Research Article

Effect of Noise Reduction Technique on Speaker Identification Using Mel-Frequency Cepstral Co-Efficients of Long Vowels

JAIISH(2016)
Vol 35, pp. 19-29

Pankaja K. R. and Hema N.

Affiliations

Department of Speech-Language Sciences, All India Institute of Speech and Hearing, Manasagangothri, Mysuru

Corresponding Author

Hema N.
Ph.D (Speech Language Pathology)
Lecturer in Speech Sciences,
Department of Speech-Language Sciences, All India Institute of Speech and Hearing,
Manasagangothri, Mysuru-570006.
Phone no: (0821) 2502254
Email: hema_chari2@yahoo.com

Key Words

Sound cleaner
Semi-automatic
Hypothetical
Distortion
Truncate

Abstract

Speech is always accompanied by noise when the speaker is talking in the environment. To improve the intelligibility of speech signal, noise should be reduced using noise reduction softwares. From the existing software the aim of the present study was to examine the effect of noise reduction technique on speaker identification using Mel Frequency Cepstral Co-Efficient (MFCC) on the long vowels in Kannada language. Ten Kannada speaking neuro-typical adults in the age range of 20-35 years (5 males and 5 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: Lab condition and Traffic condition. These samples were analyzed under two phases: Before noise reduction (BNR) and After noise reduction (ANR), using Sound Cleaner Software. Speech Science Lab Work bench software was used to extract MFCC for the truncated (PRAAT software) vowels. Results of the study revealed that in Lab condition, Traffic condition (BNR), Traffic condition (ANR), Lab condition verses traffic (BNR) and in Lab condition verses traffic condition (ANR), the vowel /i:/ is found to be better followed by /a:/ and /u:/ in the average percentage of correct speaker identification of the vowels. Overall results revealed vowel /i:/ is better for speaker identification. Hence, the 'sound cleaner' has a significant effect on percent speaker identification by reducing the influence of noise without majorly affecting the acoustical parameter of certain vowel considered for the present study.

Background

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 since long period (Atal, 1972). The telephone conversation has increased in the recent years. Due to the increased usage of mobile phones for conversational purpose, the crime rate is increasing drastically by misusing the same for many crime related activities like bomb threat, 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.

Among the biometric identifiers such as speech or handwriting, verification of individuals identity based on voice has significant advantages and practical utilizations because speech is a product of an underlying anatomical source, namely, the vocal tract and a resultant 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. The 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 context of conversational speech sample, the interesting one is the background noise. Among various factors that affect speaker recognition, background noise is one. Since the speaking environment is always associated with one or more types of noise, the considered speech sample may be accompanied with 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, 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:

Kalman filtering (Fingscheid, 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. 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 robustness of the signal (Bechler, Grimm & Kroschel, 2003).

It is also possible to ignore some parts of speech which is corrupted by background noise using the missing feature theory (Bonastre, Besacier & Fredouille, 2000). For example, consider a spectrum which has been passed through high-pass filter. If we assume that the first eight spectral magnitude features are below threshold and is labelled as ‘missing’. Once each spectral magnitude feature in a frame is labelled as present or missing, a computationally simple modification of probability models discards missing features and forms densities which would have been obtained by training without missing features.

Based on the same principal 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). A Gaussian Mixture Model (GMM) is defined as a parametric probability density function which is represented as a weighted sum of Gaussian component densities. GMMs are commonly used in biometric systems, such as vocal-tract related spectral features in speaker recognition systemsthus it has the competency of symbolizing a large class of sample distributions.

With reference to voice coding system, normally low bit rate of voice coding system will not have its own mechanism to reduce background noises from the target voice signal. This is due to constraint in the scope of voice coding systems and the complexity of the target signal which is a voice signal. Hence, it is very essential to include a specific method for removal of background noises. This process will happen by passing the voice signal to a pretreatment process. Here the background noises which disturb the identification process of the speech signals will be removed, in view of the fact that the presence of such random noises will degrade the voice signals (Yeldener & Rieser, 2000). Therefore in agreement with the view of removal of background noise a study was conducted by Rozeha and Adib (2008). It was recommended that the removal of any unwanted signal could be done by passing the entire sample through Digital Filter Design block of MATLAB (SIMULINK) software which serves as digital infinite-impulse response (IIR) bank-pass filter which is based on complex computational methods.

Apart from voice coding system and with reference to the computational complexity, Spectral Subtraction is relatively inexpensive. This Spectral subtraction method is used to remove background noise from voice recognition signal. According to Udrea and Coichina (2003), this method involves the basic principles of spectral subtraction method, wherein from the entire speech sample, the short term spectral magnitude of noise is subtracted and an attempt will be made to estimate the average signal and average noise measures. Following this, the same will be subtracted from each other. This results in improvement of signal to noise ratio and thus provide better quality of speech signal to carry out the process of speaker identification.

In addition to spectral subtraction method to remove background noise, the quality of speech signal can be improved by passing the signal through low-pass filter and by use of Fourier method for processing the signal. The processing of digital signal can be divided into FIR (finite-impulse response) and IIR (infinite-impulse response). A recursive filter has an output that is a function of an input sample and previous output samples and FIR is a non-recursive filter in which, the output is a function of input samples and is not a function of previous output samples. . In general, FIR filters have better performances in analyzing the signal but perform slower, because the process of Fast Fourier Transform takes a longer time. The IIR filter on the other hand, functions faster but has low performances (Gold, Morgan & Ellis, 2011). IIR filters are digital filters with infinite impulse response. Unlike FIR filters, they have the feedback (a recursive part of a filter) and therefore known as recursive digital filters. The IIR can be designed using different methods and however the commonly used filters are designed to be a low-pass filter (passing frequencies below some cut-off point). The FIR filters have linear phase characteristics, which is not typical of IIR filters. The FIR filters are the only choice when it is necessary to have linear phase characteristic. When the linear phase is not necessary, in other cases like speech signal processing, IIR filter is a good solution. Thus, the preference to choose IIR filters is better compared to FIR.

In continuation with the processing of digital signal, Darren (2001) in his study on 'Design of Speaker Recognition' proposed a method of removing background signal. First, signal was converted to frequency domain through the use of a shifted FFT. Then, using 3rd order Butterworth low-pass filter which was also an IIR filter, the higher frequency signal was removed. The cutoff frequency was selected to remove maximum noise signal, while still preserving the original shape of the signal. From the above mentioned methods of noise reduction techniques, it can be assumed that for example, if we consider the voice signal to be a phoneme, the acoustical parameter of the voice signal will be definitely altered with reference to the frequency and its range in correspondence to the filter settings used.

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, 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 machine method using 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 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 present era for speaker identification tasks and has been shown to yield tremendous results. Mel-frequency cepstrum is a cepstrum with its spectrum mapped onto the Mel- Scale before log and inverse fourier transform is taken. 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 are linearly spaced filters and logarithmically spaced filters. To incorporate the phonetically essential characteristics of speech, MFCCs would be used in speech signal. A series of calculation takes place which uses cepstrum with a nonlinear frequency axis following mel scale. To get Mel-Frequency Cepstrum, the speech signal will be windowed first using analysis window and then Discrete Fourier Transform will be computed. The main rationale behind MFCC is to mimic the behavior of human ears. 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 usefulness of Mel- Frequency Cepstral Coefficients (MFCC) on speaker identification.

Review

Generally, in most the 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 (Douglas O' Shaughnessy, 1987; Sigmund, 2003). Vowels have proven to be effective for characterizing individual speakers and been widely used for speaker recognition and in forensic analysis.

The foremost review with reference to the parameter considered for speaker identification is Mel-Frequency Cepstral Coefficients.A study using Mel-Frequency Cesptral Coefficients for feature extraction and vector quantization in security system based on speaker identification was conducted by Hasan, Jamil, Rabbani and Rahman (2004). 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 window. The hamming window yielded a better result when compared to triangular and rectangular window. Hamming window is the sum of rectangle and hanning windowand it is amplitude weighting of the time signal which is used with gated continuous signals that gives a slow onset and cut-off in turn to decrease the making of side lobes in their frequency spectrum. This window has similar properties to the Hanning window with the supplementary feature which suppresses the first side lobe and gives the best results for large signal. The study revealed that when codebook size was 1, speaker identification score was 57.14%. As 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.

To list out few Indian reviews for example, Jakhar (2009) studied the benchmark for text dependent speaker identification in Hindi language using cesptrum. 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 Kannada language by text independent speaker identification method using cepstrum in both direct and mobile recording conditions. The results of the study showed in direct speech and reading, vowel /u:/ had highest score (70 and 80%) and vowel /i:/ had highest score as (70 and 67%). Also the study quoted that for both the direct verse mobile recordings, for all vowels and for 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 Hindi language. Among 20 Hindi speakers 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 condition, 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 speaker verification system using MFCCs. In Kannada language. Ten Speakers participated in the studyand the material consisted of long vowels (/a:/, /i:/, and /u:/) in medial position occurring in five target 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 vowel /i:/ whereas, the accuracy of performance of vowel /i:/ was marginally better than /a:/ and /u:/. 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 filter bank. The author used 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 speaker recognition task. MFCC was also used to study the influence of the nasal co-articulation in Malayalam language samples and an attempt was made to obtain 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.

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

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

It is evident from these reviews that MFCCs is 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, till date there are limited studies on vowels as strong phonemes for speaker identification using semi-automatic methods in 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.

Aim

The aim of the present study is to investigate the effect of noise and noise reduction technique on speaker identification with reference to the parameter MFCC on the long vowels in Kannada language. The objectives of the study are to (a) evaluate the percent correct speaker identification with reference to the parameter MFCC on the long vowels in Kannada for lab recording conditions and traffic recording (embedded with or without noise) before and after the application of noise reduction technique, (b) compare speaker identification score with reference to the parameter MFCC on long vowels in Kannada for lab recording condition verses traffic recording (embedded with noise) before the application of noise reduction technique, and (c) compare the percent correct speaker identification with reference to the parameter MFCC on the long vowels in Kannada for lab recording verses traffic recording (probably embedded without noise) after the application of noise reduction technique.

Method

Participants

A total of 10 native speakers of Kannada with 5 males and 5 females in the age range of 20-35 years were considered for the study. The participants had a minimum of ten years of formal education in Kannada and were graduates and belonged to the same dialect of Kannada language usage (Mysuru dialect). The inclusion criteria for the participants were no history of speech, language, hearing problem, no associated psychological or neurological problems, and no reasonable cold or respiratory conditions at the time of recording and normal oral structure. Hearing sensitivity of all participants was screened using Ling sound test (Ling, 1979). Kannada Diagnostic Picture Articulation Test (KDPAT) (Deepa & Savithri, 2010) was administered by a Speech Language Pathologist to rule out any misarticulations present in the speech.

Procedure

Material

Hypothetical Kannada meaningful sentences (forensic speech sample) with commonly occurring long vowels /a:/, /i:/, /u:/ formed the material for reading task. The vowels were embedded in fifteen words within nineteen sentences. These target words formed the material for the present study and the same is listed in Table 1 of Appendix A.

Recording software 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. Trial I worked as a model setter for the following two trails. Participants were familiarized with the written material before recording began in laboratory condition and field condition individually. The same was recorded 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 the purpose of 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. For field condition, Olympus digital voice recorder (LS100) with attached dynamic microphone (Shure) was used for recording with the background noise of around 80 dB (A) (Kalaiselvi & Ramachandraiah, 2010). The traffic field recording was done using Olympus digital voice recorder. The recorded data was transferred from digital voice recorder to a computer using an USB cable. The samples were stored in .wav files so that the analysis could be carried out efficiently.

Analysis Software

Sound Cleaner: The individually recorded samples were analyzed under two phases: Phase I, the audio files of condition I and condition II was not subjected to any noise reduction algorithm. In Phase 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). 'Street Noise' scheme, one of the in built modules in Sound Cleaner software was used for the present study. 'Street noise' scheme consists a series of in built sub-modules such as, 'Input', 'Waveform-input', 'Broad band Filter', 'Dynamic Filter', 'Output/file' and 'Speaker'. 'Broad band Filter' was set at its default settings and for 'Dynamic Filter', which has options such as 'strong signal' and 'weak signal', where 'strong signal' remained as 'strong' and weak signal was 'weakened' and the threshold was kept at 4kHz. Data flows from the starting 'Input' process module (.wav file) to the final one (.wav file) through intermediate modules such as 'Broad band Filter' and 'Dynamic Filter' and thus the sample was processed and saved as output file.

To explain further, generally in the dynamic processing module there would be alteration in the dynamic range of signal. The common process of operations would be the 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). In the present study, after passing the signal through Sound Cleaner (with reference to the above selected inbuilt module). The samples before and after noise reduction were subjected to perceptual judgment to note the differences. Based on perceptual ratings, it was found that the selected in built module was effective in eliminating the embedded noise.

The samples of Phase I and Phase II stored in a separate folder in the CSL 4500 (original sampling frequency of 16kHz) were opened in PRAAT software (Boersma & Weenink, 2009) and down sampled to 8 kHz since the analysis could be done up to 4 KHz (frequency distribution of an individual's speech frequency ranges till 4 KHz). Of the three recordings, 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 as mentioned in the previous section. From the down sampled speech material, the long vowels /a:/, /i:/ and /u:/ in medial position of the target words were truncated from the wide band bar type of spectrograms using 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 1, speaker 1, first occurrence, vowel and first session was given the file name as “LB_ SPM1_ (thupaaki)_(a)_2.wav” and saved in a folder with the name SPM1.

Speech Science Lab (SSL) Work bench: This is Semi-Automatic vocabulary dependent speaker recognition software used in the present study to extract Mel-Frequency Cepstral Coefficients (MFCC) for the truncated (PRAAT software) vowels. Initially the file was specified using notepad in Workbench software and .dbs file, the extension of notepad file was created by specifying the phoneme, speaker, number of sessions and occurrences and was then segmented. Followed by this, the truncated samples for analysis were segmented to the workbench software. As soon as all files were segmented, the software select some samples randomly as trials. The trail/repetitions and utterances of each recording were randomized on 5:5 distribution by the software and were considered as test set and training set on equal distribution. Thus, the SSL Pro.V4 software was used to test the performance of distance based, semiautomatic speaker recognition system, which is vocabulary dependent. After training, 13 MFCC was selected and the samples for identification were tested. The software automatically generated the speaker identification threshold in terms of Euclidian Distance and thus, the correct percentage of speaker identification was calculated. This data was stored and the same procedure was repeated at least 15 times by randomizing the training and testing samples and the speaker identification thresholds was noted for the highest score and the lowest score. All the speech samples were non-contemporary, as all the recordings of the same person were carried out in two different conditions. Closed set speaker identification tasks were performed, in which the examiner was aware that the ‘unknown speaker’ is one among the ‘known’ speakers.

Results

The aim of the present study was to examine the effect of noise and noise reduction technique on speaker identification with reference to the parameter Mel-Frequency Cepstral Co-Efficients (MFCCs) on long vowels in Kannada language. The Euclidean distance of the samples used in test and reference samples of each speaker were calculated and was then tabulated as a distance matrix comparing all the speakers. Following this, the correct percentage of speaker identification scores were obtained. The same process was repeated for 15 times and among this the highest correct percentage of speaker identification (HPI) was noted. Thus, in the present study, the HPI ranged from 70% to 100%. The objective of the present study was speaker identification using noise reduction method. The speaker identification results could be explained with reference to the parameter MFCC on the long vowels in Kannada language. Thus, the effectiveness of vowels in speaker identification is the final implication with reference to the noise reduction method. Although with reference to the recording conditions and their comparisons, the results of the present study are descriptively discussed under five sections. 1) Lab condition. 2) Traffic condition. 3) Traffic condition followed with noise reduction technique. 4) Lab recording verses traffic recording preceding noise reduction technique. 5) Lab recording verses traffic recording followed with noise reduction technique.

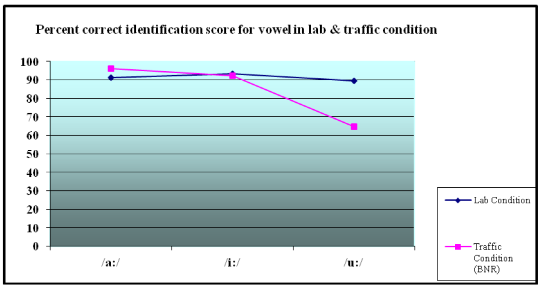

Figure 1: Percent correct speaker identification score for vowels of lab verse traffic condition .

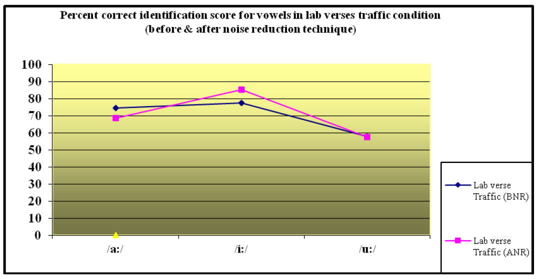

Figure 2: Percent correct speaker identification score for vowels of lab verse traffic condition.

1) Lab condition

The task under this section was to evaluate the percent correct Speaker Identification using MFCC on the long vowels in Kannada language for lab recording (test sample) verses lab recording (reference sample). Here the results revealed that the highest percent correct identification (HPI) was 100% for /a:/, /i:/ and /u:/ vowels. The frequency of the occurrence of HPI for the three vowels was 6, 11 and 6 respectively. On an average of 15 times randomization score of the percent correct speaker identification of three vowels /a:/, /i:/ and /u:/ was 91.33% (SD: 8.33), 93.33% (SD: 11.75) and 89.33% (SD: 13.34) respectively. This indicates /i:/ to be better followed by /a:/ and /u:/. The descriptive data of speaker identification scores obtained for all fifteen randomized trials for vowels is depicted in Table 1 of Appendix B.

2) Traffic condition

The task under this section was to evaluate the percent correct Speaker Identification using MFCC on the long vowels in Kannada language for traffic recording (test sample) verses traffic recording (reference sample). Where, these samples contained some amount of traffic noise embedded in it during analysis. Here the results revealed that the highest percent correct identification (HPI) was 100% for /a:/ and /i:/ and 90% for /u:/ vowel. The frequency of the occurrence of HPI for the three vowels was 9, 7 and 2 respectively. On an average of 15 times randomization score of the percent correct speaker identification of three vowels /a:/, /i:/ and /u:/ was 96% (SD: 4.07), 92.33% (SD: 7.98) and 64.66% (SD: 20.30) respectively. This indicates /i:/ has better speaker identification scores followed by /a:/ and /u:/. These findings are similar to the lab condition. The descriptive data of speaker identification scores obtained for all fifteen randomized trials for vowels are depicted in Table 2 of Appendix B.

On comparison among the average percent correct speaker identification score for lab verses traffic recording condition, the differences were observed only for the vowel /u:/ when compared to /a:/ and /i:/. The same is represented graphically in figure (1).

3) Traffic recording followed with noise reduction technique.

The task under this section was to evaluate the percent correct Speaker Identification with reference to the parameter MFCC on the long vowels in Kannada language for traffic recording conditions following the application of noise reduction technique (test sample) verses traffic recording conditions following the application of noise reduction technique (reference sample). Where, these samples were subjected to noise reduction scheme in the sound cleaner software. The results of this sample revealed high percent of correct speaker identification to be 100% for /a:/, /i:/ and /u:/. The frequency of the occurrence of HPI for the three vowels was 9th , 8th and 1st trail among the 15 randomized trail respectively. On an average of 15 times randomization, the percent correct speaker identification of three vowels /a:/, /i:/ and /u:/ was 94% (SD: 8.28), 92.66% (SD: 10.99) and 66% (18.43) respectively. This indicated /i:/ to be better identified followed by /a:/ and /u:/. The descriptive data of speaker identification scores obtained for all fifteen randomized trials for vowels is depicted in Table 3 of Appendix B.

4) Lab recording verses traffic recording preceding noise reduction technique.

The task under this section was to evaluate the percent correct Speaker Identification using MFCC on the long vowels in Kannada language for lab recording (reference sample) verses traffic recording (test sample) preceding noise reduction technique. Here, the lab sample was absolutely speech with no embedded noise, whereas the traffic samples contained some amount of traffic noise embedded in it during analysis. The results of this comparison revealed that the high percent of correct speaker identification for vowel /a:/, /i:/ and /u:/ to be 90%, 100% and 80% respectively. The frequency of the occurrence of HPI for the three vowels was 2, 1 and 1 respectively. On an average of 15 times randomization, the percent correct speaker identification of three vowels /a:/, /i:/ and /u:/ was 74.66% (SD: 12.45), 77.33% (SD: 15.79) and 58% (SD: 12.07) respectively. This indicated /i:/ to be better identified followed by /a:/ and /u:/. The descriptive data of speaker identification scores obtained for all fifteen randomized trials for vowels is depicted in Table 4 of Appendix B.

Lab recording condition verses traffic recording condition followed with noise reduction technique.

The task under this section was to evaluate the percent correct Speaker Identification using MFCC on the long vowels in Kannada language for lab recording (reference sample) verses traffic recording (test sample) with the application of noise reduction technique. Here, the lab sample was absolutely speech with no embedded noise, and the traffic samples containing some amount of traffic noise was removed with the sound cleaner software during the analysis. The results of this comparison revealed that the high percent of correct speaker identification for vowel /a:/, /i:/ and /u:/ to be 80%, 100% and 70% respectively. The frequency of the occurrence of HPI for the three vowels was 6, 1 and 7 respectively. On an average of 15 times randomization the percent correct speaker identification of three vowels /a:/, /i:/ and /u:/ was 68.66% (SD: 13.55), 85.33% (SD: 8.33) and 57.33% (SD: 14.37) respectively. This indicated /i:/ to be better identified followed by /a:/ and /u:/. The descriptive data of speaker identification scores obtained for all fifteen randomized trials for vowels is depicted in Table 5 of Appendix B.

On comparison among the percent correct speaker identification score in section 4 and 5, there is increment in the percent correct speaker identification scores in vowel /i:/ and slight decrement in vowel /a:/ after the application of noise reduction technique. The same is represented graphically in figure (2).

Table 1: ***Summary of the results-*** *Average and standard deviation (SD) of the percentage of speaker identification for condition I, II, III, IV and V*

Conditions/Comparisons	/a:/		/i:/		/u:/	
	Mean	SD	Mean	SD	Mean	SD
1 Lab v/s Lab	91.33	8.33	93.33	11.75	89.33	13.34
2 Traffic (BNR) v/s Traffic (BNR)	96	4.07	92.33	7.98	64.66	20.30
3 Traffic (ANR) v/s Traffic (ANR)	94	8.28	92.66	10.99	66	18.43
4 Lab v/s Traffic (BNR)	74.66	12.45	77.33	15.79	58	12.07
5 Lab v/s Traffic (ANR)	68.66	13.55	85.33	8.33	57.33	14.37

Average and standard deviation (SD) of the percentage of speaker identification

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

Discussion

The aim of the present study was to examine the effect of noise and noise reduction technique on speaker identification with reference to the parameter Mel-Frequency Cepstral Co-Efficient (MFCC) on the long vowels in Kannada language. Results of the study revealed that for (1). Lab condition, HPI is 100% for /a:/, /i:/ and /u:/ vowels. On an average the percent correct speaker identification of the vowels, the vowel /i:/ is found to be better followed by /a:/ and /u:/. For (2) Traffic condition (BNR- Phase I) HPI is 100% for /a:/ and /i:/ and 90% for vowel /u:/. On an average the percent correct speaker identification scores of the vowels revealed, the vowel /i:/ is to be better followed by /a:/ and /u:/. For the next (3). Traffic condition compared across traffic condition (ANR Phase II) the results of traffic condition revealed HPI to be 100% for /a:/, /i:/ and /u:/ vowels. On an average the percent correct speaker identification of the vowels, the vowel /i:/ is found to be better followed by /a:/ and /u:/. Following was (4). Lab condition compared across traffic (BNR), the results revealed HPI for /a:/, /i:/ and /u:/ vowels are 90%, 100% and 80% respectively. On an average the percent correct speaker identification of the vowels, the vowel /i:/ is found to be better followed by /a:/ and /u:/. The final was (5). Lab condition compared across traffic condition (ANR), the results revealed HPI for /a:/, /i:/ and /u:/ vowels are 80%, 100% and 70% respectively. On an average the percent correct speaker identification of the vowels, the vowel /i:/ is found to be better followed by /a:/ and /u:/.

Among all the above conditions, the vowel /i:/ is found to be better identified compared to vowel /a:/ and /u:/. From this result of the present study, it is clear that the vowel /i:/ is more effective in speaker identification with reference to the parameter MFCC for the samples (embedded with or without noise) before and after the application of noise reduction technique. The results of the present study are in support with the previous studies. To mention a few, Jakhar (2009), found that vowel /a:/ yielded better results in live recording and vowel /i:/ in mobile recording. Medha, (2010) found vowels /a:/ and /i:/ were useful in speaker identification in males. Chandrika, (2010) found better performance of vowel /i:/ compared to /a:/ and /u:/. In contrast to the present study Arjun, (2015) found vowel /a/ preceding nasals performed better compared to /i/ and /u/.

From the above discussions, it was clear that speaker identification scores were poorer in condition (Lab v/s Traffic (BNR)) and (Lab v/s Traffic (ANR)) compared to condition (Lab v/s Lab), (Traffic (BNR) v/s Traffic (BNR)) and (Traffic (ANR) v/s Traffic (ANR)). This could be because of the sound cleaner contributing a significant affect in reducing the influence of noise without majorly affecting the acoustical parameter of certain vowels. Interestingly it was found that there is slight increment in the percent correct identification scores in vowel /i:/ after the application of noise reduction technique. From the present study, it is observed that in a semi-automatic method of speaker identification the vowel /i:/ is considered to be the strongest phoneme which is not majorly affected by the influence of noise and the noise reduction technique. This was with reference to the sound cleaner software. In general, the reasons for this difference in the results could be as follows:

Reason (1) 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 other audio environment since there will be complete silence in the lab.

However, in traffic condition it is a different situation. The recording equipment does focus on certain audio streams (specialized microphone) 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.

Reason (2). 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 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 is estimated and subtracted from each other (Udrea & Coichina, 2003). Hence, there might be a chance of signal getting distorted.

Reason (3). The quality and accuracy of spectral picture is the most important factor for both experts and automatic systems (Barinov, Koval, Ignatov, 2010; Goldstein, 1975; 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, nonlinearity of frequency response and sampling frequency and bit rate. This might have contributed for poor percent correct identification score of traffic condition in the present study.

Therefore, to conclude the study, the outcome after the application of noise reduction technique on speaker identification for traffic noise has not shown significant effect on the acoustical characteristics of the speech sounds. The speech sounds considered are being vowels only and the average percent correct speaker identification scores was better for vowel /i:/ followed by /a:/ and /u:/. Hence to conclude from the present study, the vowel /i:/ acts as a better cue for speaker identification irrespective of before and after the application of noise reduction technique As an implication from the present study, there is also a future need to study the effect of other speech sounds in speaker identification under other noise reduction technologies.

Conclusion

The present study aimed to investigate the effect of noise and noise reduction technique on speaker identification with reference to the parameter MFCCs on the long vowels in Kannada language. Results revealed vowel /i:/ is better for speaker identification as there was slight increment in the percent correct identification scores in vowel /i:/ after the application of noise reduction techniquewhich indicates that the phonemic cue has not been altered much after noise reduction for vowel /i:/. Whereas in vowels /a:/ and /u:/, there were changes observed. Therefore, the sound cleaner has a significant affect in reducing the influence of noise without majorly affecting the acoustical parameter of certain vowel. Though, this study is a preliminary study which stepped to see the effect of noise reduction technique using Sound Cleaner, further studies have to be conducted to check the effect of Sound Cleaner or other technology related to noise reduction techniques in reducing the background noise with reference to certain variables like increased participants, stimulus in different languages and considering the same in different environmental noise.

Acknowledgments

The authors are grateful to the current Director, Dr. S. R. Savithri, Director, AIISH, Mysore for permitting to conduct the study and extending support. This article is a part of the ARF project titled ‘Effect of noise and noise reduction technique on speaker identification’. The authors express their sincere gratitude to all the participants of the study.

Conflict of Interest: NIL

Source of Funding: AIISH Research Fund - ARF 2015-16.

References

Arjun. M. S. (2015). *Benchmark for speaker identification using MFCC on vowels preceding the nasal continuants in Kannada.* Dissertation submitted to the University of Mysore, Mysore, India.

ASHA Monographs Number 16 (American Speech and Hearing Association, Washington D.C.).

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.* Dissertation submitted to the 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.

Boersma, P., & Weenink, D. (2009). *Praat: doing phonetics by computer* (Version 5.1. 12)[Computer program]. Retrieved August 4, 2009.

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

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.* Dissertation submitted to the University of Mysore, Mysore, India.

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.

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

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

Stan, S., Fingscheidt, T., & Beaugeant, C. (2003). *An evaluation of VTS and IMM for speaker verification in noise.* In Eighth European Conference on Speech Communication and Technology.

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

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

Goldstein, U.G. (1975). Speaker-identifying features based on formant tracks. *Journal of Acoustical Society of America, 59*(1), 176-182.

Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and audio signal processing: processing and perception of speech and music.* John Wiley & Sons.

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. *IEEE transactions on speech and audio processing, 2*(4), 578-589.

Hollien (1990). *The Acoustics of Crime: The New Science of Forensic Phonetics.* New York and London: Plenum Press. xiv+370 pp.

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

Kalaiselvi, R., & Ramachandraiah, A. (2010, August). 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.

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

Lakshmi, P., and 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.

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.

Medha, S. (2010). Benchmark for speaker identification by Cepstrum measurement using text-independent data. Dissertation submitted to the University of Mysore, Mysore, India.

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

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* (pp. 744-767).

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of speech recognition.* Englewood cliffs. NJ: PTR Prentice Hall.

Ramya. B.M. (2011). Bench mark for speaker identification under electronic vocal disguise using Mel Frequency Cepstral Coefficients. Dissertation submitted to the University of Mysore, Mysore, India.

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. Dissertation submitted to the University of Mysore, Mysore, India.

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

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.

Sreevidya, M. S. (2010). Speaker identification using Cepstrum in Kannada language. Dissertation submitted to the 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.

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

Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America, 51*(6B), 2044-2056.