ the difference between the lower and upper bound is smaller (9.04) in comparison with the vowel /i:/ and /u:/ which is wider (10.42 &10.83). Thus, the interpretation for this condition with reference to the percent correct speaker identification score is more consistent when the difference between the lower and upper bound is minimal compared to the wider difference.

Table 4.3: Speaker identification of vowels in Traffic condition (ANR)

Traffic (ANR) condition v/s Traffic (ANR) Condition				
No. of Randomization	**Test samples from randomization**	**Percentage of speaker identification score**		
		/a:/	**/i:/**	**/u:/**
1	2,4,5,7,1	60	55	30
2	2,3,6,7,10	75	63.33	53.33
3	2,3,5,7,9	85	81.67	53.33
4	**2,4,6,8,10**	96.67	**100**	76.67
5	**1,3,4,5,7**	95	96.67	**85**
6	2,4,5,8,9	73.33	71.67	50
7	2,5,7,8,10	65	60	38.33
8	1,3,6,9,10	80	78.33	58.33
9	3,4,7,8,9	85	66.67	45
10	1,2,5,6,8	80	81.67	48.33
11	2,3,4,9,10	66.67	83.33	43.33
12	2,3,4,7,8	70	85	40
13	1,2,8,9,10	78.33	66.67	80
14	2,6,7,8,9	95	**100**	55
15	1,4,8,9,10	80	78.33	53.33
16	2,3,4,6,10	76.67	71.67	45
17	**3,5,7,8,10**	**98.33**	71.67	83.33
18	3,4,5,8,10	78.33	100	43.33
19	1,3,6,9,10	93.33	65	33.33
20	2,3,5,7,9	78.33	71.67	53.33
21	3,7,8,9,10	83.33	55	53.33
22	4,6,7,8,9	71.67	63.33	56.67
23	2,4,5,6,9	96.67	81.67	63.33
24	2,6,7,8,10	81.67	**100**	36.67
25	6,7,8,9,10	53.33	96.67	48.33
26	2,3,5,7,9	53.33	71.67	75
27	2,3,6,8,9	76.67	60	53.33
28	3,6,7,8,9	80	78.33	48.33
29	1,2,3,5,8	93.33	65	56.67
30	1,3,6,7,9	81.67	81.67	36.67
Average		**79.38**	**76.72**	**53.22**
SD		**12.10**	**13.95**	**14.50**

*Note** SD= Standard deviation, ANR= After noise reduction

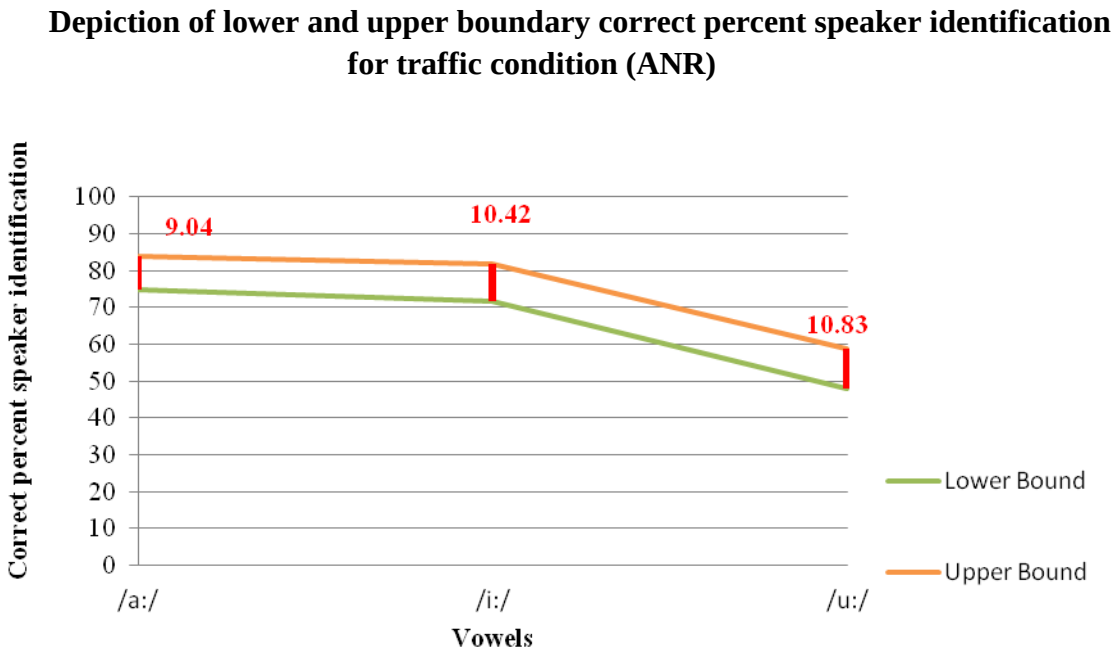


Figure 4.4: 95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for traffic condition (ANR)

Condition IV: Comparison of MFCCs of speakers' lab recording verses traffic recording (BNR) of vowel /a:/, /i:/ and /u:/

In this condition, non-contemporary speech samples were used where lab (reference sample) recording was compared with traffic recording (test sample) before the application of the noise reduction technique. Here, the lab sample was absolutely speech and no noise, whereas the traffic samples contain some amount of traffic noise embedded in it during analysis. Here the results revealed that the highest percent correct identification (HPI) for vowel /a:/, /i:/ and /u:/ was noted to be 83.33%, 86.67 and 66.67 respectively. The lowest percent correct identification (LPI) for vowel /a: /, /i: / and /u: / was noted to be 36.67%, 21.67% and 20% respectively. On an average of 30 times of randomization, the percent correct speaker identification score for the vowel /a:/, /i:/ and /u:/ was 65.77% (SD: 14.05), 62.27% (SD: 16.37) and 42.61% (SD: 14.34) respectively. This indicates /a:/ to be better followed by /i:/ and /u:/. Table 4.4 depicts descriptive data of speaker identification scores obtained for all 30 randomized trials for vowels. For example the trail with the test sample (2, 4, 6, 8, 10) (2, 6, 7, 8, 9) (3, 7, 8, 9, 10) with the highest percent speaker identification score with reference to distance matrix with Euclidian Distance for the vowel /a:/. /i:/ and /u:/ is shown in Table 10, 11, 12 of Appendix D respectively. The green color in the tables indicates the correct identification of the speaker sample as belonging to the same speaker as the reference sample whereas the red color indicates the wrong identification of the test sample as belonging to a different reference speaker. The 95% Confidence Interval for Mean was also calculated using descriptive statistics. The lower and upper bound for vowels /a:/, /i:/ and /u:/ are 60.52%-71.02%, 56.16%-68.39% and 37.25%-47.96% respectively which is depicted in Figure 4.5. From the figure, it can be observed that for the vowel /a:/ and /u:/ the difference between the lower and upper bound is smaller (10.5 & 10.71) in comparison with the vowel /i:/ which is wider (12.23). Thus, the interpretation for this condition with reference to the percent correct speaker identification score is more consistent when the difference between the lower and upper bound is minimal compared to the wider difference.

Table 4.4: Speaker identification of vowels in Lab condition v/s Traffic (BNR) condition

Lab Condition v/s Traffic (BNR) Condition				
No. of Randomization	**Test samples from randomization**	**Percentage of speaker identification score**		
		/a:/	**/i:/**	**/u:/**
1	2,4,5,7,1	71.67	55	20
2	2,3,6,7,10	55	66.67	53.33
3	2,3,5,7,9	71.67	61.67	56.67
4	**2,4,6,8,10**	**83.33**	81.67	63.33
5	1,3,4,5,7	43.33	55	25
6	2,4,5,8,9	63.33	78.33	53.33
7	2,5,7,8,10	58.33	65	48.33
8	1,3,6,9,10	73.33	65	53.33
9	3,4,7,8,9	80	73.33	38.33
10	1,2,5,6,8	68.33	71.67	50
11	2,3,4,9,10	80	71.67	51.67
12	2,3,4,7,8	66.67	78.33	48.33
13	1,2,8,9,10	73.33	58.33	60
14	**2,6,7,8,9**	76.67	**86.67**	55
15	1,4,8,9,10	36.67	41.67	26.67
16	2,3,4,6,10	68.33	43.33	13.33
17	3,5,7,8,10	80	71.67	25
18	3,4,5,8,10	66.67	**86.67**	46.67
19	1,3,6,9,10	73.33	63.33	38.33
20	2,3,5,7,9	76.67	78.33	21.67
21	**3,7,8,9,10**	36.67	73.33	**66.67**
22	4,6,7,8,9	68.33	61.67	38.33
23	2,4,5,6,9	71.67	41.67	31.67
24	2,6,7,8,10	81.67	33.33	31.67
25	6,7,8,9,10	61.67	73.33	41.67
26	2,3,5,7,9	53.33	38.33	31.67
27	2,3,6,8,9	80	21.67	51.67
28	3,6,7,8,9	71.67	61.67	55
29	1,2,3,5,8	40	66.67	26.67
30	1,3,6,7,9	41.67	43.33	55
Average		**65.77**	**62.27**	**42.61**
SD		**14.05**	**16.37**	**14.34**

*Note** SD= Standard deviation, BNR= Before noise reduction

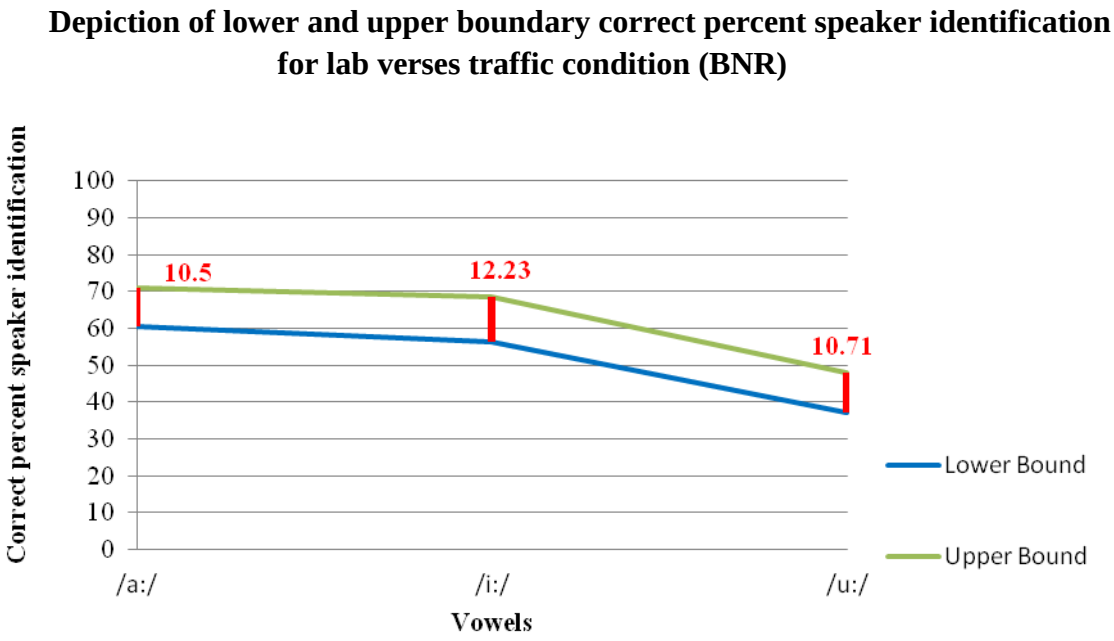


Figure 4.5: 95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for lab verses traffic condition (BNR)

Condition V: Comparison of MFCCs of speakers' lab recording verses traffic recording (ANR) of vowel /a:/, /i:/ and /u:/

In this condition, non-contemporary speech samples were used where lab recording (reference sample) was compared with traffic recording (test sample) after the application of noise reduction technique. Here, the lab sample was absolutely speech and no noise, whereas the traffic samples containing some amount of traffic noise embedding in it was removed with the sound cleaner software during analysis. Here the results revealed that the highest percent correct identification (HPI) for vowel /a:/, /i:/ and /u:/ was noted to be 86.67%, 83.33% and 63.33% respectively. The lowest percent correct identification (LPI) for vowel /a: /, /i: / and /u: / was noted to be 16.67%, 33.33% and 16.67% respectively. On an average of 30 times of randomization the percent correct speaker identification score for the vowel /a:/, /i:/ and /u:/ was 58.61% (SD: 17.11), 58.94% (SD: 15.11) and 38.11% (SD: 10.67) respectively. This indicates /a:/ to be better followed by /i:/ and /u:/. Table 4.5 depicts descriptive data for speaker identification scores obtained for all 30 randomized trials for vowels. For example the trail with the test sample (3, 4, 5, 8, 10) (1, 4, 8, 9, 10) (1, 2, 8, 9, 10) with the highest percent speaker identification score with reference to distance matrix with Euclidian Distance for the vowel /a:/. /i:/ and /u:/ is shown in Table 13, 14, 15 of Appendix D respectively. The green color in the tables indicates the correct identification of the speaker sample as belonging to the same speaker as the reference sample whereas the red color indicates the wrong identification of the test sample as belonging to a different reference speaker. The 95% Confidence Interval for Mean was also calculated using descriptive statistics. The lower and upper bound for vowels /a:/, /i:/ and /u:/ are 52.22%-65%, 53.30%-64.58% and 34.12%-42.09% respectively which is depicted in Figure 4.6. From the figure, it can be observed that for the vowel /i:/ and /u:/ the difference between the lower and upper bound is smaller (11.28 & 7.97) in comparison with the vowel /a:/ which is wider (12.78). Thus, the interpretation for this condition with reference to the percent correct speaker identification score is more consistent when the difference between the lower and upper bound is minimal compared to the wider difference.

Table 4.5: Speaker identification of vowels in lab condition v/s Traffic (ANR) condition

Lab condition v/s Traffic Condition (ANR)				
No. of Randomization	**Test samples from randomization**	**Percentage of speaker identification score**		
		/a:/	**/i:/**	**/u:/**
1	2,4,5,7,1	58.33	48.33	36.67
2	2,3,6,7,10	68.33	53.33	25
3	2,3,5,7,9	51.67	65	26.67
4	2,4,6,8,10	75	33.33	23.33
5	1,3,4,5,7	35	53.33	45
6	2,4,5,8,9	56.67	68.33	36.67
7	2,5,7,8,10	63.33	76.67	48.33
8	1,3,6,9,10	75	78.33	38.33
9	3,4,7,8,9	75	48.33	36.67
10	1,2,5,6,8	66.67	50	36.67
11	2,3,4,9,10	71.67	80	61.67
12	2,3,4,7,8	50	46.67	35
13	**1,2,8,9,10**	66.67	71.67	**63.33**
14	2,6,7,8,9	80	66.67	36.67
15	**1,4,8,9,10**	35	**83.33**	51.67
16	2,3,4,6,10	75	73.33	43.33
17	3,5,7,8,10	65	70	38.33
18	**3,4,5,8,10**	**86.67**	68.33	36.67
19	1,3,6,9,10	71.67	60	43.33
20	2,3,5,7,9	55	61.67	30
21	3,7,8,9,10	75	41.67	31.67
22	4,6,7,8,9	51.67	56.67	36.67
23	2,4,5,6,9	48.33	36.67	41.67
24	2,6,7,8,10	30	33.33	16.67
25	6,7,8,9,10	55	60	43.33
26	2,3,5,7,9	36.67	61.67	46.67
27	2,3,6,8,9	16.67	40	23.33
28	3,6,7,8,9	51.67	35	23.33
29	1,2,3,5,8	75	65	41.67
30	1,3,6,7,9	36.67	81.67	45
Average		**58.61**	**58.94**	**38.11**
SD		**17.11**	**15.11**	**10.67**

*Note** SD= Standard deviation, ANR= After noise reduction

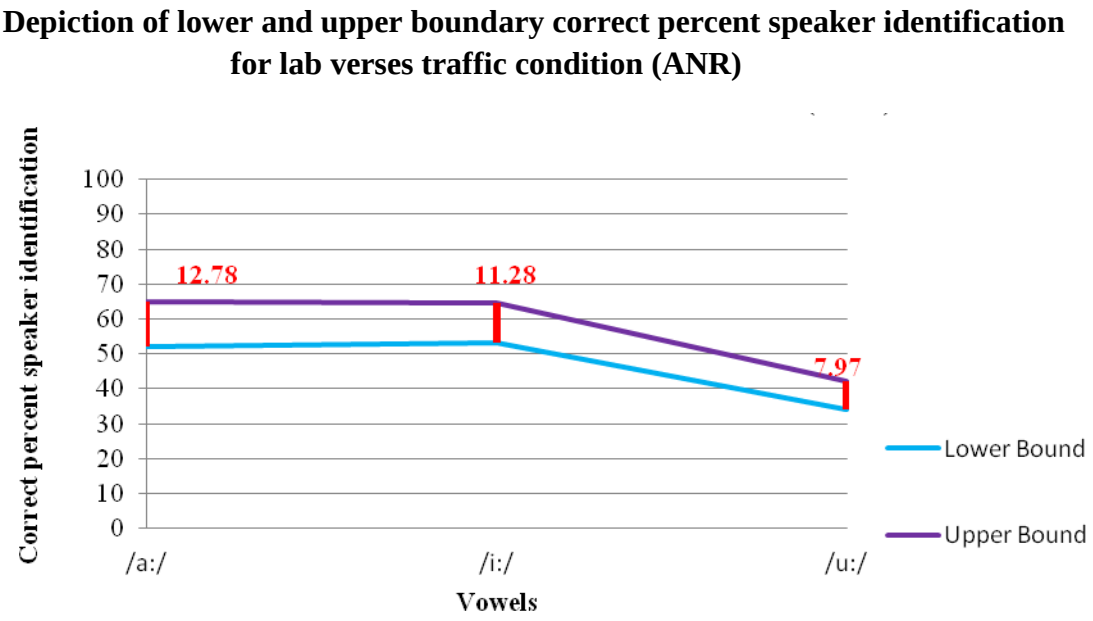


Figure 4.6: 95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for lab verses traffic condition (ANR)

Comparison on observation among the percent correct speaker identification score for conditions IV and V, there is a decrement in the percent correct speaker identification scores for all the vowels after the application of noise reduction technique. The same is represented graphically in Figure (4.7).

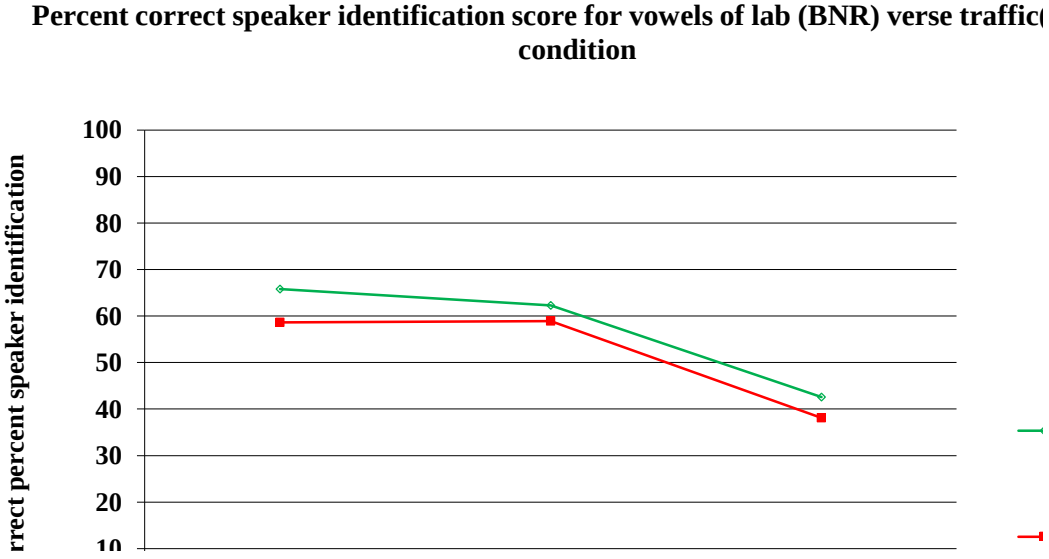


Figure 4.7: Percent correct speaker identification score for vowels of lab verse traffic condition

Condition VI: Comparison of MFCCs of speakers' traffic recording (BNR) verses traffic recording (ANR) of vowel /a:/, /i:/ and /u:/

In this condition, non-contemporary speech samples were used where traffic recording (reference sample) before the application of noise reduction technique was compared with traffic recording (test sample) after the application of noise reduction technique. Here, the traffic samples containing some amount of traffic noise embedding in it were compared with the traffic samples containing some amount of traffic noise embedding in it, and the same was removed with the sound cleaner software during analysis. Here the results revealed that the highest percent correct identification (HPI) for vowel /a:/, /i:/ and /u:/ was noted to be 96.67%, 95% and 68.33% respectively. The lowest percent correct identification (LPI) for vowel /a: /, /i: / and /u: / was noted to be 43.33%, 35% and 8.33% respectively. On an average of 30 times of randomization the percent correct speaker identification score for the vowel /a:/, /i:/ and /u:/ was 70.25% (SD: 13.78), 63.83% (SD: 14.93) and 35.72% (SD: 14.28) respectively. This indicates /a:/ to be better followed by /i:/ and /u:/. Table 4.6 depicts descriptive data for speaker identification scores obtained for all 30 randomized trials for vowels. For example the trail with the test sample (2, 4, 6, 8, 10) (2, 3, 4, 9, 10) (2, 3, 4, 9, 10) with the highest percent speaker identification score with reference to distance matrix with Euclidian Distance for the vowel /a:/. /i:/ and /u:/ is shown in Table 16, 17, 18 of Appendix D. The green color in the tables indicates the correct identification of speaker sample as belonging to the same speaker as the reference sample whereas red color indicates wrong identification of test sample as belonging to a different reference speaker. The 95% Confidence Interval for Mean was also calculated using descriptive statistics. The lower and upper bound for vowels /a:/, /i:/ and /u:/ are 64.90%-75.20%, 58.25%-69.40% and 30.39%-41.05% respectively which is depicted in Figure 4.8. From the figure, it can be observed that for the vowel /a:/ and /u:/ the difference between the lower and upper bound is smaller (10.3 & 10.66) in comparison with the vowel /i:/ which is wider (11.15). Thus, the interpretation for this condition with reference to the percent correct speaker identification score is more consistent when the difference between the lower and upper bound is minimal compared to the wider difference.

Table 4.6: Speaker identification of vowels in Traffic (BNR) condition v/s Traffic (ANR) condition

Traffic (BNR) condition v/s Traffic (ANR) condition				
No. of Randomization	**Test samples from randomization**	**Percentage of speaker identification score**		
		/a:/	**/i:/**	**/u:/**
1	2,4,5,7,1	56.67	53.33	26.67
2	2,3,6,7,10	75	43.33	13.33
3	2,3,5,7,9	73.33	68.33	26.67
4	**2,4,6,8,10**	**96.67**	41.67	8.33
5	1,3,4,5,7	48.33	70	41.67
6	2,4,5,8,9	68.33	70	53.33
7	2,5,7,8,10	76.67	70	40
8	1,3,6,9,10	81.67	70	41.67
9	3,4,7,8,9	85	53.33	26.67
10	1,2,5,6,8	75	51.67	36.67
11	**2,3,4,9,10**	71.67	**95**	**68.33**
12	2,3,4,7,8	60	53.33	30
13	1,2,8,9,10	71.67	75	60
14	2,6,7,8,9	91.67	76.67	36.67
15	1,4,8,9,10	50	88.33	45
16	2,3,4,6,10	76.67	65	46.67
17	3,5,7,8,10	71.67	68.33	31.67
18	3,4,5,8,10	85	70	36.67
19	1,3,6,9,10	78.33	56.67	33.33
20	2,3,5,7,9	43.33	66.67	36.67
21	3,7,8,9,10	85	68.33	26.67
22	4,6,7,8,9	73.33	63.33	28.33
23	2,4,5,6,9	75	51.67	10
24	2,6,7,8,10	53.33	35	36.67
25	6,7,8,9,10	71.67	56.67	43.33
26	2,3,5,7,9	48.33	81.67	30
27	2,3,6,8,9	51.67	55	10
28	3,6,7,8,9	73.33	35	46.67
29	1,2,3,5,8	78.33	75	53.33
30	1,3,6,7,9	55	86.67	46.67
Average		**70.25**	**63.83**	**35.72**
SD		**13.78**	**14.93**	**14.28**

*Note** SD= Standard deviation, BNR= Before noise reduction, ANR= After noise reduction

Depiction of lower and upper boundary correct percent speaker identification for traffic (BNR) verses traffic condition (ANR)

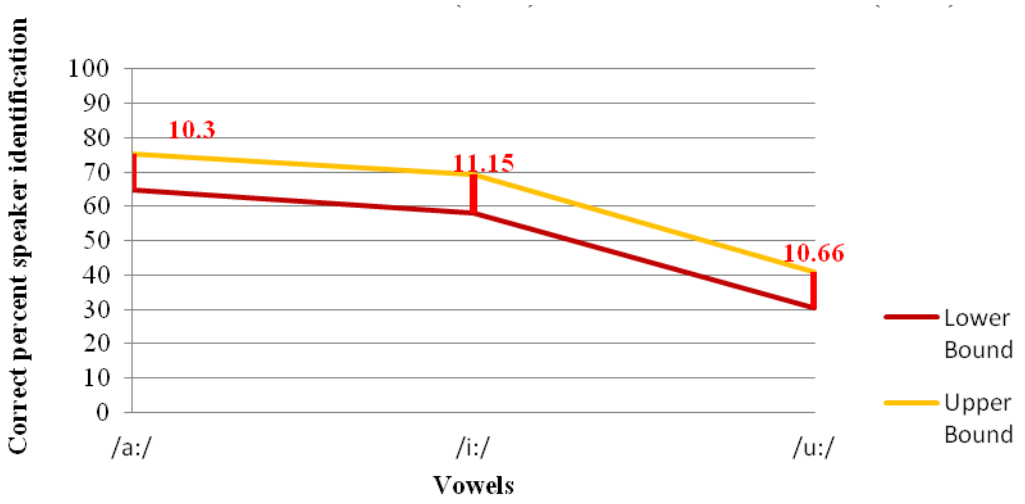


Figure 4.8: 95% Confidence Interval for Mean of /a:/, /i:/ and /u:/ vowels for traffic (BNR) verses traffic condition (ANR)

Summary of the results:

Average in terms of mean and standard deviation (SD) of the percentage of speaker identification for a condition I, II, III, IV, V, and VI are represented in Table 4.7 as a summary.

Table 4.7: Mean and standard deviation (SD) of the percent correct speaker identification for a condition I, II, III, IV, V & VI

Conditions	**Percent correct speaker identification scores**					
	/a:/		**/i:/**		**/u:/**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
I. Lab v/s Lab	**88.34**	9.34	87.61	10.07	77.11	14.89
II. Traffic (BNR) v/s Traffic (BNR)	**87.22**	8.28	81.99	12.14	66.77	15.14
III. Traffic (ANR) v/s Traffic (ANR)	**79.38**	12.10	76.72	13.95	53.22	14.50
IV. Lab v/s Traffic (BNR)	**65.77**	14.05	62.27	16.37	42.61	14.34
V. Lab v/s Traffic (ANR)	58.61	17.11	**58.94**	15.11	38.11	10.67
VI. Traffic (BNR) v/s Traffic (ANR)	**70.25**	13.78	63.83	14.93	35.72	14.28

Note: SD= Standard deviation, BNR= Before noise reduction, ANR= After noise reduction

The average percent correct speaker identification score for the vowel /a:/ was better in all conditions like **I**- Lab v/s Lab, **II**- Traffic (BNR) v/s Traffic (BNR), **III**- Traffic (ANR) v/s Traffic (ANR), **IV**- Lab v/s Traffic (BNR), **VI**- Traffic (BNR) v/s Traffic (ANR) and except **V**- Lab v/s Traffic (ANR) with the vowel /i:/ being better with very minimal difference (at decimal value) in comparison with the vowel /a:/ and /u:/.

The 95% Confidence Interval for Mean was calculated which gave lower and upper bound for all thirty randomized trials. The summary of this finding is represented in Table 4.8 for all the conditions (I, II, III, IV, V, and VI). With reference to the upper and lower bound value, the difference is minimal for the vowel /a:/ followed by /i:/ in conditions **I**, **II,** and **III** which indicates the percent correct speaker identification score is more consistent compared to vowel /u:/ which has a wider difference since these were all the contemporary samples. For conditions **IV** and **VI**, vowel /a:/ had minimal difference which indicates the percent correct speaker identification score is more consistent compared to /u:/ followed by /i:/. For condition **V** vowel /u:/ followed by /i:/ had minimal difference which indicates the percent correct speaker identification score is more consistent compared to /a:/. Therefore according to the statistical method also, the vowel /a:/ is better for speaker identification followed by /i:/ and then /u:/ for all the conditions except condition V, where the vowel /u:/ is better for speaker identification followed by /i:/ and /a:/. The details regarding these findings are discussed in the following section of the discussion.

Table 4.8: Difference value between the lower and upper boundary calculated for 95% confidence interval of the mean

Conditions	**The difference value between the upper and lower bound**		
	/a:/	**/i:/**	**/u:/**
I. Lab v/s Lab	**6.55**	7.53	11.13
II. Traffic (BNR) v/s Traffic (BNR)	**6.19**	9.07	11.31
III. Traffic (ANR) v/s Traffic (ANR)	**9.04**	10.42	10.83
IV. Lab v/s Traffic (BNR)	**10.5**	12.23	10.71
V. Lab v/s Traffic (ANR)	12.78	11.28	**7.97**
VI. Traffic (BNR) v/s Traffic (ANR)	**10.3**	11.15	10.66

Note: SD= Standard deviation, BNR= Before noise reduction, ANR= After noise reduction

CHAPTER V

DISCUSSION

The present study aimed to investigate the effect of noise and noise reduction techniques on speaker identification using MFCCs on the long vowels in the Kannada language. Results of the study revealed that for Condition I (*lab recording*), on an average of 30 randomized trails the percent correct speaker identification for vowel /a:/, /i:/ and /u:/ were 88.34%, 87.61% and 77.11% respectively. For the Condition II (*traffic recording - before noise reduction technique*) on an average of 30 randomized trials the percent correct speaker identification for vowel /a:/, /i:/ and /u:/ were 87.22%, 81.99% and 66.77% respectively. Subsequent to this was Condition III (*Traffic recording compared across traffic recording- after noise reduction technique*) on an average of 30 randomized trials the percent correct speaker identification for the vowel /a:/, /i:/ and /u:/ were 79.38 %, 76.72 %, and 53.22 % respectively. Then for Condition IV (*lab recording versus traffic recording- before noise reduction technique*) on an average of 30 randomized trials the percent correct speaker identification for vowel /a:/, /i:/ and /u:/ were 65.77%, 62.27 % and 42.61% respectively. Finally for Condition V (*lab recording compared across traffic recording- after noise reduction technique*) on an average of 30 randomized trials the percent correct speaker identification for vowel /a:/, /i:/ and /u:/ were 58.61%, 58.94% and 38.11% respectively. The last Condition VI (*traffic recording- before noise reduction versus traffic recording- after noise reduction)* on an average of 30 randomized trials the percent correct speaker identification for the vowel /a:/. /i:/ and /u:/ were 70.25%, 63.83% and 35.72% respectively.

As a summary (Table 4.7) the average percent correct speaker identification score for the vowel /a:/ was better in Condition I, II, III, IV, and VI except for Condition V with the vowel /i:/ being better with very minimal difference (at decimal value) in comparison with the vowel /a:/ and /u:/. According to the statistical method also (Table 4.8), the vowel /a:/ is better for speaker identification followed by /i:/ and then /u:/ for all the conditions except Condition V, where the vowel /u:/ is better for speaker identification followed by /i:/ and /a:/. When the comparison was made between the same recording fields irrespective of the samples is not subjected and/or subjected to noise reduction vowel /a:/ had better speaker identification score followed by vowel /i:/ and /u:/.

Any speech signal, for example, consists of various parameters. This can be broadly divided into three groups; the first is with reference to the quantity. For example, the duration of the voice sample. The second is with reference to the quality, which includes a signal to noise ratio, frequency range, clipping, etc and the third group is with reference to comparability, this includes the speaker being in the same emotional state, etc. These are the general variables that can be considered in studies related to speech signals.

With this background, therefore the foremost point to discuss with reference to the results of the present study could be the **vowels** contributing significantly in speaker identification with reference to their exceptional **acoustical characteristics when compared to consonants.** Several studies using automatic and semiautomatic methods of speaker identification have proved vowels to be effective speech sound for speaker identification compares to consonants. Since the vowels are the speech sounds produced by voiced excitation of the open vocal tract. In the production of a vowel, the vocal tract normally maintains a relatively stable shape and offers minimal obstruction to the airflow. The energy produced can be radiated through the mouth or nasal cavity without audible friction or stoppage. Thus, acoustically vowels are characterized by formant pattern, spectrum, duration, and fundamental frequency. However, in the present study, these factors have contributed differently amongst the three vowels, and therefore results being represented differently. The present study focused on long vowels such as /a:/, /i:/ and /u:/ in Kannada language and found vowel /a:/ to be better for speaker identification compared to /i:/ and /u:/. In conditions such as I -Lab v/s Lab, II- Traffic (BNR) v/s Traffic (BNR), III- Traffic (ANR) v/s Traffic (ANR), IV- Lab v/s Traffic (BNR), VI- Traffic (BNR) v/s Traffic (ANR) and V- Lab v/s Traffic (ANR) vowel /a:/ scored better compared to /i:/ and /u:/.

Similar results were found in the study done by Arjun (2015), where the author attempted to obtain a benchmark for speaker identification using Kannada vowels preceding nasal continuants and found vowel /a:/ preceding both nasals /m/ and /n/ to be best for speaker identification compared to other vowels (/i:/ and /u:/). Aswathy (2016) studied the effect of native versus non-native languages in speaker identification using MFCCs. The study concluded that vowel /a:/ symbolized as a better cue for speaker identification irrespective of the language used when compared to /i:/ and /u:/ vowel. These studies used only a target word list with preceding and following specific consonants. Whereas in the present study, sentences were used and within a sentence, a target word was selected and the vowel in the medial position of the target words was only considered for analysis. The factor related to preceding and following any specific consonant is not considered in the present study and this could be considered as one of the contributing factors.

The efficacy of a speaker verification system was studied using MFCCs by Chandrika (2010) and found vowel /i:/ to be the better cue for speaker identification. The correct speaker identification score obtained was 90-95% for vowel /i:/ and thus the overall performance of vowel /i:/ was better compared to /a:/ and /u:/. From these studies, it is observed that the vowel /a:/ and /i:/ resulting in a better cue for speaker identification in different conditions.

When the text-independent data was used and speaker identification was checked in native Kannada speakers by Sreevidya (2010). The results obtained was high percent correct speaker identification for vowel /u:/ (70%) in comparison with vowel /i:/ and /a:/ (50% each). In addition to this, Ramya (2011) also reported higher percent speaker identification (96.66%) for vowel /u:/ in comparison with vowel /a: / 93.33 %, and /i: / 93.33%. Here, the speaker identification was checked using MFCCs under electronic vocal disguise condition in females.

Thus, with more specific to the automatic method of speaker identification the second contributing factor could be the **parameter MFCC**. Several authors have found MFCCs to be the best parameter for speaker identification (Pruthi & Epsy-Wilson, 2007; Singh & Rajan, 2011; Sukor & Syafiq, 2012). However in the present study also, with the application of noise reduction technique to the field recording samples, it is found that the MFCC being an effective parameter for speaker identification. To list some supporting studies, a study by Hasan, Jamil, Rabbani, and Rahman (2004) used MFCCs for feature extraction and vector quantization and found 57.14% of speaker identification score when codebook size (number of co-efficient) was 1 and increased to 100% when codebook size (number of co-efficient) became 16. Hence conclude MFCCs to be best for speaker identification.

In another study by Mao, Cao, Murat, and Tong (2006) considering LPC and MFCCs for speaker identification found speaker recognition rate for 50 speakers to be increased from 42% to 80% for text-dependent and for text-independent recognition rate increased from 60% to 72%. With reference to the reason behind the decreased recognition rate, a study by Wang, Ohtsuka, & Nakagawa (2009) on consideration of new phase information integrated with MFCCs found a reduction in the error rate. The other variable of increasing the number of filters to 32 in MFCCs, Tiwari (2010) found 85% accuracy in speaker recognition task and Chandrika (2010) reported overall performance of speaker verification system using MFCCs was about 80%.

MFCCs were used in comparison with cepstral coefficients for speaker identification in Malayalam nasal coarticulation by Jyotsna (2011) and as a result, the author obtained 80% speaker identification when cepstral coefficients were used and the percentage increased to 90% when MFCCs was used. Following this, the electronic vocal disguise for females using MFCC for speaker identification was studied by Ramya (2011) and it was reported that the percent correct identification was above chance level. From all these reviews, it is observed that the speaker identification was better irrespective of various variables in accordance with the MFCC.

As an advanced study Ridha (2014) and Ayesha (2016) studied the benchmark for speaker identification using MFCCs in nasal continuants in Hindi and Urdu speakers respectively. They found /ŋ/ to be the best for speaker identification among the nasals /m/, /n/ and /ŋ/. Nithya (2015) reported a benchmark for speaker identification using three Tamil nasal continuants in live recording and mobile network recording conditions and found /m/ to be reliable for speaker identification compared to /n/ and /n̪/. Chandrika (2015) reported a benchmark for speaker identification using three Kannada language nasal continuants in live recording and mobile network recording conditions. Results revealed nasal continuant /n. / having the highest percent of correct speaker identification for direct recording and /m/ and /n/ for network recording. Thus, the above-listed studies and the present findings with reference to the parameter MFCCs, the speaker identification percentage seems to be better and higher.

However, when the samples were recorded in different conditions the accuracy of speaker identification scores was reduced due to various factors like speaker distortion, system distortion, the influence of background noise, and so on. The influence of these factors is varied with reference to the comparison made between two similar recording conditions and between two varied recording conditions in any speaker identification process. Thus, a **comparison made between any recording conditions** could be considered as the third contributing factor. Accordingly, in the present study, there was a decline in the speaker identification scores of field recording (traffic) conditions compared to the lab recording condition. The average of percent correct speaker identification scores were 88.34%, 87.61% and 77.11% for vowel /a:/, /i:/ and /u:/ respectively when lab condition was compared with lab condition. The average of percent correct speaker identification scores decreased to 87.22%, 81.99% and 66.77% for vowels /a:/, /i:/ and /u:/ respectively when traffic (BNR) condition was compared with traffic (BNR) condition and in addition average of percent correct speaker identification scores declined to subsequent level as 79.38%, 76.72% and 53.22% for vowels /a:/, /i:/ and /u:/ when traffic (ANR) condition was compared with traffic (ANR) condition and also the average of percent correct speaker identification scores later decreased to 70.25%, 63.83% and 35.72% for vowels /a:/, /i:/ and /u:/ when traffic (BNR) condition was compared with traffic (ANR) condition.

These results of the present study correlate with the previous studies, where Jakhar (2009) obtained better results when similar recording conditions were compared that is when the live recording was compared with live recording and also when the mobile recording was compared with mobile recording and the poor result was obtained when the live recording was compared with mobile recording. A study on automatic speaker identification using Workbench software by Ridha (2014) reported 100%, 90%, and 100% for /m/, /n/, and /ŋ/ nasal sounds of Hindi language samples when the live recording was compared with live recording and 50%, 80% and 90% for the same nasal sounds /m/, /n/ and /ŋ/ when mobile network recordings were compared with mobile network recordings. Thus, it was concluded that live recording was better compared to mobile recording. Nithya (2015) conducted similar study in Tamil language samples and found 97.6%, 85.6% and 76.5% of speaker identification scores for /m/, /n/ and /n̪/ in live recording condition. For mobile network conditions, the scores were 83.5%, 65.8%, and 68.3% respectively. Proving that, the live recording consists of better samples for speaker identification compared to the samples from mobile network recording conditions. Similarly Suman (2015) conducted a study in Kannada language by considering vowel /a:/, /i:/ and /u:/ following the nasal consonant /m/ and /n/. It was reported that speaker identification scores were 71.16%, 73% and 65.66% for vowels following the nasal consonant /m/ and 77.83%, 81.33% and 68.83% for vowels following the nasal consonant /n/ in live condition. Whereas for the mobile network condition the scores reduced to 68%, 67% and 48.33% for /m/ and 75%, 63% and 67% for /n/. The contributing reasons for these changes could be with reference to the communication channels. For example, through any communication channels during the transmission of voice signals, the errors are associated with the signals which are reproduced due to the distortions caused by the microphone, channel, and noises, electromagnetic and acoustical interferences thereby

affecting the transmitting signal for example in the mobile network. The network used was Vodafone and Airtel (GSM 900/GSM 1800 MHz) and the GSM (Global System for Mobile Communications) was the pan-European cellular mobile standard. Here the speech signals were compressed before transmission because the speech coding algorithms are part of GSM. It also reduces the number of bits in digital representation at the same time however maintaining the acceptable quality of the signal. Thus, this process modifies the speech signal and can have an influence on speaker recognition performance along with perturbations introduced by the mobile cellular network (background noise, channel errors) (Barinov, Koval, Ignatov & Stolbov, 2010). This could be one of the possible contributing reasons when it is concerned with the different modes of speaking, however, the present study used a live mode of recording.

These distortions modify the formant's energy and position which are fundamental for speaker identification. Another study by Barinov, Koval, Ignatov, and Stolbov (2010) examined the characteristics of speech transmitted over a mobile network. The noticeable changes in the energy distribution were caused by the non-linearity of the GSM channel's frequency response in the range 750-2000 Hz. Thus, affecting the 2nd and the 3rd formants (F2 and F3) and also reported a shift in the fourth formant (F4) which was due to the fall-off in the channel's frequency response at 3500 Hz. As a result, the parameter MFCC was affected.

Similar study by considering vowel /a:/, /i:/ and /u:/ preceded by the nasal consonant /m/ and /n/ in Indian context was carried out by Arjun (2015) in Kannada language. It was reported that the correct percent speaker identification score was 92%, 80% and 80% for /m/ and 93%, 78% and 80% for /n/ in live condition and scores reduced to 75%, 58% and 51% for /m/ and 72%, 49% and 53% for /n/ in mobile network condition.

Following this when the live recording was compared with the live recording by Ayesha (2016), it was reported that the percent correct speaker identification was 70%, 80%, and 100% for /m/, /n/, and /ŋ/ and when mobile network recordings were compared with mobile network recordings the scores were reduced to 60%, 70% and 60% for /m/, /n/ and /ŋ/. Therefore the speaker identification scores in live condition were better than in-network condition, this difference is contributed by the differences in the recording characteristics of mobile network versus live voice. From the above discussion, it was clear that speaker identification scores were poor in field conditions compared to lab conditions. When lab condition was compared with field condition the performance reduced to a greater extent compared to lab and field condition.

From the above-supporting studies, it was clear that speaker identification scores reduced in-network condition which can be due to the transmission of speech signal through communication channels, the signals which are reproduced with errors caused by distortions from the microphone and channel, acoustical, electromagnetic interferences, and noises affecting the transmitting signal. These distortions affect the acoustical parameters/properties of speech sounds (Examples: formant energy and position) which are crucial for speaker identification.

Apart from the above-mentioned factors, the fourth contributing factor for the results of the present study is the **individual variability of the speakers** considered for the study. The speech/linguistic aspects of any individuals are influenced to change with reference to varied personality, emotional status, educational level, world knowledge, speaking situations, and interview method, etc. In the present study, there was a decrease in the score of speaker identification in field conditions compared to lab conditions. The above-mentioned variables could have contributed to this difference in the results even though the recording procedure was counterbalanced. Scores were poorer in field conditions compared to lab conditions. This could be due to the speaker variability factor, the unnoticed variations in the speaker's emotional state playing an important role during the recording procedure, and despite the structured stimulus material being used in specific recording settings and with varied trails of tasks. The repeated speech utterances (recorded speech) cannot be replicated with reference to the trials or repetitions since the **speech is very complex**. Similarly, most of the forensic case speech samples are non-contemporary samples, which consist of the questioned sample and suspected (reference) samples which are recorded or extracted in two contexts by the police personnel where the criminals' emotional state would not be similar.

For example, a study was done by Devi, Srinivas, and Nandyala (2014) and Ghiurcau, Rusu, and Astola (2011) reported that the performance of the speaker recognition system reduced significantly when the emotional state altered in a human voice. Meanwhile, the environment also plays a major role. Recording of suspect's (reference) speech would be done in the recording room which is noise-free in the police station but the test sample will be in field condition which will be distorted due to various factors. Background noise also plays an important role where it will be present in wide-frequency range and filtering of speech from noise will be difficult hence alters the speaker's acoustic features.

Gong (1995) found the low performance by the speech recognizers when the reference and the test samples environment altered. Das et al., in 1993 found 1% error rate when the system trained under quite conditions and the error rate increased to more than 50% in a cafeteria environment. Singh and Rajan (2011) also reported that the accuracy of the speaker recognition system degraded because of the presence of background noise which was the dominating factor that affected the speech signal.

Therefore the **use of the noise reduction technique would resolve the issue related to the influence of background noise on speaker identification**. Thus, in the present study, the recorded speech samples were subjected to a module called street noise reduction and later introduced for speaker identification. The results revealed that after noise reduction (ANR) the speaker identification scores reduced [Traffic (ANR) v/s Traffic (ANR)- (/a:/- 79.38%, /i:/- 76.72% and /u:/- 53.22%)] compared to before noise reduction (BNR) [(Traffic (BNR) v/s Traffic (BNR)- /a:/- 87.22%, /i:/- 81.99% and /u:/- 66.77%)] for the comparison between similar recording conditions. In contrast, when the comparison was made between two different recording conditions, that is when lab condition was compared with field condition the identification score was better in only lab condition comparison [(Lab v/s Lab- /a:/- 88.34%, /i:/- 87.61% and /u:/- 77.11%)] compared to lab verses field condition

comparison. In this lab verses field recording condition also after the application of noise reduction technique the speaker identification scores reduced [(Lab v/s Traffic (ANR)- /a:/- 58.61%, /i:/-58.94% and /u:/- 38.11%)] compared to before noise reduction [(Lab v/s Traffic (BNR)- (/a:/- 65.77%, /i:/- 62.27% and /u:/- 42.61%)]. Therefore, there is no considerable improvement in the speaker identification score after the application of the noise reduction technique for the comparison between similar and different recording conditions for the vowels/a: /, /i: /, and /u:/. Therefore the module called 'street noise reduction' of Sound Cleaner software is not very effective in resulting in a higher speaker identification score.

However, since the first voice identification is based on many different approaches to speaker identification. Except for the one based on listening (auditory, psychological, linguistic, etc), all the other is based on spectral analysis (formant matching, microanalysis, voiceprint, etc). Therefore, the most important factor for both the automatic systems and the experts is the accuracy and the quality of the spectral image according to Kersta (1962), Goldstein (1976), and Barinov, Koval, and Ignatov (2010). Thus, with reference to the results of the present study the **possible parameters which affect instrumental identification analysis will be discussed in this section.** The parameters affecting the spectrum also affect the quality of speech. According to Barinov (2010), for example, the parameter called *overloading, signal-to-noise ratio, reverberation, the nonlinearity of frequency response, sampling frequency and bit rate* are the few possible parameters which are important as speech signal's parameters and one should follow the guidelines during the application of any noise reduction technique. The following are the illustrations given for each parameter by the same author. The sound cleaner software also follows relatively similar parameters in the process of noise reduction, but in the present study, the module considered did not result in a higher speaker identification score. However, the list of few important parameters to be considered during the noise reduction process is discussed with illustrations in detail under the following sections.

Overloading or clipping would occur in terms of signal amplitude limitation that is when the dynamic range of one of the elements in the recording chain does not match the dynamic range of the recorded sound signal. Figure 5.1, showing the average spectra of the original 100 Hz sine wave and after clipping the spectrum is changed with additional frequencies. In any automatic voice identification system, the difficulty will initiate when the maximum amplitude level reaches 10-15% in corresponding to the degree of clipping. Once the recording has been closed, there would be no chance to compensate for the influence of such clipping. Thus, the remaining option would be the adjustment and proper selections of equipment in the recording chain.

To add on, since the speech signal is dynamic with varying amplitude; the above-mentioned problem is very negligible. This is because very often the speech fragments are found without any clippings even if the quality of recorded files is poor. This segment is recommended to consider for further processing and analysis.

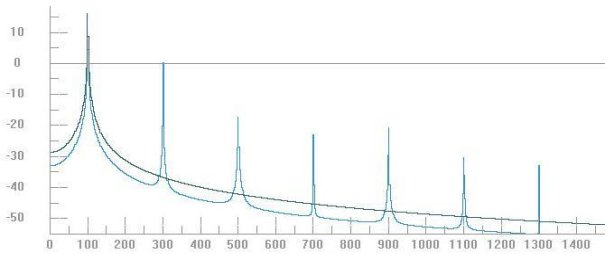


Figure 5.1: Average spectra of the original 100 Hz sine wave (Blackline) and after clipping (blue line)

Another example of a single sound is depicted in Figure 5.2. On observation, the level of the original third formant (2900 Hz) is seen to be towards the level of additional frequencies in the range 1500-2500 Hz. These frequencies as a formant will be accepted as the false formant by both experts and automatic feature extraction algorithms.

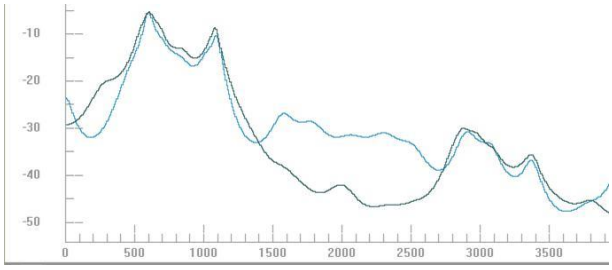


Figure 5.2: Instantaneous LPC spectra of initial “o”-like sound (Blackline) and after clipping (blue line)

The s*ignal-to-noise ratio* describes how much higher the level of useful signal is than the level of the unwanted signal (noises). In Figure 5.3 (a) High-quality signals, (b) Signal with the SNR at 55-60 dB.

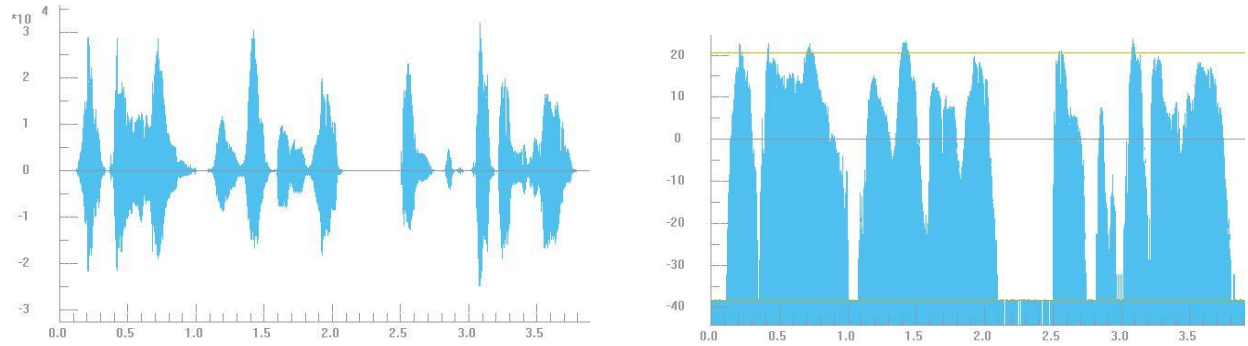


Figure 5.3 (a) and (b): High-quality speech signal without any noises and with SNR 55-60 dB

In some instances the signal will be noisy, this includes a combination of various types of noises which mask the identification features in a form of visible speech and mask the useful signal as perceived audio. Thus, from a noisy speech sample, it becomes sometimes impossible to extract individual features (almost all of them are spectral based) of

the speaker even in the noise cancellation processing as shown in Figure 5.4. With reference to the dynamic LPC of the spectrogram, in quiet recording and a noisy environment is shown in Figure 5.5.

To run formant's based identification in any automatic feature extraction algorithms the SNR of 10dB is not always sufficient. The type of noise and the pronounced sound at the moment plays an important role. But a minimum level of SNR equal to 10dB is required for the formant analysis of speech samples by experts and automatic systems. But in the present study, the SNR was higher.

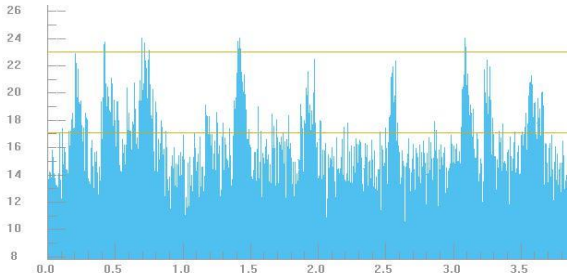


Figure 5.4: SNR of noisy recording very low and even negative

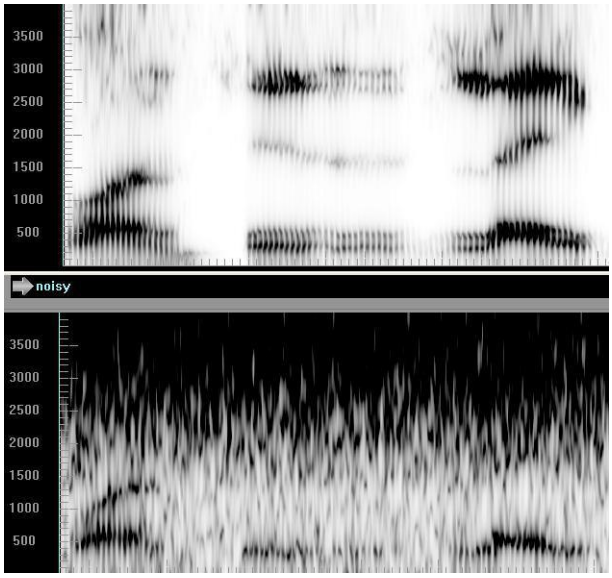


Figure 5.5: Dynamic LPC spectrograms of a clean recording (on top) and a noisy one (at the bottom).

Reverberation happens when the initial signal gets reflected due to different surfaces and the reflected signals combine with the original signal. Therefore a combination of the reflected signal which is picked by the microphone with a time delay corresponding to the distance of the reflecting surface and its initial signal form the reverberated signal. The quality of the reflecting surface decides the amount of similarity between the initial signal

and the reflected signal. Thus, reverberation changes the sounding, waveform, and spectrum of every sound signal. The extent of reverberation is decided by the parameter called reverberation time that is the reflections that need to decay to the level of initial signal minus 60 dB in a millisecond as depicted in Figure 5.6. The dynamic FFT spectrograms of a signal recorded in a sound-treated room (on top) and the reverberated one in a different environment (at the bottom) are represented in Figure 5.7.

From the above-mentioned point of view, it is important to consider two different sounds separately during any type of speech analysis. Since there would be a possibility of overlapping of the spectrum from one sound to another leading to feature extraction mistakes. Therefore it is recommended that on 20dB level the maximum time needed for an impulse to decay should be no larger than the duration of the average sound. This value is noted to be around 100ms and 300ms for reverberation time and the initial value of up to 60dB suppression.

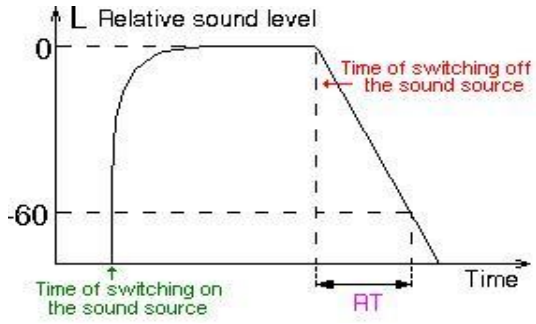


Figure 5.6: The reverberation time measurement

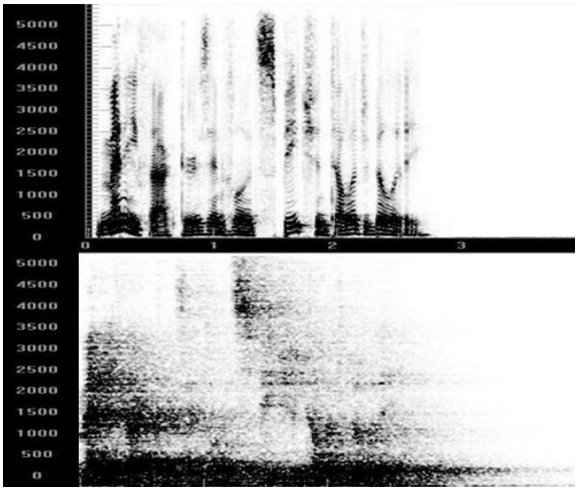


Figure 5.7: Dynamic FFT spectrograms of a quiet recording (on top) and the reverberated one (at the bottom)

The non-linearity of the frequency response is related to the recording device. Every device in the recording chain will have its amplitude-frequency response which executes its function as a filter thus alter the initial spectrum of the speech signal. Due to this, the recording chain output of the original signal's spectrum significantly changes as shown in Figure 5.8. It is recommended that the frequency response of any speech recording device

(from phone or radio channels) should be linear in the frequency range from 100Hz to 5500Hz or at least to 4000Hz or flat. In the present study also the recording frequency response was 4000Hz for lab condition whereas it was different when the recording was done in field condition. These differences were due to the recording devices used in these conditions.

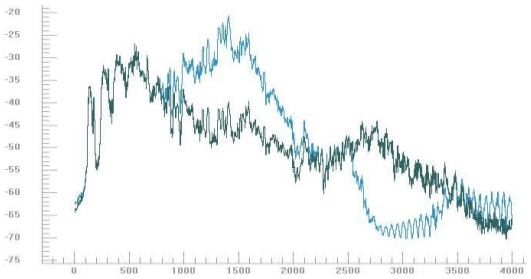


Figure 5.8: Average FFT spectra of the original sound signal and the same signal recorded through the devices with a non-linear amplitude-frequency response.

Sampling frequency and bit rate will determine the amount of information being recorded from the initial analog form and the same saved as a digital audio file. Here, the sampling frequency limits the maximum existing frequency in a digital file according to the Nyquist theorem. The frequency as 8000, 11025, 16000, 22050, 32000, 44100 Hz, etc is the common sampling rate. Therefore, the frequency range of the original signal will be limited up to 4000, 5512, 8000, 11025, 16000, 22050 Hz, etc. respectively. Thus, resulting in the exclusion of the high formants from the speech signal's spectrum, and the same is represented in Figure 5.9.

In any practical scenario of the recorded speech signal, the maximum frequency with which it can be heard or recorded will be around 5500Hz and it goes a bit higher in very rare cases. This implies that it is sufficient that in any speech analysis and voice identification the sampling rate should be twice bigger than its maximum frequency. Therefore, for recording from microphone channels the sampling frequency should be 11025Hz or 16000Hz and for any type of recording from phone or radio channels, it should be 8000Hz. In the present study, this particular point of using a proper sampling rate might have not resulted in losing the high-frequency range and all highest formants which are important for speaker identification.

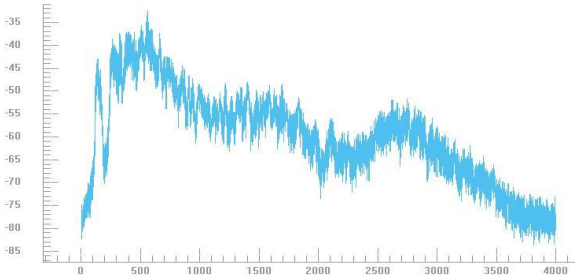


Figure 5.9: Average FFT spectrum of a speech signal sampled with 8000Hz.

The dynamic range of the recorded audio signal will be limited because of the bit rate. Every bit of quantization carries information about 6 dB of the dynamic range of the signal. Generally, 8-bit, 16-bit, 24-bit, etc are the popular bitrates. With reference to the dynamic range of the input signal, these correspond to 48, 96, 144 dB, etc. The digital clipping of the speech signal will be caused due to the incorrect bit rate and this will cause a similar problem as in the case of overloading. For all the types of speech analysis or voice identification, it is proposed to use 16-bit coding or a maximum of 24-bit coding since the dynamic range of human speech is less than 96dB. For instance, the digital clipping and corresponding side effects can happen when the lower non-sufficient bit rate is recommended.

Therefore all these above-mentioned alterations in the speech signal will have a varied influence on the speaker identification task. However, in the present study, a default noise reduction technique called the "Street Noise Reduction" module of Sound Cleaner software was used and is the product of Speech Technology Center Ltd. The software doesn't illustrate the changes happening with the speech signal after the application of the noise reduction technique as shown in the study by Barinov (2010). However, Barinov (2010) has carried out a series of experiments involving forensic experts and automatic voice identification system Voice Net, produced by Speech Technology Center Ltd. In general, the points under the parameters listed in Barinov's (2010) study can be considered as a supporting variable for the difference in speaker identification score obtained for the conditions and/or comparisons before and after the application of the noise reduction technique in the present study.

But, some studies are addressing the correlation between the influences of noise reduction technique on the speech parameters, and however, the results of the present study show a positive correlation. The noise reduction technique would have contributed some changes in the speech parameters involved for speaker identification since there is a reduction in the speaker identification scores after the application of the noise reduction technique irrespective of the comparisons (two similar recording situation before noise reduction/after noise reduction, two different recording situation before noise reduction/after noise reduction) and the vowels (/a:/, /i:/ and /u:/) considered. However, in contradiction to the results of the present study, the following are the studies in support of the noise reduction technique not affecting the speech parameters.

For example, the known fact is, the speech gets corrupted when it is influenced by any kind of noise. But with the use of any kind of noise reduction methods the speech properties will be enhanced (no pictorial illustrations are provided). Thus, the spectral noise subtraction method was used by Berouti, Schwartz, and Makhoul (1979) to enhance speech which was corrupted by broadband noise, and found no loss of intelligibility in the speech. According to Udrea and Coichina (2003), the spectral noise subtraction method is defined on the basic principles of the spectral subtraction method that is to subtract the short-term spectral magnitude of noise from the signal. Average signal and average noise are estimated and subtracted from each other. This will make the signal-to-noise ratio improved. For instance, in vowels since the frequency properties are known; the noise separation properties would become much easier according to Davis (2002).

On the other hand, in the frequency domain, the Fourier method details the spectral content of the signal. The time-domain information for a particular event will be lost because the preservation of time instances is not considered while using the Fourier transformation. If the signal is stationary this condition can be unnoticed. Whereas, the speech signal comprising acoustic waves carry information in a non-stationary way. To overcome this problem another alternative method was proposed by Shonda and Simon (2003) called the Wavelet analysis. To get the accurate signal representation by producing precise decompositions of signal a concept called multi-resolution analysis is used in Wavelet. Here the higher-order derivatives, small discontinuities, and self similarities can be revealed. Like many traditional methods, this does not remove noise by low-pass filtering but includes the nonlinear function. The low-pass filtering method which is a linear time-invariant, might blur the sharp features in the signal and sometimes difficult to separate noise from the signal where the Fourier spectra overlap. The noise is removed based on the threshold of the Wavelet coefficients. This is because; the values of the signal that has energy concentrated in a small number of Wavelet coefficients will be large in comparison to the noise. This process allows features in the original signal to remain sharp. The lack of shift-invariance is the only disadvantage of Wavelet de-noising which means the Wavelet coefficients will not move by the same amount that the signal has shifted. But all the de-noising results can be averaged over all possible shifts of the signal to overcome the problem. This process of noise reduction using Wavelet coefficients threshold can be obtained in MATLAB command wdencmp.

The other noise reduction technique could be the use of inverse filtering by Barinov et al. (2010) and found no change in the formant structure of the speech sound. On the other hand with the unknown temporal-spectral characteristics of speech signal combined multi-condition model training and missing-feature theory was used to model noise by Md Imdad, Akhtar, & Md Imran (2012). There was a positive result where the speech signal was not affected by the removal of background noise. In missing feature theory, for example, consider a spectrum that has been passed through a high-pass filter. If we assume that the first eight spectral magnitude features are below the threshold and are labeled as “missing.” Once each spectral magnitude feature in a frame is labeled as present or missing, a computationally simple modification of probability models discards missing features and forms densities that would have been obtained by training without missing features. Another study by Sukor & Syafiq (2012) passed the signal to the pre-treatment process where the background noise was removed. The results showed that the speaker recognition system was able to identify the voice pattern correctly.

Therefore, these various noise reduction methods working based on different principles incorporate certain advantages and disadvantages during the process of noise reduction. Each method has their distinct abilities to reduce noise sources from the given signal. Thus, in the present study, the existing Sound Cleaner Software was used to apply noise reduction technique to the traffic field recording signal and the speaker identification scores were obtained. Therefore to summarize the results, the average percent correct speaker identification scores reduced drastically in traffic conditions compared to lab conditions, and

also the average percent correct speaker identification scores were poorer in traffic conditions after the application of noise reduction technique compared to before the application of noise reduction technique. Hence to conclude, the module called ‘street noise reduction’ of Sound Cleaner software was sensitive to some extent only in reducing the traffic noise. This is because during the noise reduction processes the speaker-specific acoustic features would have been altered to a greater extent which resulted in revealing poorer scores in traffic conditions after the application of noise reduction technique when compared to before noise reduction technique. Therefore to conclude the study, the outcome after the application of noise reduction technique on speaker identification for traffic noise was not effective in comparison with lab recording condition. Among vowels, the average percent correct speaker identification scores were better for the vowel /a:/ followed by /i:/ and /u:/. Hence to conclude vowel /a:/ acts as a better cue for speaker identification.

CHAPTER VI

SUMMARY AND CONCLUSIONS

Every matured voice has unquestionably a unique character dependent upon the structure of the head, neck, and face of the individual. The coordination among these structures in association with the nervous system results in speech production. The most natural and common way used to communicate information by humans is through speech. Speech signal conveys several types of information. For example, speech signal conveys linguistic information (language and message) and speaker information (regional, emotional, and physiological characteristics). With reference to speaker information, different individuals sound different with respect to their voice, which is a known fact. This can be illustrated with an example of how an individual is identified through his voice in any telephone conversation. This is due to the property of individuals' speech being speaker-specific. The same principle is considered in one type of speaker identification method. The method in which a person is recognized exclusively (perceptually) from his voice and is known as speaker recognition which is known for long period (Atal, 1972). Among the biometric identifiers such as speech or handwriting, verification of individuals' identity based on the voice has significant advantages and practical utilizations because speech is a product of an underlying anatomical source, namely, the vocal tract and a result of natural production. Thus, comprising inherent constrained biometric feature where it does not require a specialized input device, therefore the user acceptance of the system would be high. In recent advances to improve the performance and flexibility of speaker recognition, new tools have been produced in speech technologies. Telephone conversation has increased in recent years. Due to the increased usage of mobile phones for conversational purposes, the crime rate is increasing drastically by misusing the same for many crime-related activities like bomb threats, ransom demand, sexual abuse, and hoax emergency call. In these conditions, voice is the only evidence available for analysis. Hence there is a need in the measurement of the voice for the establishment of legal proof by police and magistrates.

Therefore a method called speaker identification aims 'to identify an unknown voice as one or none of a set of known speakers on comparison' (Naik, 1994; Nolan, 1983). Speaker verification is another common task in speaker recognition in which 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). Hence, Forensic Speaker Identification is seeking an expert opinion in the legal process as to whether two or more speech samples are of the same person. Thus, according to some set of authors speaker recognition can be studied under two headings: a) speaker identification and b) speaker verification (Fururi, 1994; Nolan, 1997; Rabiner & Juang, 1993; Rose, 2002).

Speaker recognition is affected by various factors. With reference to the different contexts of the conversational speech sample, the interesting one is the background noise. Since the speaking environment is always associated with one or more types of noise, the considered speech sample may be accompanied by some noise. Thus, for the listeners, the speech will not be heard clearly. Thus, background noise also plays a major role in forensic speaker identification. Most of the speech recognition instrument will have difficulty in identifying speech signal when it is accompanied by background noise. To overcome this problem, the noise has to be filtered so that the required speech signals will be free from noise and the same will be used for further analysis. Various approaches have been implemented to improve the noise robustness of speaker recognition. The following are the techniques which can be listed in general: Techniques such as Kalman filtering (Fingscheidt. Suhadi, Stan, 2003) or Spectral Subtraction (Garcia & Rodriguez, 1996) can be used to filter noise from speech, based on the prior knowledge of the noise characteristics. It is also possible to extract noise-robust features, e.g. relative spectral features (Hermansky & Morgan, 1994) from speech signals instead of removing the background noise. It is also possible to ignore the parts of speech corrupted by background noise using missing feature theory (Bonastre, Besacier, & Fredouille, 2000). The above approaches are used in statistical speakers' models (e.g Hidden Markov Models (HMMs) or Gaussian Mixture Models (GMMs).

However, the global leader in Speech Technologies Center is a leading developer of voice and multimodal biometric systems, as well as the solutions for audio and video recording, processing, and analysis. For over 20 years, the SpeechPro under STC has been developing specialized tools for efficient noise reduction and text transcription of low-quality recordings. Various studies on the perception of poor audio recordings and noisy speech signals carried out by SpeechPro have resulted in the formation of the unique sound filtering algorithms that are now presented in the software and hardware products like Sound Cleaner, ANF II, and The Denoiser Box. In the present study, the Sound Cleaner Signal Enhancement Program Model 5142 (Noise Cancellation Software) was used to reduce the background noise and an attempt has been made to see its effect on speaker identification score for the samples which was subjected to noise reduction.

Thus, in the present study, speaker identification was carried out using the machine method using the semi-automatic speaker identification process. This has been selected from the classification of Hecker (1971) and Bricker and Pruzansky (1976) speaker identification as: (i). Speaker identification by listening, (ii). Speaker identification by visual method & (iii) Speaker identification by the machine which is subdivided into (a) Semi-automatic speaker identification and (b) Automatic speaker identification.

Therefore, the present study focuses on the Semi-automatic Speaker Identification (SAUSI) where the known and the unknown samples from the speaker are selected by the examiner and are processed by the computer program to extract certain parameters. And the final interpretation will be made by the examiner. Few examples of such studies are with the

parameter-first and second formants (Atal, 1972; Hollien, 1990; Kuwabara & Sagisaka, 1995; Lakshmi & Savithri, 2009; Nolan, 1983; Stevens, 1971), higher formants (Wolf, 1972), fundamental frequency (Atkinson, 1976), fundamental frequency contours (Atal, 1972), Linear prediction coefficients (Markel & Davis, 1979; Soong, Rosenberg, Rabiner & Juang, 1985), Cepstral coefficients and Mel-Frequency Cepstral Coefficients (Atal, 1974; Fakotakis, Anastasios & Kokkinakis, 1993; Rabiner & Juang, 1993; Reyond & Rose, 1995), Long-Term Average Spectrum (Kiukaanniemi, Siponen & Matilla, 1982).

Among these short and long term acoustical parameters, Mel-Frequency Cepstral Coefficients (MFCCs) are extensively used in the present era for speaker identification tasks and has been shown to yield tremendous results (Hasan, Jamil, Rabbani & Rahman, 2004; Jyotsna, 2011; Mao, Cao, Murat & Tong, 2006; Singh & Rajan, 2011; Tiwari, 2010; Wang, Ohtsuka, & Nakagawa, 2009). Mel-frequency cepstrum is a cepstrum with its spectrum mapped onto the Mel- Scale before the log and inverse Fourier transform is taken. As such, the scaling in Mel-frequency cepstrum mimics the human perception of distance in frequency, and its coefficients are known as the MFCC. The present study will be focusing on the usefulness of Mel -Frequency Cepstral Coefficients (MFCC) on speaker recognition.

It is evident from these reviews that MFCCs are perhaps the best parameter for speaker identification and less susceptible to variation of the speaker's voice and surrounding environment (noise). Also, the vowels may be the most suitable among speech sounds for speaker identification. However, to date, there are limited studies on vowels as strong phonemes for speaker identification using semi-automatic methods in the presence and absence of noisy situations and after the application of speech signal to any noise reduction techniques. In the present study, the Sound Cleaner software (speaker recognition instrument) is used to reduce the noise and study the effect of the same on speaker identification. In forensic sciences, the scientific testimony has to be provided to impress any court of law and from whichever country the research would have been executed. However, for any result to be called scientific, it has to be measured, quantified, and reproducible if and when the need arises. Therefore, a method to carry out these analyses becomes a must. In this context, the present study was conducted.

Accordingly, the current study aimed to investigate the effect of noise and noise reduction technique on speaker identification using MFCCs on long vowels in the Kannada language for lab condition (Condition I), traffic condition (Before Noise Reduction- BNR) (Condition II), traffic condition (After Noise Reduction- ANR) versus traffic condition (ANR) (Condition III), lab versus traffic condition (BNR) (Condition IV), lab versus traffic condition (ANR) (Condition V) and Traffic condition (BNR) versus traffic condition (ANR) (Condition VI).

A total of 60 participants with 30 males and 30 females in the age range of 20-40 years were considered for the study. All the participants were native speakers of the Kannada language with no history of speech, language, hearing problems, no associated psychological or neurological problems, and no reasonable cold or respiratory conditions at the time of

recording and normal oral structure. Commonly occurring hypothetical Kannada meaningful sentences with long vowels /a:/, /i:/, /u:/ was used as material for a reading task. The same was recorded in two different conditions: I- Laboratory condition and II- Traffic (Field) condition. These recorded samples were analyzed under two phases: 1- Before noise reduction (BNR) and 2- After noise reduction (ANR), using Sound Cleaner- Universal Noise Cancellation Software. In Sound Cleaner software, the 'Street Noise' scheme was selected for the present study. Further, Speech Science Lab Workbench, a Semi-Automatic vocabulary dependent speaker recognition software was used to extract Mel-Frequency Cepstral Coefficients (MFCC) for the truncated (PRAAT software) vowels.

Thus, the MFCCs derived from the vowels were used to compute the Euclidian distance between the test and reference samples. For the present study, the feature vector chosen was MFCCs 13 coefficients. Upon choosing the feature vector, the system computes a measure of Euclidian 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 scores was displayed. The analyses were performed in terms of obtaining correct percent speaker identification scores separately for lab v/s lab condition, traffic (BNR) v/s traffic (BNR) condition, traffic (ANR) v/s traffic (ANR) condition, lab v/s traffic (BNR) condition, lab v/s traffic (ANR) condition and traffic (BNR) v/s traffic (ANR) condition. Repetitions were done by randomizing the testing and training samples and the speaker identification thresholds were noted for the highest score and the lowest score.

To explain in brief, for an average of 30 trials of randomization the results revealed that, (I). When lab condition was compared with lab condition, the percent correct speaker identification for vowels /a:/, /i:/ and /u:/ were 88.34%, 87.61% and 77.11%. (II) When traffic condition (BNR) was compared with traffic condition (BNR) the percent correct speaker identification for vowels /a:/, /i:/ and /u:/ were 87.22%, 81.99% and 66.77%. (III). When lab condition was compared with traffic condition (BNR) the percent correct speaker identification for vowels /a:/, /i:/ and /u:/ were 65.77%, 62.27% and 42.61%. (IV). When traffic condition (ANR) was compared with traffic condition (ANR) the percent correct speaker identification for vowels /a:/, /i:/ and /u:/ were 79.38 %, 76.72 % and 53.22 %. (V). When traffic (BNR) condition was compared with traffic condition (ANR), the percent correct speaker identification for vowels /a:/, /i:/ and /u:/ were 70.25 %, 63.83 % and 35.72 %. (VI). When lab condition was compared with traffic condition (ANR) the percent correct speaker identification for vowels /a:/, /i:/ and /u:/ were 58.61 %, 58.94 % and 38.11 %.

To summarize the results, among vowels /a:/, /i:/ and /u:/, the average percent correct speaker identification for the vowel /a:/ was better in all the conditions (I, II, III, IV, and V). The present study is in consonance with the previous studies where Arjun (2015) and Aswathy (2016) found vowel /a:/ to be better compared to /i:/ and /u:/, Jakhar (2009), found vowel /a:/ to be better in live condition and vowel /i:/ in mobile network condition, Medha (2010), found vowels /a:/ and /i:/ to be better. But in contradiction Chandrika (2010), found vowel /i:/ to be better compared to /a:/ and /u:/ and Sreevidya (2010) and Ramya (2011)

found /u:/ to be better for speaker identification. When different conditions were considered performance in lab conditions was better compared to field conditions (embedded with or without noise). In specific the performance scores decreases in the following order; lab v/s lab condition, traffic (BNR) v/s traffic (ANR) condition, traffic (ANR) v/s traffic (ANR) condition, traffic (BNR) v/s traffic (ANR) condition, lab v/s traffic (BNR) condition, and lab v/s traffic (ANR) condition. Therefore average percent correct speaker identification scores were better in lab condition and poor in traffic condition and poorer in lab v/s traffic condition. The studies which support the present study are Jakhar (2009), Ridha (2014), Nithya (2015), Suman (2015), Arjun (2015), Ayesha (2016) where the authors found better scores in live condition and poor scores in mobile network condition and poorer scores in live v/s mobile network condition.

In the present study also there was a decline in the score of speaker identification for field condition compared to lab condition. The following reasons could be contributing factors. (1). From the above-supporting studies, it was clear that speaker identification scores reduced in-network condition which can be due to the transmission of speech signal through communication channels, the signals which are reproduced with errors caused by distortions from the microphone and channel, acoustical, electromagnetic interferences, and noises affecting the transmitting signal. These distortions affect the formant energy and position which are crucial for speaker identification. Thus, the quality and accuracy of the spectral picture is the most important factor for both experts and automatic systems (Barinov, Koval, Ignatov, 2010; Goldstein, 1976; Kersta, 1962). These authors describe only those parameters which affect instrumental identification analysis and this is one of the objectives of the present study. Thus, each of these parameters, affecting spectrum, also affects the perceived quality of speech. The parameters listed are overloading, signal-to-noise ratio, reverberation, the nonlinearity of frequency response and sampling frequency, and bit rate. This might have contributed to the poor percent correct identification score of traffic conditions in the present study.

(2). Different recording situations- During a real speech a person can recognize the surrounding sounds and concentrate on the speech of another person thus filtering the desired information out of various audio environments. Therefore, the ability of a human to recognize and filter sounds significantly increases the intelligibility and comprehension of the speech even if communication takes place in a noisy environment, situation, or condition. This is not in the case of lab condition, where the individuals concentrate on their own speech with no task of filtering another audio environment since there will be complete silence in the lab.

However, in traffic conditions, it is a different situation. The recording equipment does focus on certain audio streams (specialized microphones) and impartially record everything that happens in the audio spectrum. As a product, we receive a ‘flat picture’ of all recorded sounds which often makes the speech partially unintelligible, quiet, and buried in the noises. In addition, the scores were poorer in field condition which can be due to the speaker variability factor where variations in speaker’s emotional state also play an important

role which might not be the same during field recording and also speech cannot be replicated in the same way which was produced earlier during lab recording.

(3). Background noise also plays an important role where it will be present in wide-frequency range and filtering of speech from noise will be difficult hence alters the speaker's acoustic features. The signal in the lab condition does not contain noise and is not subjected to undergo the removal of background noise from voice recognition signal, for example, using the spectral subtraction method. Here, in this method, the short term spectral magnitude of noise will be subtracted from the signal. That is the average noise and average signal are estimated and subtracted from each other (Udrea & Coichina, 2003). Hence, there might be a chance of the signal getting distorted. The phonemic effect on speaker identification is assessed and it is found that the level of correct perceptual identification varies as a function of vowel production, consonant-vowel transitions, vocal tract turbulence, and inflections. Henceforth studies on voice quality, speech prosody/ timing, and many other speaking characteristics have to be considered as an important factor in the identification process.

Davis (2002) reported that the noise separation properties would become much easier in vowels because the frequency properties of vowels are known. The present study focused on long vowels (/a:/, /i:/, and /u:/) and used Sound Cleaner software for noise reduction. The results revealed that after noise reduction the speaker identification scores reduced [Traffic (ANR) v/s Traffic (ANR)] compared to before noise reduction [Traffic (BNR) v/s Traffic (BNR)].

Therefore, the average percent correct speaker identification scores reduced drastically in traffic conditions compared to lab conditions, and also the average percent correct speaker identification scores were poorer in traffic conditions after the application of noise reduction technique compared to before the application of noise reduction technique. Hence to conclude, the Sound Cleaner software was sensitive enough to reduce the traffic noise. But, certain reasons and observations are contributing to the poor speaker identification scores after the application of the noise reduction technique to the traffic condition sample. (1). The signal in the lab condition probably does not contain noise whereas the signal in traffic conditions is improbably embedded with noise. (2). During the noise reduction processes, the speaker-specific acoustic features might have been removed, and (3). The resultant speaker-specific acoustic feature might not be very effective for the calculation of MFCCs in the Workbench software. These variables might have resulted in poorer scores in traffic condition after the application of the noise reduction technique when compared to before noise reduction technique.

Therefore the study can be concluded that the outcome after the application of noise reduction technique on speaker identification for traffic noise was not effective in acquiring 100% correct speaker identification. However, the correct percent speaker identification was relatively higher within the range of 58.62% to 79.38% for the vowel /a: /. Thus, among vowels, the average percent correct speaker identification scores were better for the vowel /a:

/ followed by /i:/ and /u:/. Hence to conclude vowel /a:/ acts as a better cue for speaker identification.

Limitations and future directions

- Sound Cleaner software was relatively sensitive enough to reduce the traffic noise in the present study. However, the extended study is required to note and confirm the changes in the phonemic cue after the application of the noise reduction technique to the speech signal recorded in any conditions.
- Further research should be executed using different recording conditions (cafeteria, market, etc.) and compare among them. And in addition, other advanced noise reduction software could also be used in further studies.
- The study can be extended to reduce the background noise with reference to certain variables like increased participants, stimulus from different languages, and considering the other phonemes like consonants.
- The present study used a semi-automatic speaker identification system (Workbench) to obtain the correct speaker identification scores, but it is also recommended to use an automatic speaker identification system to obtain the correct speaker identification scores.

REFERENCES

Arjun. M. S. & Hema. N. (2014). *Speaker Identification using Fricatives in Kannada Speaking Individuals: A Preliminary Study.* Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Arjun. M. S. (2015). *Benchmark for speaker identification using MFCC on vowels preceding the nasal continuants in Kannada*. Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Aswathy, A. (2016). *Effect of native verses non native languages (Kannada and Malayalam) in a closed set of speaker identification using Mel-Frequency Cepstral Co-Efficient*. Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Atal, B. S. (1972), Automatic speaker recognition based on pitch contours. *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 the Acoustic Society of America, 55* (6), 1304- 1312.

Atkinson, J. E. (1976). Inter and intra speaker variability in fundamental voice frequency. *Journal of the Acoustical Society of America, 60* (2), 440-445.

Ayesha (2016). *Benchmarks for speaker identification for nasal continuants in Urdu in direct and mobile network recording*. Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Barinov, A. S., Koval, S. L., & Ignatov, P. V. (2010). Forensic Speaker Identification based on the Formants Matching Approach. *Forensic Science International Journal.*1-10.

Bechler, D., Grimm, M., & Kroschel. K. (2003). Speaker tracking with a microphone array using Kalman filtering. *Advances in Radio Science, 1,* 113–117.

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP' 79.* (4), 208-211.

Besacier, L., Bonastre, J. F., & Fredouille, C. (2000). Localization and selection of speaker-specific information with statistical modeling. *Speech Communication*, *31*(2), 89-106.

Boersma & Weenink, D. (2009). PRAAT S.1.14 software, restricted from http://www.goofull.com/au/program/14235/speedytunes.html.

Bolt, R. H. Cooper, F. S., Green, D. M., Ham-let, S. L., Hogan, D. L., McKnight, J. G., Pickett, J. M., Tosi, O., & Underwood, B. D. (1979). On the theory and practice of voice identification. National Academy of Sciences, Washington, DC.

Bonastre, J.F., Besacier, L., & Fredouille, C. (2000). "Localization and selection of speaker-specific information with statistical modeling." *Speech Communication, 31*, 89–106.

Bricker, P. D., & Pruzansky, S. (1996). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, *40*(6), 1441-1449.

Bricker, P.S., & Pruzansky, S. (1976). *Speaker recognition: Experimental Phonetics*. London: Academic press.

Chandrika, S. (2010). *The influence of handsets and cellular networks on the performance of a speaker verification system*. Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Das, S., Bakis, R., Nadas, A., & Picheny, M. (1993). Influence of background noise and microphone on the performance of the IBM Tangora speech recognition system. In *Acoustics, Speech, and Signal Processing conference, IEEE International Conference*, Vol.2

Davis, G. (2002). "*Noise Reduction in Speech Applications*", Electrical Engineering & Applied Signal Processing Series, CRC Press. Publication.

Deepa, A., & Savithri, S. R. (2010). Re-standardization of Kannada articulation test. *Student research at AIISH (Articles based on dissertation done at AIISH), 8*, 53-55.

Devi, J. S., Srinivas, Y., & Nandyala, S. P. (2014). Automatic speech emotion and speaker recognition based on hybrid GMM and FFBNN. *International Journal on Computational Sciences & Applications*, *4*(1), 35-42.

Fakotakis, N., Anastasios, T., & Kokkinakis, G. (1993). A text-independent speaker recognition system based on vowel spotting. *Speech Communication, 12*(1), 57-68.

Fingscheidt, T ., Suhadi, C. B., & Stan, S. (2003). An evaluation of VTS and IMM for speaker verification in noise. *Eurospeech,* 1669–1672.

Fururi. S. (1994). An overview of speaker recognition technology. *Proceeding of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 1-8.

Ghiurcau, M. V., Rusu, C., & Astola, J. (2011). Speaker recognition in an emotional environment. *Proceedings of SPAMEC,* 1517-1520.

Gill, M. K., Kaur, R., & Kaur, J. (2010). Vector quantization based speaker identification. *International journal of computer applications*, *4*(2), 1-4.

Goldstein, U. G. (1976). Speaker-identifying features based on formant tracks. *The Journal of the Acoustical Society of America, 59*(1), 176-182.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech communication*, *16*(3), 261-291.

Hansen, J., & Proakis, J. (2000). Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York.

Hasan, R., Jamil, M., Rabbani, G., & Rahman, S. (2004). Speaker identification using Mel-Frequency Cepstral Coefficients. *3rd international conference on electrical & computer engineering, 565-568.*

Hecker, M. H. (1971). Speaker recognition- An interpretive survey of the literature. *ASHA monographs*, *16*, 1.

Hermansky, H., & Morgan, N. (1994). Rasta processing of speech (RASTA) processing of speech. *IEEE transactions on speech and audio processing*, *2*(4), 578-589.

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

Hollien, H. F. (2002). *Forensic voice identification*. Academic Press.

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

Jakhar, S. S. (2009). *Benchmark for speaker identification using Cepstrum*. Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Jyotsna. (2011). *Speaker Identification using Cepstral Coefficients and Mel-Frequency Cepstral Coefficients in Malayalam Nasal Coarticulation.* Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Kalaiselvi, R., & Ramachandraiah, A. (2010). Environmental noise mapping study for heterogeneous traffic conditions. In *Proceedings of 20th International Congress on Acoustics, ICA* (pp. 23-27).

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

Kiukaanniemi, H., Siponen, P. & Mattila, P. (1982). Individual differences in the Long-Term Speech Spectra. *Folia Phoniatrica, 34,* 21-28.

Künzel, H. (1994). On the problem of speaker identification by victims and witnesses. *International Journal of Speech, Language and the Law*, *1*(1), 45-57.

Kuwabara, H. & Sagisaks, Y., (1995). Acoustic characteristics of speaker individuality: control and conversion. *Journal of Speech Communication, 16*, 165-173.

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

Lavner, J. M. D. (1994). Principles of Phonetics, Cambridge: Cambridge University Press.

Linde, Y., Buzo, A., & Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Transactions on communications*, *28*(1), 84-95.

Ling, D. (1978). Auditory coding and recording: an analysis of auditory training procedures for hearing-impaired children. In M. Ross and T. Giolas (Eds.), *Auditory Management of Hearing-Impaired Children*. Baltimore: University Park Press, 181-218.

Luck, J. E. (1969). Automatic speaker verification using cepstral measurements. *The Journal of the Acoustical Society of America*, *46*(4B), 1026-1032.

Mao, D., Cao, H., Murat, H., & Tong, Q. (2006). Speaker identification based on Mel frequency cepstrum coefficient and complexity measure. *Journal of Biomedical Engineering*, *23*(4), 882-886.

Markel, J. & Davis, S. (1979). Test independent speaker recognition from a large linguistically unconstrained time-spaced data base. *IEEE Transcations on Acoustics, Speech, and Signal Processing. 27*(1), 74-82.

Mcghee, F., (1937). "The reliability of the identification of voice". *Journal of General Psychology, 17,* 249-271.

Md Rashidul Hasan, (2004). "Speaker Identification System using MFCC procedure and Noise Reduction Method" sited by Syafiq, B. A. & Sukor. A. In Unpublished Dissertation, University Tun Hussein Onn, Malaysia. 2012.

Md. Imdad, N., Akhtar, S. N., & Md. Akhtar, I. (2012). Speaker recognition in noisy environment. *International Journal of Advanced Research in Computer Science and Electronics Engineering, 1(4).*

Medha, S. (2010). *Benchmark for speaker identification by Cepstrum measurement using text-independent data.* Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

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

Nithya (2015). *Benchmark for speaker identification using Tamil nasal continuants in live recording and mobile network recording.* Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Nolan, F. (1983). *Phonetic bases of speaker recognition*, Cambridge: Cambridge University.

Nolan, F. (1997). "Speaker recognition and forensic phonetics" In: Hardcastle & Laver (Eds.): The Handbook of Phonetic Sciences.

Noll, M. (1964). Short-Time Spectrum and Cepstrum Techniques for Vocal-Pitch Detection. *Journal of the Acoustical Society of America, 36*(2), 296-302.

Noll, M. (1967). Cepstrum Pitch Determination. *Journal of the Acoustical Society of America, 41*(2), 293-309.

Ortega-García, J., & González-Rodríguez, J. (1996). Overview of speech enhancement techniques for automatic speaker recognition. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference of IEEE* (2), 929-932.

O'shaughnessy, D. (1987). *Speech communication: human and machine*. Universities press.

Pamela, S. (2002). Reliability of voice prints. Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Plumpe, M. D., Quateri, T. F. & Reynolds, D. A. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Transactions on Speech and Audio Processing, 7(5), 569-586.

Pollack, I., Pickett, J. M., Sumby, W. H. (1954). On the identification of speakers by voice. *Journal of Acoustical Society of America, 26*(3), 403-406.

Pruthi, T. & Epsy-Wilson, C. Y. (2007). Simulation and analysis of nasalized vowels based on magnetic resonance imaging data. *Journal of Acoustic Society of America, 121*(6), 3858-3873.

Rabiner, L. & Juang, B. H. (1993). Fundamentals of Speech Recognition, Signal Processing, editor A. Oppenheim. Englewood Cliffs: Prentice-Hall.

Ramya. B. M. (2011). *Bench mark for speaker identification under electronic vocal disguise using Mel Frequency Cepstral Coefficients.* Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Ranganathan, M. (2003). Speaker identification in disguised speech. Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

Reich, A. R. & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America, 66*(4), 1023-1028.

Reich, A. R., Moll, K. L., & Curtis, J. F. (1976). Effects of selected vocal disguises upon spectrographic speaker identification. *The Journal of the Acoustical Society of America, 60*(4), 919-925.

Reyond. A. D. & Rose. R. (1995). Robust text-independent speaker identification using Gaussian Mixture speaker models. *IEEE Transaction Speech Audio Process, 3,* 72-83.

Rida, Z, A. (2014). *Benchmarks for speaker identification using nasal continuants in Hindi in direct mobile and network recording.* Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

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

Rose, P. (2006). Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language*, *20*(2), 159-191.

Rothman, H. B. (1977). Decoding speech from tape recordings. Proceedings of Carnahan Conference of Crime Counter measures, Lexington, KY, 63-67.

Seddik, H., Rahmouni, A., & Sayadi, M. (2004). Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier. In *Control, Communications and Signal Processing. First International Symposium of* IEEE (pp. 631-634).

Shonda & Simon (2003). "Speaker Identification System using MFCC procedure and Noise Reduction Method" sited by Syafiq, B. A. & Sukor. A. in Unpublished Dissertation, University Tun Hussein Onn, Malaysia. 2012.

Sigmund, M. (2008). Gender distinction using short segments of speech signal. *International Journal of Computer Science and Network Security*, *8*(10), 159-162.

Singh, S., & Rajan, E. G. (2011). MFCC VQ based Speaker Recognition and Its Accuracy Affecting Factors. *International Journal of Computer Applications, 21* (6).

Soong. F., Rosenberg. A., Rabiner. L., & Juang. B. H. (1985). A vector quantization approach to speaker recognition. *Proceedings in the International Conference on Acoustic Signal Processing*, 387-390.

Sreedevi, N. (2012). *Frequency of Occurrence of Phonemes in Kannada*. Unpublished Project, Department of Speech-Language Sciences, All India Institute of Speech and Hearing, Mysore, India.

Sreevidya, M. S. (2010). *Speaker identification using Cepstrum in Kannada language.* Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

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.

Sukor, A., & Syafiq, A. (2012). *Speaker identification system using MFCC procedure and noise reduction method*. Doctoral dissertation, University Tun Hussein Onn, Malaysia.

Suman, S. & Hema. N. (2015). *Benchmark for speaker identification using MFCC on vowels following nasal continuants in Kannada*. Unpublished Dissertation, Department of Speech-Language Sciences, University of Mysore, Mysore, India.

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

Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies, 1*(1), 19-22.

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

Udrea, R. M., & Ciochina, S. (2003). Speech enhancement using spectral over-subtraction and residual noise reduction. In *Signals, Circuits and Systems in International Symposium of IEEE* (1), 165-168.

Wang, L., Ohtsuka, S., & Nakagawa, S. (2009). High improvement of speaker identification and verification by combining MFCC and phase information. *In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on Acoustics, Speech and Signal Processing*, 4529-4532.

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