FORENSIC SCIENCE JOURNAL SINCE 2002

Forensic Science Journal
2011;10(1):1-10
Available online at:fsjournal.cpu.edu.tw

Speaker identification using cepstrum in Kannada language

C. Srividya[1], M.Sc.,; S.R. Savithri[2*], Ph.D.

[1]Physics Section Forensic Science Laboratory Bangalore – 560068.
[2]Department of Speech Language Sciences All India Institute of Speech and Hearing, Mysore-570006.

Received: May 19, 2011 / Accepted: June 16 , 2011

Abstract

Speaker identification in forensic cases involves careful estimation of the features that are more specific to speaker. The anatomical differences in vocal tract depict speaker related differences. Cepstrum has been investigated as a possible parameter for speaker identification. Fundamental frequency obtained from cepstral coefficients indirectly depicts the shape and size of vocal tract. In the present study quefrency and amplitude was extracted using cepstrum spectral analysis technique under various recording conditions. 30 normal males were selected. For reading intention, four paragraphs were selected having long vowels /a:/ , /i:/ and /u:/ embedded in the words of the paragraphs of Kannada passage. To elicit spontaneous speech from the subjects, six Kannada words having long vowels in the medial position were considered. Subject's speech and reading paragraphs were recorded in field conditions to suit realistic forensic situations. Using CSL-4500 software, cepstral coefficients quefrency and amplitude were extracted for the long vowels /a:/, /i:/ and /u:/. Extracted parameters were normalized and Euclidian distances between speakers were measured. The results indicated that the percent correct identification was above chance level for direct vs. direct and mobile vs. mobile recording (DS Vs. DS = 68%, MS Vs. MS = 64%, DR Vs. DR = 75%, and MR Vs. MR = 62%). Results of 4-way repeated measures ANOVA revealed that significant difference between speakers on quefrency and interaction between speaking style, recording, set and vowels on quefrency. Results of paired t-test reveals the significant difference between direct and mobile recording for long vowels /a:/, /i:/ and /u:/ on quefrency.

Keywords: Forensic, Cepstrum, Quefrency, Amplitude, Fast Fourier Transform (FFT), Euclidean distance, Correct Identification.

Introduction

Forensic Speaker identification is obtaining an expert opinion in the legal process as to whether speech recordings to be compared are of same person. Forensic speaker identification is to identify an unknown speaker whose voice has been recorded during the committing of a crime, for example a bomb threat, ransom demand, sexual abusement, child abusement, terrorist attack, hoax emergency call or drug deal. The expert compares the incriminating recording with suspect speech samples with a view to identify the perpetrator or eliminating the suspect. Speaker identification is deciding if a speaker belongs to group of known speaker population. Speaker verification is verifying the speaker from his previous recordings.

Speaker identification has been used in a variety of criminal cases, including murder, rape, extortion, drug smuggling, wagering-gambling investigations, political corruption, money-laundering, tax evasion, burglary, bomb threats, terrorist activities and organized crime activities. Forensic acoustic analysis also involves tape filtering and enhancement, tape authentication, gunshot acoustics, reconstruction of conversations and the analysis of any other questioned acoustic event.

There are three methods of speaker identification [1] - (a) speaker identification by listening (subjective

*Corresponding author: savithri_2k@yahoo.com

method), (b) speaker identification by visual examination (subjective method), and (c) speaker identification by machine (objective method). In the first method, the expert hears the voices and decides whether two voices belong to the same person or not. In the second method spectrograms of two speech samples are visually matched to identify a speaker. The third method can be semiautomatic or automatic. In semiautomatic methods, features of the voice signal are extracted and decision is made by the experimenter. In automatic methods, features as well as the final conclusion is made by the machine.

Previous study

For speaker identification, features used were pitch contour[2], first and second formant frequencies[2-7] higher formants [8], Fundamental frequency [9] , Linear Prediction coefficients [10] , Cepstral Coefficients & Mel Frequency Cepstral coefficients [11,12] , Long term average spectrum [13] and Cepstrum [14,15] have been used in the past.

The word "Cepstrum" is an anagram (type of word play) of the word "Spectrum". This word was first introduced by Bogert et al in 1963 to denote the data in frequency domain to time domain. New data set called quefrencies in units of seconds indicates the variations in the frequency spectrum. The cepstrum is a measure of the periodicity of a frequency response plot. Cepstral analysis performs deconvolution of the speech signal by use of FFT (Fast Fourier Transform) techniques. Deconvolution is a process of separation of signal from an impulse response that has been convolved with it [16].The ***cepstrum*** is a common transform used to gain information from a person's speech signal. It can be used to separate the ***excitation*** signal (which contains the words and the pitch) and the transfer function (which contains the voice quality). The cepstrum spectral analysis is a non invasive technique of extraction of pitch from voice signal. Pitch is reciprocal to the fundamental period of the vocal details of the speaker.

Cepstrum has been investigated as a possible parameter for speaker identification. Luck[15] reports 94 % verification accuracy for 10 subjects using same sentence in English Language and it is an automatic approach of speaker verification. In Atal's [17] study, the speech data consisted of 60 utterances consisting of six repetitions of the same sentence spoken by 10 speakers. Among all the parameters investigated, the cepstrum was found to be the most effective providing an identification accuracy of 98 % for speech of 0.5 sec duration. However, this study was restricted to 0.5 sec duration material with only 10 subjects. Also, this was a text dependent method. An effort using text dependent method of telephonic speech for ten subjects was done by Furui [18] who obtained an error rate of 0.2% for 3 sec speech duration. Li & Wrench[14] investigated cepstrum in text independent speech in laboratory condition for 11 subjects and obtained identification accuracy of 69 % for 3 sec duration. Higgins & Wohlford [19] studied cepstrum for speaker population of 11 using text independent method in laboratory condition and obtained verification accuracy of 80% for 2.5sec speech. Che & Lin [20] reported identification accuracy of 94.44% for 2.5sec in text dependent method and office recordings for 138 speakers. The above mentioned studies include only automatic method and either speech or reading recorded directly on to the recorder. Jakkar [21] developed benchmark for speaker identification using cepstrum for speaker population of 20 in Hindi language (Jaipur dialect) and obtained identification accuracy of 88.33% for direct reading vs. direct reading , 81.67% for mobile reading vs. mobile reading and 78.3% for direct vs. mobile reading.

All these studies have used either speech or reading material and the number of participants is small except in Che & Lin [20] study. It is a well known fact in forensic that the two samples may not be reading or speech. One may be speech and the other may be reading. Therefore, it is necessary to examine if cepstrum provides high percent of identification if a speech sample is compared with reading sample. In this context the present study investigated speaker identification using cepstrum in Kannada language in field conditions. The aims of the study were multifold and as follows:

(a) To compare cepstrum in speech and reading sample.
(b) To compare cepstrum in direct and mobile recording.
(c) To derive a benchmarking for cepstrum.

Materials and methods

Subjects

Thirty males in the age range of 35-55 years

participated in the study. All the speakers knew to read and write Kannada language. None of the speakers had any complaints of speech and or hearing problems. All speakers were free from respiratory infections at the time of recording. Written consent was obtained prior to recording.

Recording procedure

Four paragraphs from a Kannada passage [22] and six Kannada words each having a long vowel (/a:/, /i:/, or /u:/) in the medial position formed the material. Table1 shows six Kannada words used for eliciting speech among subjects. Paragraphs were used for reading and words were used to elicit spontaneous speech. Paragraphs and words were written on cards. The card bearing the material was displayed to each subject and voice was recorded simultaneously from mobile and digital recorder. To create realistic forensic situation the recordings were done at different environment at subject's convenience and in noisy environment. Recordings were done in two conditions - direct and mobile phone. A digital tape recorder (Olympus voice recorder, WS-100) and two mobile phones (LG and Maxx company mobile, MX463) were used for recording. Digital tape recorder was kept at a distance of 10 cm from the mouth of the subject. Initially the subjects were instructed to receive the call made by the experimenter using the LG mobile phone. The experimenter used the Maxx company mobile (MX463) to make the call. DOCOMO network was used for both mobile phones. After the subject received the call they had to start reading the paragraphs in a natural manner. Following this they made at least two sentences using the 6 words visually presented to them. The reading and speech were recorded using the Maxx company mobile by the experimenter.

Table 1. words used to elicit speech

Sl. NO.	Word with vowel In bold italics
1)	bh**a**:rata
2)	h**a**:sana
3)	s**ri**:rangapattana
4)	hale:b**i**:d.u
5)	bengal**u**:ru
6)	mais**u**:ru

Analyses

The recorded reading and speech material from the mobile and digital recorder was transferred on the computer memory. The transferred digital data was converted to wave file format using Adobe Audition. The audio files were down sampled from 44 kHz to 8 kHz to suit the extraction of cepstrum. The digitized wave file format of the audio files of different speakers will be stored separately. Key words containing long vowels were verified for correct production. The steady state of the vowels was displayed as waveforms on the computer screen and cepstrum (quefrency and amplitude) was extracted at six points for each vowel using CSL-4500 (Kay Pentax, New Jersey). Figures 1, 2, 3, 4, 5 and 6 represent pitch extraction.

Cepstral coefficients extraction using CSL- 4500 software is mentioned below:

(i) The key words were displayed as waveform and wide band spectrogram.

(ii) The steady state portion of each vowel was selected from the spectrogram and displayed as waveform on the window A.

(iii) To obtain Fast Fourier Transform (FFT) for selected waveform of the vowel, FFT settings (Analysis size: 512 poiwnts, window weighting: Hamming, Pre emphasis : 1.0) were performed. Cursor was placed at the steady state portion of the selected vowel belonging to the key word and Fast Fourier Transform (FFT) was extracted on window B.

(iv) The spectrum of spectrum, i.e cepstrum of the Fast Fourier Transforms (FFT) on window B was extracted in window C and cepstral coefficients quefrency in ms (milliseconds) and amplitude in dB (Decibel) was noted.

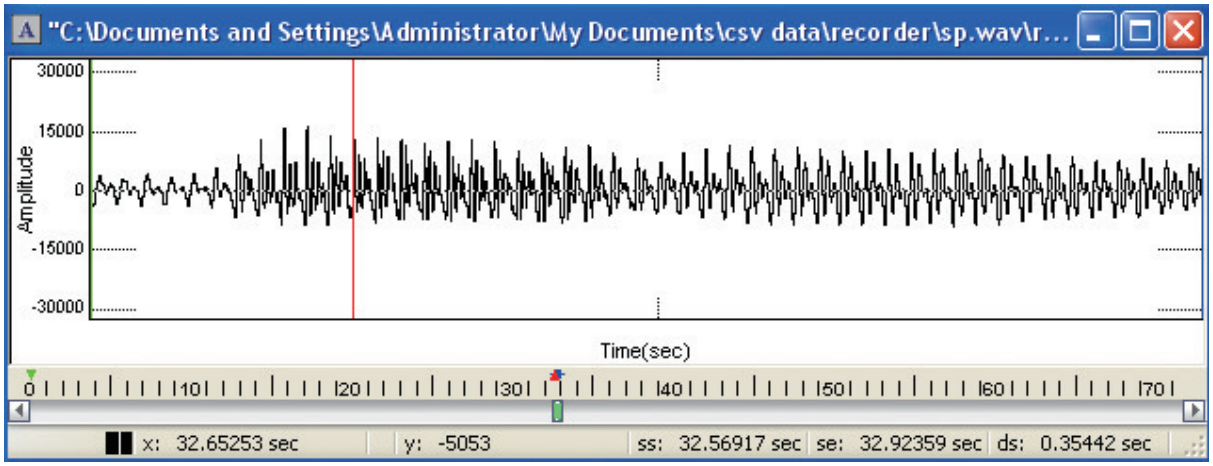

Fig.1waveform for key word /ba:ri/ in direct reading.

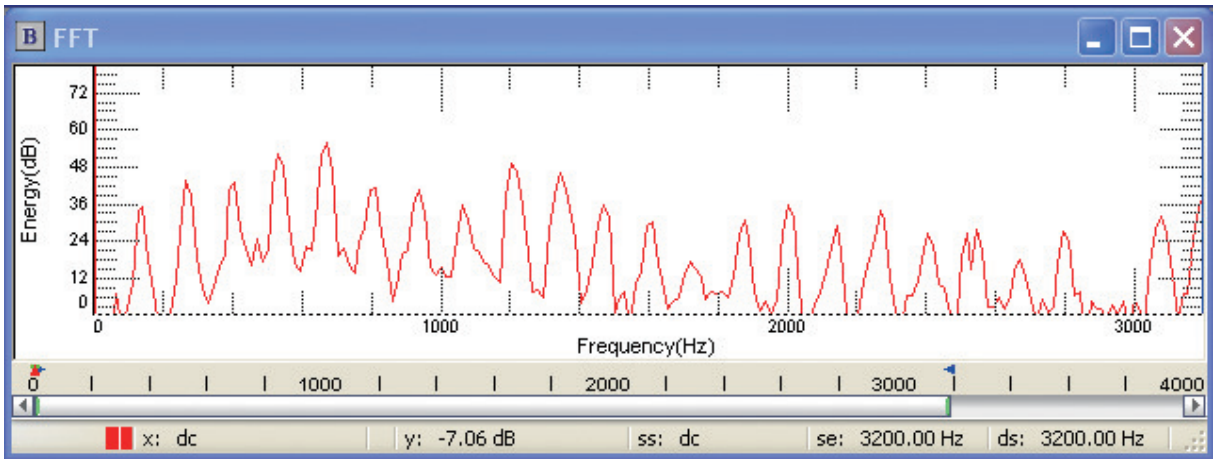

Fig.2FFT for vowel /a:/ at cursor point in direct reading.

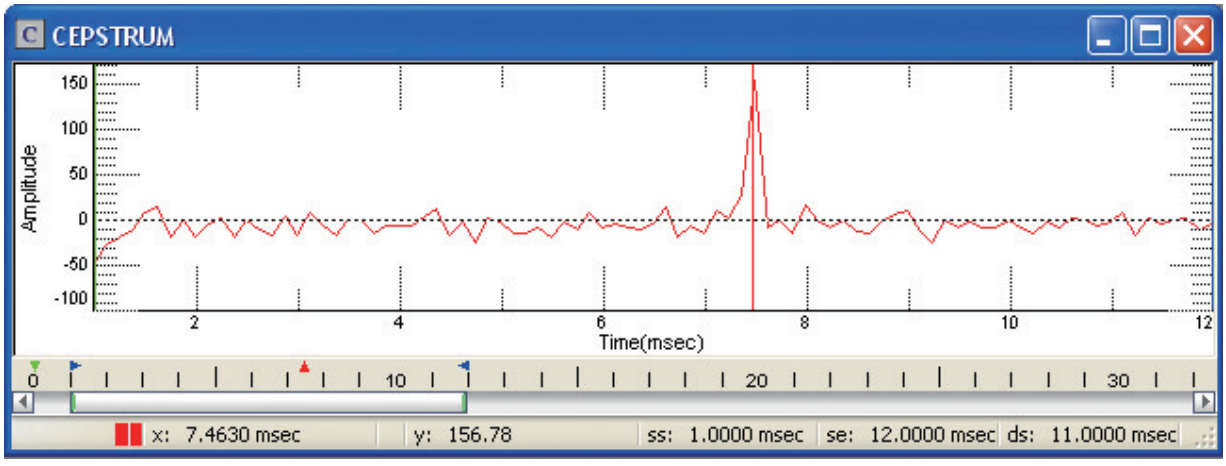

Fig.3cepstrum for vowel /a:/ in direct reading.

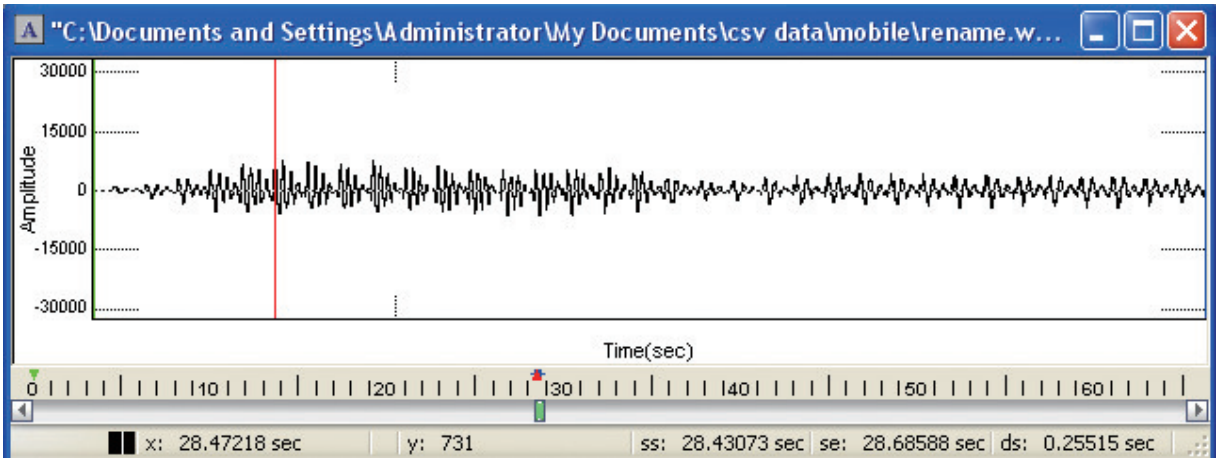

Fig.4waveform for key word /ba:ri/ in mobile reading.

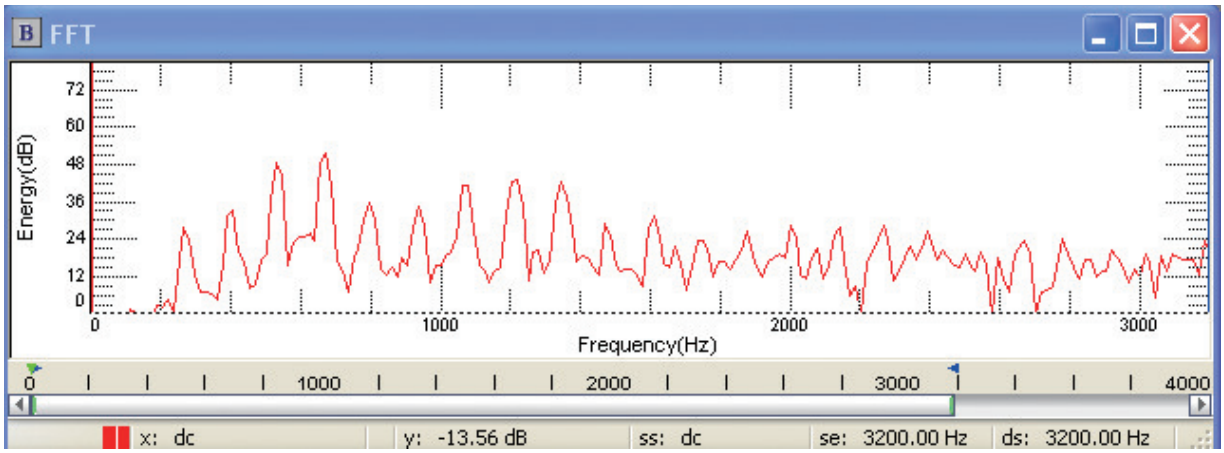

Fig.5FFT for vowel /a:/ at cursor point in mobile reading.

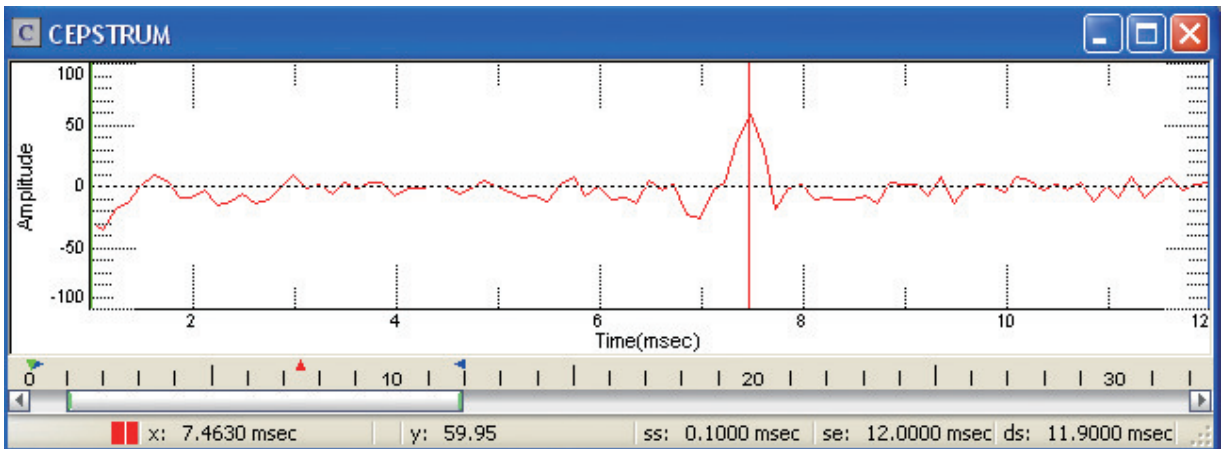

Fig.6cepstrum for vowel /a:/ in mobile reading.

Extracted quefrency and amplitude for three long vowels, four recording conditions and thirty speakers were normalized and Euclidian distances were calculated within and between speakers. 50% of each subject data was used as training set and the remaining as test set. Euclidian distances between the training set of one subject and the test sets of the same subject and other subjects were calculated. Speaker with least Euclidian distance was identified. Speakers were grouped in to sub groups of 30, 20, 10, and 5. Percent identification was calculated based on Euclidian distance.

All possible comparison were studied depending on the recording conditions, they are

1) Direct speech vs. Direct speech (DS vs. DS)
2) Mobile speech vs. Mobile speech (MS vs. MS)
3) Direct reading vs. Direct reading (DR vs. DR)
4) Mobile reading vs. Mobile reading (MR vs. MR)
5) Direct speech vs. Mobile speech (DS vs. MS)
6) Direct reading vs. Mobile reading (DR vs. MR)
7) Direct speech vs. Direct reading (DS vs. DR)
8) Direct speech vs. Mobile reading (DS vs. MR)
9) Mobile speech vs. Mobile reading (MS vs. MR)
10) Mobile speech vs. Direct reading (MS vs. DR)

Statistical Analysis

Statistical analyses were done using commercially available SPSS 10.0 software. 4 –way repeated measure ANOVA was used to determine the significance difference between set (training and test), vowels (a:, i:, and u:), recording condition (direct and mobile) and speaking style (speech and reading). Further paired t-test was performed to find significant difference between conditions.

The results of 4 – way repeated measure ANOVA indicated significant differences between subjects on quefrency. {Training and test set: [F (29, 1) 9.288, $p<0.00$]; style: [F (29, 1) = 35.726, $p<0.00$]; vowel: [F (29, 2) = 8.502, $p<0.00$] }. Significant interaction between set * type * style [F (1, 29) = 13.013, $p<0.00$], type * vowel [F (2, 58) = 12.770, $p<0.00$], style * vowel [F (2, 58) = 8.288, $p<0.00$] , type * style * vowel [F (2, 58) = 5.173, $p<0.00$] and set * type * style * vowel [F (2, 58) = 22.012, $p<0.00$] } was noticed. No significant difference within subjects was observed on amplitude. However, interaction between set * style [F (1, 29) = 7.759, $p<0.00$], set * vowel [F (2, 58) = 14.546, $p<0.00$], type * vowel [F (2, 58) = 6.202, $p<0.00$], set * type * vowel [F (2, 58) = 13.739, $p<0.00$], set * style * vowel [F (2, 58) = 6.417, $p<0.00$] and type *style * vowel [F (2, 58) = 9.870, $p<0.00$]} was noticed.

Results and Discussions

The results indicated that the benchmarking depended on the recording conditions, number of speakers and vowels. Percent correct identification was above chance level for DS vs. DS, DR vs. DR, MS vs. MS, MR vs. MR and DS vs. MS conditions. However, percent correct identification was below chance level across other conditions. Table 2 shows the percent correct identification of speakers across conditions. Figure 7 show correct identification based on Euclidian distance for DS vs. DS condition on long vowel /a:/ , Figure 8 show incorrect identification based on Euclidian distance for DS vs. DS condition on long vowel / a:/ Figure 9 show percent of correct identification for vowels across conditions. Figure 10 show percent of correct identification for same and different recording conditions. Figure 11 show mean percent of correct identification for three long vowels in same and different recording conditions.

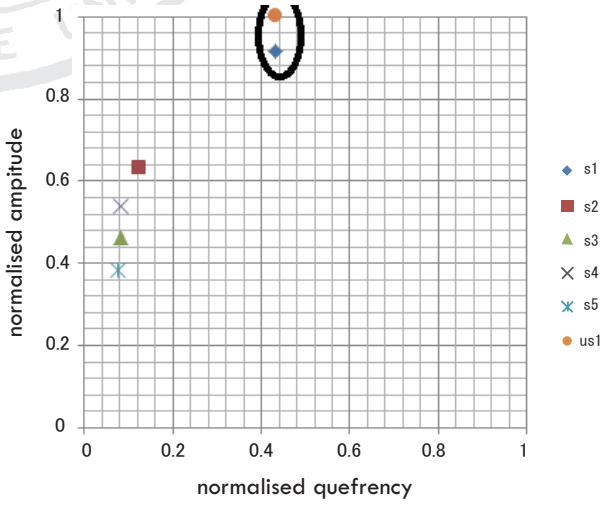

Fig.7 correct identification for 5 subjects-→ DS vs. DS .

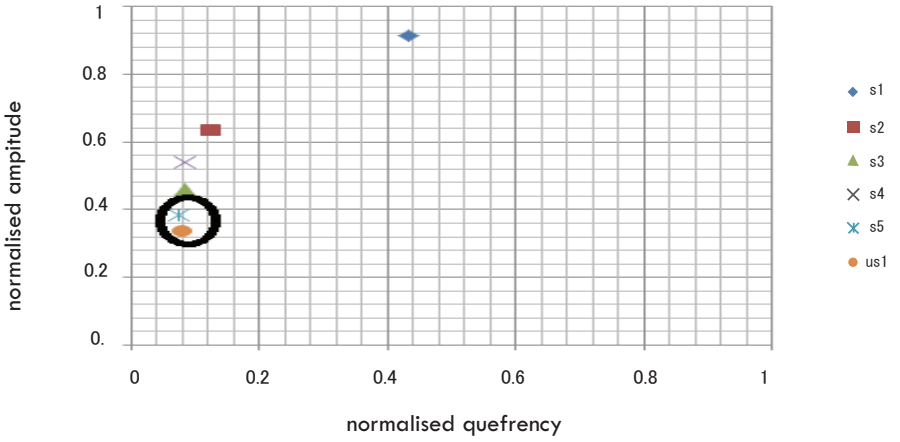

Fig.8 Incorrect identification for 5 subjects-→ DS vs. DS

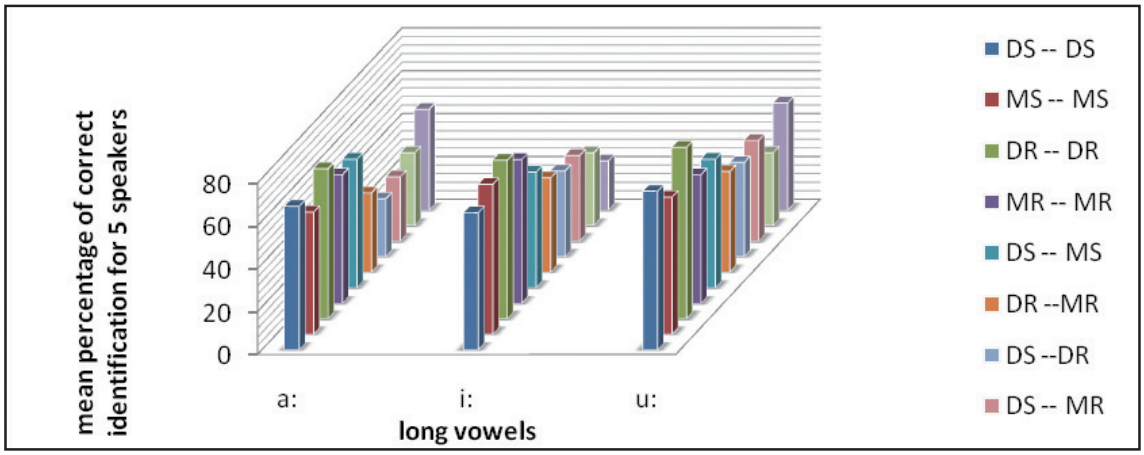

Fig.9 Percent correct identification for vowels /a:/, /i:/ and /u:/ across conditions.

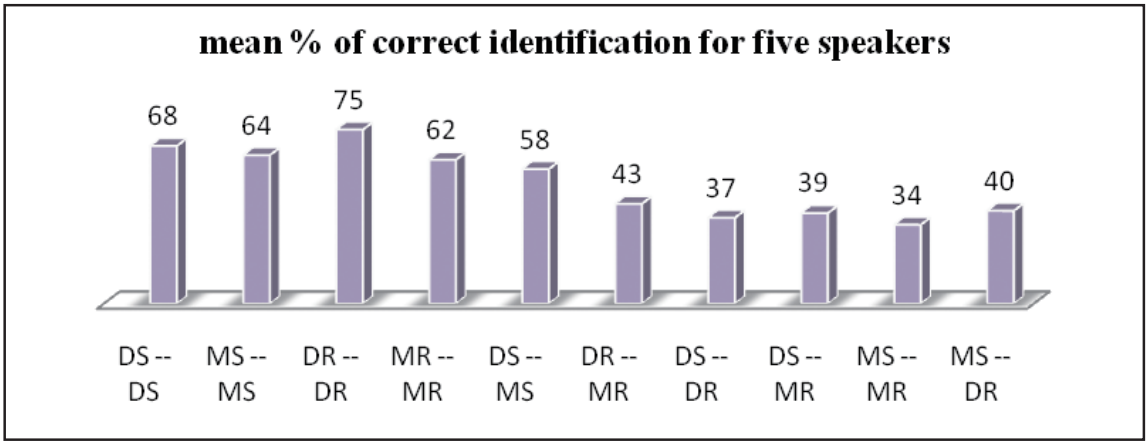

Fig.10 Percent of correct identification for same and different recording conditions.

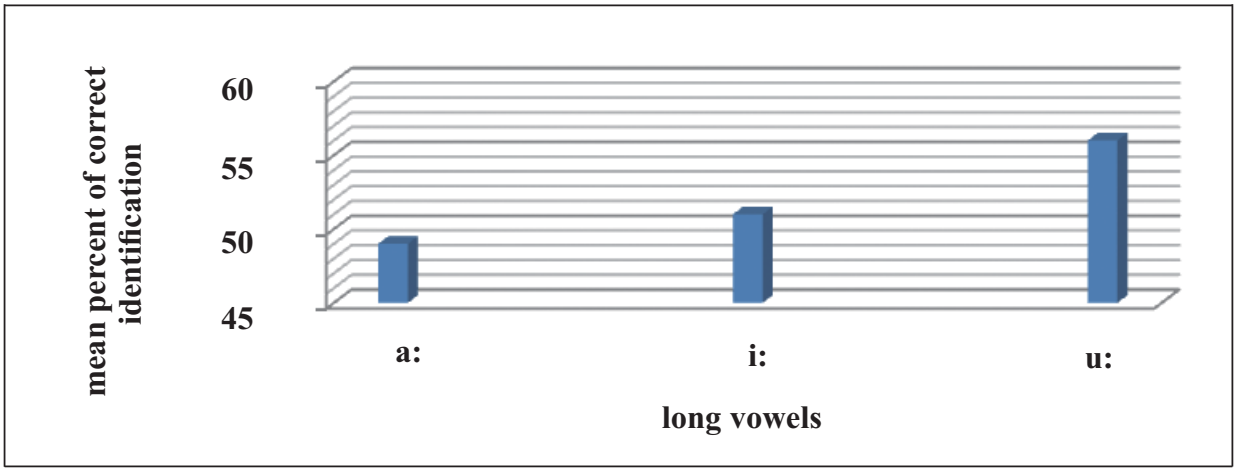

Fig.11 Mean percent of correct identification for three long vowels in same and different recording conditions.

Table 2. Mean Percent correct identification of vowels /a:/, /i:/ and /u:/ for five speakers for same and across conditions.

Sl. No.	conditions	%Correct identification			
		/a:/	/i:/	/u:/	Mean
1	DS -- DS	67	64	74	**68**
2	MS -- MS	57	70	64	**64**
3	DR -- DR	70	74	**80**	75
4	MR -- MR	60	67	60	**62**
5	DS -- MS	60	54	60	**58**
6	DR -- MR	37	44	47	43
7	DS -- DR	27	40	44	37
8	DS -- MR	30	40	47	39
9	MS -- MR	34	34	34	34
10	MS -- DR	47	23	50	40

The results indicated several points of interest. ***First of all, the percent correct identification was above chance level for direct vs. direct and mobile vs. mobile recording (DS Vs. DS = 68%, MS Vs. MS = 64%, DR Vs. DR = 75%, and MR Vs. MR = 62%).*** Of these, DR Vs. DR condition had highest percent correct identification. The results are not in consonance with the results of Luck [15] who used cepstral measurement and reported error rates of 6% to 13%, Wolf [8] who used F0, features of vowel and nasal consonant spectra, estimation of glottal source spectrum slope, word duration, and voice onset time and reported 2% error rates, Atal [1] who examined the temporal variations of pitch in speech as a speaker identifying characteristics in 10 speakers and reported a percentage of correct identifications of 97%. However, it is in consonance with the results of Atal [17] who determined twelve predictor coefficients in 10 speakers and reported that the cepstrum was found to be the most effective parameter, providing an identification accuracy of 70% for speech 50 msec in duration, which increased to more than 98% for a duration of 0.5 sec. Using the same speech data, the verification accuracy was found to be approximately 83% for a duration of 50 msec, increasing to 98% for a duration of 1sec. It is also not in consonance with the results of Doddington[23] who used six spectral/time matrices located within a test phrase in 50 "known" speakers and 70 "casual impostors" (including 20% female speakers in each session) and reported a rejection rates of 5% and 15%, Furui [18] who used the time pattern of both the fundamental frequency and log-area-ratio parameters and the other used several kinds of statistical features derived from them and reported an accuracy of 95%, and Jakkar [21] who reported 88.33% correct identification

using ceptrum in Hindi male speakers. The reasons for this might be the differences in the language and subjects. In the current study recordings were done in noisy conditions whereas in other studies it was in lab conditions.

Second, high vowels /i:/, and /u:/ had higher percent correct identification compared to vowel / a:/. Vowels /u:/, and /i:/ had higher percent correct identification in direct and mobile recording, respectively. High vowels have higher F0 compared to low vowels. Vowels /u:/, and /i:/ had highest and lowest mean normalized quefrency in direct and mobile recording, respectively in the present study. Quefrency is inversely proportional to F0. The vowels which have highest and lowest normalized quefrencies are identified better than vowel /a:/ which is in between.

Third, percent correct identification increased as the number of speakers decreased. This is in consonance with the results of Hollien [24] who also reported decrease in error rate with increase in the number of subjects Jakkar [21] and Glenn and Kleiner [25]. However, the percent correct identification in their studies was higher.

Conclusions

The present study has contributed to the area of speaker identification. Cepstral measurements quefrency and amplitude can be used if one is comparing DS Vs. DS, MS Vs. MS, DR Vs. DR, and MR Vs. MR. The investigating officers can be educated on the type of recording to be done and the vowels to be recorded. Further, the present study was restricted to Kannada and a specific network (DOCOMO). Future research on other networks and handsets, other Indian languages, disguised conditions, subsets of speakers, non-contemporary samples, and text-independent samples is warranted.

In most of the forensic cases Direct Vs. Mobile recording or Speech Vs reading are encountered. In such cases the results of the present study will throw light on whether speech and reading can be compared and whether Direct Vs Mobile recording can be compared. The investigating officer, while collecting the control recording for comparison, has to keep in mind that the benchmarking obtained in the study was above chance level for similar type of recording environment.

Acknowledgments

The first author expresses sincere thanks to her mother Smt S.H. Bhagyamma for all her support and motivation provided throughout the study.

References

1. Markel, J. D., & Davis, S. B. (1979), Text independent Speaker Recognition from a Large Linguistically Unconstrained Time spaced Data Base, IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-27, 74-82.
2. Atal, B. S,(1972), Automatic speaker recognition based on pitch contours, The Journal of the Acoustical Society of America, Vol. 52, 1687-1697.
3. Hollien, H (1990), The acoustics of Crime, The New Science of Forensic Phonetics, Plenum, Nueva York.
4. Kuwabara, H. & Sagisaka, Y., (1995), Acoustic characteristics of speaker individuality: control and conversion, Speech Communication, 16, 165-173.
5. Lakshmi, P., & Savithri.S. R (2009), Bench mark for speaker Identification using Vector F1 & F2, Proceedings of the international symposium, Frontiers of Research on Speech & Music, FRSM-2009 , 38 - 41.
6. Nolan, F (1983), The phonetic Bases of Speaker Recognition, Cambridge University press, Cambridge.
7. Stevens, K. N. (1968), Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material, The Journal of the Acoustical Society of America, Vol.44: 1596–1607.
8. Wolf, J. J. (1972), Efficient acoustic parameter for speaker recognition, The Journal of the Acoustical Society of America, 2044–2056.
9. Atikinson, E. J. (1976), Inter and Intra Speaker variability in Fundamental voice frequency, The Journal of the Acoustical Society of America, 440-445.
10. Soong, F., Rosenberg, A. E., Rabiner, L., & Juang, B.H. (1985), A vector Quantisation Approach to Speaker Recognition, In International Conference on Acoustics, Speech and Signal Processing in Florida, IEEE, 387-390.

11. Rabiner, L., & Juang, B.H. (1993), Fundamentals of Speech Recognition, Prentice Hall PTR.
12. Reynold, D.A. (1995), Speaker Identification and verification using Gaussian mixture speaker models, Speech Communication, 17, 91-108.
13. Kiukaanniemi, H., Siponen, P., & Mattila, P. (1982), Individual Differences in the Long term Speech Spectrum, Speech Communication, 21-28.
14. Li, K. P., & Wrench, E. H. (1983), Text Independent Speaker Recognition with short Utterances, In international Conference on Acoustics, Speech and Signal Processing in Boston, IEEE, 555-558.
15. Luck, J. E. (1969). Automatics speaker verification using cepstral measurements.The Journal of the Acoustical Society of America, 46, 1026-1032.
16. Bogert. B.P, M.J. R. Healy and J.W. Tukey, (1963), The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Psuedo-Autocovariance, Cross-cepstrum and Saphe Cracking, Proceedings of the Symposium on Time Series Analysis.
17. Atal, B. S, (1974), Effectiveness of Linear prediction characteristics of the speech wave for Automatic Speaker Identification and Verification, The Journal of the Acoustical Society of America, Vol. 55, 1304-1312.
18. Furui, S. (1981), Cepstral Analysis Technique for Automatic Speaker Verification, IEEE Transactions on Acoustics, Speech and signal Processing, Vol-29, 254-272.
19. Higgins, A., & Wohlford, R. E. (1986), A new method of text Independent Speaker Recognition, In International Conference on Acoustics, Speech and Signal processing in Tokyo, IEEE ,869-872.
20. Che, C., & Lin, Q., (1995), Speaker recognition using HMM with experiments on the YOHO database, In EUROSPEECH, 625-628.
21. Jakkhar, S. S. (2009), Bench mark for speaker Identification using Cepstrum, Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore.
22. Santosh (2006), Personal communication.
23. Doddington.G (1971), A method of speaker verification, The Journal of the Acoustical Society of America, Vol-49, p-139(A).
24. Hollien, H. (2002). Forensic Voice Identification. San Diego, CA: Academic Press.
25. Glenn. J.W and N. Kleiner (1968), Speaker identification based on nasal phonation, The Journal of Acoustical Society of America., 43, 368-372.