Benchmark for Speaker Identification Using Long Term Average Spectrum in Kannada Speaking Individuals

Jyothi S. [1] & S. R. Savithri[2]

Abstract

The identification of people by their voices is a common practice in everyday life. In the last four decades, speaker recognition research has advanced a lot. The aim of the study was to generate benchmarking for speaker identification using Long Term Average Spectrum of speech in Kannada speaking individuals. Ten female Kannada speaking normal subjects in the age range of 18-25 years participated in the study. Material included two standard sentences in Kannada developed such that it embedded most of the phonemes in Kannada. Subjects were informed to speak the sentence in a normal modal voice. Samples were analyzed using Long Term Average Spectrum (LTAS) of speech of Computerized Speech Lab (CSL). From the LTAS, kurtosis and skewness were extracted and noted for each speaker. The results revealed several interesting points. Skewness and kurtosis appear to be robust when the number of subjects is limited. Its efficiency drops when the number of subjects increased. However, a 90% benchmarking was obtained for a group of 5 speakers. The results of the present study are restricted to female speakers and Kannada language. Hence generalization of the results to other languages and gender is questionable. Future studies with five speakers in other Indian languages, indirect or mobile recording and disguise conditions are warranted.

Key words: speaker recognition, kurtosis, skewness, Long Term Average Spectrum (LTAS)

Spoken language is the most natural way used by humans to communicate information. The speech signal conveys several types of information. From the speech production point of view, the speech signal conveys linguistic information (e.g., message and language) and speaker information (e.g., emotional, regional, and physiological characteristics). Most of us are aware of the fact that voices of different individuals do not sound alike. This important property of speech of being speaker-dependent is what enables us to recognize a friend over a telephone. The ability of recognizing a person solely from his voice is known as speaker recognition. Hecker (1971) suggests that speaker recognition is any decision-making process that uses the speaker-dependent features of the speech signal.

Hecker (1971) and Bricker and Pruzansky (1976) recognize three major methods of speaker recognition - (1) by listening (2) by visual inspection of spectrograms, and (3) by machine. More recently, with the availability of digital computers, automatic and objective methods can be devised to recognize a speaker uniquely from his voice.

Speaker identification by listening is entirely a subjective method. Hecker (1971) reported that speaker recognition by listening appears to be the most accurate and reliable method at that time. The second method of speaker recognition is based upon the visual examination and comparison of the spectrograms. Kersta (1960) coined the word "voice print" in a report discussing identification of speaker by visual inspection of spectrograms and concluded this method seemed to offer good possibility. Stevens (1968) compared aural with the visual examination of spectrogram using a set of eight talkers and found that error rate for listening is 6% and for visual is 21%. These scores depended upon the talker, phonetic content and duration of the speech material.

In speaker identification by machine, acoustic parameters from the signals are extracted and are analyzed by the machines. The objective methods can be further classified into (a) semi-automatic method, and (b) automatic method. In the semi-automatic method, there is extensive involvement of the examiner with the computer, whereas in the automatic method, this contact is limited. In the last four decades, speaker recognition research has advanced a lot. The applications of speaker recognition technology are quite varied and continually growing. Some commercial systems have been applied in certain domains. Speaker recognition technology makes it possible to use a person's voice to control the access to restricted services (automatic banking services), information (telephone access to financial transactions), or area (government or research facilities). It also allows detection of speakers, for example, voice-based information retrieval and detection of a speaker in a multiparty dialog.

[1]e-mail: jyo.ammu@gmail.com; [2]Professor of Speech Sciences, AIISH, Mysore, savithri2k@ gmail. com.

There have been several studies on the choice of acoustic features in the speech recognition tasks. In these methods first and second formant frequencies (Stevens, 1971; Atal, 1972; Nolan, 1983; Hollien, 1990; Kuwabara & Sagisaka, 1995 and Lakshmi & Savithri, 2009), higher formants (Wolf, 1972), Fundamental frequency (Atkinson, 1976), F0 contour (Atal, 1972), LP coefficients (Markel & Davis, 1979; Soong, Rosenberg, Rabiner & Juang, 1985) , Cepstral Coefficients & MFCC (Atal, 1974; Fakotakis, Anastasios & Kokkinakis, 1993; Rabiner & Juang, 1993; Reynold, 1995), LTAS (Kiukaanniemi, Siponen & Mattila, 1982), Cepstrum (Luck, 1969; Atal, 1974; Furui, 1981; Li & Wrench, 1983; Higgins & Wohlford, 1986; Che & Lin, 1995; Jakkar, 2009) & glottal source parameters (Plumpe, Quatieri & Reynolds, 1999), and long-term average spectra (Hollien & Majewski, 1977 among others) have been used in the past.

Long Term Average Spectrum (LTAS) is computed by calculating consecutive spectra across the chosen segment and then taking the average of each frequency interval of the spectra. However, it may be unstable for short segments (Pittam & Rintel, 1996). A range of factors have been correlated or found to be important in speaker recognition. These are all related to the original set of indices that was defined by Abercrombie (1967). The features presented include the speaker's gender, age, and regional or foreign accent. In addition, other factors not related to the voice production impact upon the listeners' ability to detect speaker identity. These include retention interval, sample duration and speaker familiarity. Further, acoustic features that are immediately available from the voice signal can be used to separate speakers. These include LTAS, fundamental frequency and formant transitions.

Hollien and Majewski (1977) concluded that n-dimensional Euclidian distance among long-term speech spectra (LTS) can be utilized as criteria for speaker identification at least under laboratory conditions. Its power as identification tool is somewhat language dependent. The LTS technique constitutes a reasonable robust tool in the laboratory but its efficiency is quickly reduced when distorting effect of the type found in more realistic environment impinge on the process. It has been argued to be effective in speaker discrimination processes (Hollien & Majewski, 1977; Doherty & Hollien, 1978; Kiukaanniemi, Siponen & Mattila, 1982; Hollien, 2002). It has however, also been argued to display voice quality differences (Hollien, 2002; Tanner, Roy, Ash & Buder, 2005), to successfully differentiate between genders (Mendoza, Valencia, Muñoz & Trujillo, 1996) and to display talker ethnicity (Pittam & Rintel, 1996).

The advantage of LTAS from a forensic perspective is that it has more or less direct physical interpretation, relating to the location of the vocal tract resonances. This makes LTAS more justified as evidence than Mel Frequency Cepstral Coefficients (MFCC). LTAS vectors of the questioned speech sample and the suspect's speech sample can be plotted on top of each other for visual verification of the degree of similarity. The advantages of LTAS from automatic speaker recognition perspective would be simple implementation and computational efficiency. In particular, there is no separate training phase included; the extracted LTAS vector will be used as the speaker model directly and matched with the test utterance LTAS using a distance measure. In view of this, and in view of the lack of benchmark of LTAS for Kannada speakers, the present study was undertaken. The aim of the study was to generate benchmarking for speaker identification using Long Term Average Spectrum of speech in Kannada speaking individuals. Specifically, skewness and kurtosis were extracted from LTAS for which the percent correct identifications were determined.

Method

Subjects: Ten female Kannada speaking normal subjects participated in the study. The subjects were in the age range of 18-25 years. They had passed at least 10^{th} standard and all speakers belonged to the same dialect. The inclusion criteria of subjects were (a) no history of speech, language and hearing problem (b) normal oral structures and (c) no other associated psychological and neurological problems.

Material: Two standard sentences in Kannada formed the material. These sentences were developed such that it embedded most of the phonemes in Kannada. The sentences were written on a separate card. The sentences are given below.

Namma u:ru karnataka ra:dzjadalliruva shivamogga dzilleja chikkada:da thirthahalli.
illi dzo:ga dzalapa:thavu bahu rabasava:gi entunu:ra ippathombattu adi etharadinda dhumukuttade.

Recording procedure: The testing was done in a laboratory condition. Speech samples were collected individually. The sentences were presented visually to the participants. Subjects were informed about the nature of the study and were instructed to speak the sentence in a normal modal voice. Four repetitions of the sentences were recorded. Thus forty samples were recorded from 10 speakers. The recordings were done using Computerized Speech Lab [CSL Model 4500 software (Kay Pentax, New Jersey)]. All these were recorded on a computer memory using a 12-bit A/D

(Analog to Digital) converter at a sampling frequency of 16,000 Hz.

Acoustic analysis: The pauses and noises were edited from the sample using Adobe Audition software (version 2.00, Syntrillium Software Corporation). All the four recordings of each subject were stored in separate folders. Long Term Average Spectrum (LTAS) of CSL was used to analyze the samples. A Hamming window with a Nyquist frequency sampling, and pre-emphasis of 0.8 was used to extract LTAS. Figure 1 shows the waveform and LTAS from a speech sample.

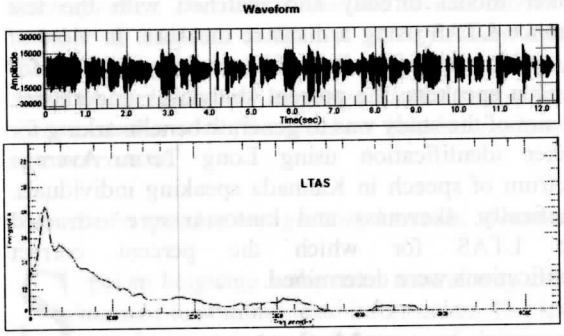

Figure 1. *Waveform (upper window) and LTAS (lower window) of a speech sample.*

From the LTAS, kurtosis and skewness were extracted and noted for each speaker. The data was normalized using the following formula.

$$N = \frac{X - Min}{Max - Min}$$

In this study all the voice samples were contemporary, as all the four recordings were carried out in one sitting. Closed-set speaker identification tasks were performed, in which the examiner was aware that the "unknown" speaker was among the "known" ones. The speakers recorded in first and second trails were considered as "known" and those done in thirds and fourth trials were considered as "unknown" speakers. All the "known" speakers were numbered from KS1 to 10 and corresponding "unknown" speakers were numbered as US1 to 10 For example, speaker KS1 (known) and speaker US1 (unknown) represent the same speaker in different trials of recording.

Two conditions were considered. In the first condition, one "unknown" speaker was compared with all the ten "known" speakers. An illustration is provided in the Table 1.

Table 1. *Unknown speaker (speaker 1) is compared with ten known speakers on skewness and kurtosis*

Speaker	Unknown speaker		Known speakers	
	Skewness	Kurtosis	Skewness	Kurtosis
KS1	0.180	0.299	0.318	0.437
KS2			0.397	0.336
KS3			0.453	0.486
KS4			0.247	0.320
KS5			0.445	0.414
KS6			0.114	0.162
KS7			1	1
KS8			0	0
KS9			0.858	0.805
KS10			0.542	0.567

The Euclidean distance was calculated in Microsoft Excel. Euclidean Distance is the most common use of distance. Euclidean distance or simply 'distance' examines the *root of square differences* between coordinates of a pair of objects. The formula to calculate Euclidean distance was as follows: Euclidean distance = $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$ where X and Y, in this study, refer to skewness and kurtosis. In Table 2, Euclidian distance is least for KS4. Therefore, US1 is likely to be KS4.

In the second condition, all the ten speakers were grouped into two sub-groups of five speakers. Only five speakers were considered in each group and one "unknown" speaker was compared with all the five "known" speakers. For example, in Table 3, the least Euclidian distance is for KS4. Therefore, it implies that US1 is likely to be KS4.

The graphs were plotted with skewness on the horizontal axis and kurtosis on vertical axis for group of different number of speakers. The unknown speaker was compared with the known speakers. Positive and negative speaker identifications were based on the Euclidian distance between the unknown and the known speakers. If the distance between unknown speaker and the respective known speaker was less, then speaker identification was deemed to be correct; if the distance between unknown speaker and any other known speaker was less, then speaker was deemed to be falsely identified or not correctly identified. The percentage correct identification was calculated by using the following formula:

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

The mean and SD of skewness and kurtosis were calculated.

Table 2. *Euclidian distances for US1 with KS1-KS10*

Unknown speaker	Skewness	Kurtosis	Known speakers	Skewness	Kurtosis	Euclidean distance
US1	0.180	0.3	KS1	0.318	0.437	0.195
			KS2	0.397	0.336	0.220
			KS3	0.453	0.486	0.331
			KS4	0.247	0.320	**0.070**
			KS5	0.445	0.414	0.289
			KS6	0.114	0.162	0.153
			KS7	1	1	1.078
			KS8	0	0	0.349
			KS9	0.858	0.805	0.845
			KS10	0.542	0.567	0.450

Table 3. *Euclidian distances for US1 with KS1- KS5*

Unknown speaker	Skewness	Kurtosis	Known speakers	Skewness	Kurtosis	Euclidean distance
US1	0.180	0.3	KS1	0.318	0.437	0.195
			KS2	0.397	0.336	0.220
			KS3	0.453	0.486	0.331
			KS4	0.247	0.320	**0.070**
			KS5	0.445	0.414	0.289

Table 4. *Mean and Standard Deviation (SD) of normalized skewness and kurtosis*

Subject No.	Skewness Trials 1,2	Kurtosis Trials 1,2	Skewness Trials 3,4	Kurtosis Trials 3,4
1.	0.180	0.299	0.318	0.437
2.	0.146	0.116	0.397	0.336
3.	0.237	0.303	0.453	0.486
4.	0.167	0.219	0.247	0.320
5.	0.192	0.195	0.445	0.414
6.	0.176	0.228	0.114	0.162
7.	1	1	1	1
8.	0	0	0	0
9.	0.551	0.543	0.858	0.805
10.	0.422	0.491	0.542	0.567
Mean	**0.307**	**0.339**	**0.438**	**0.453**
SD	**0.288**	**0.282**	**0.308**	**0.291**

Table 5. *Unknown speaker (US1) is compared with ten known speakers and is identified with KS4 (false identification)*

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
US1	0.180	0.30	0.318	0.437	KS1	0.195
			0.397	0.336	KS2	0.220
			0.453	0.486	KS3	0.331
			0.247	0.320	**KS4**	**0.070**
			0.445	0.414	KS5	0.289
			0.114	0.162	KS6	0.153
			1	1	KS7	1.078
			0	0	KS8	0.349
			0.858	0.805	KS9	0.845
			0.542	0.567	KS10	0.450

Table 6. *Unknown speaker (US6) is compared with ten known speakers and is identified with KS6 (correct identification)*

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.318	0.437	KS1	0.253
			0.397	0.336	KS2	0.247
			0.453	0.486	KS3	0.379
			0.247	0.320	KS4	0.116
			0.445	0.414	KS5	0.328
US6	0.176	0.228	0.114	0.162	**KS6**	**0.090**
			1	1	KS7	1.129
			0	0	KS8	0.288
			0.858	0.805	KS9	0.893
			0.542	0.567	KS10	0.499

Table 7. *Mean (M) and Standard Deviation (SD) of skewness and kurtosis in groups of 5 subjects*

S. No	Skewness	Kurtosis	Skewness	Kurtosis	S No.	Skewness	Kurtosis	Skewness	Kurtosis
	Trials 1,2	Trials 1,2	Trials 3,4	Trials 3,4		Trials 1,2	Trials 1,2	Trials 3,4	Trials 3,4
1)	0.180	0.230	0.318	0.437	1)	0.176	0.228	0.114	0.162
2)	0.146	0.116	0.397	0.336	2)	1	1	1	1
3)	0.237	0.303	0.453	0.486	3)	0	0	0	0
4)	0.167	0.219	0.247	0.320	4)	0.551	0.543	0.858	0.805
5)	0.192	0.195	0.445	0.414	5)	0.422	0.491	0.542	0.567
M	**0.185**	**0.227**	**0.372**	**0.399**		**0.430**	**0.452**	**0.503**	**0.507**
SD	**0.034**	**0.078**	**0.088**	**0.070**		**0.384**	**0.376**	**0.441**	**0.422**

Table 8. *Unknown speaker US5 is compared with ten known speakers and is identified with KS5 (Correct identification)*

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.114	0.162	KS1	0.450
			1	1	KS2	0.770
			0	0	KS3	0.647
			0.858	0.805	KS4	0.537
US5	0.422	0.491	0.542	0.567	**KS5**	**0.142**

Table 9. *Unknown speaker US4 is compared with ten known speakers and is identified with KS5 (false identification)*

Unknown speaker	Skewness	Kurtosis	Skewness	Kurtosis	Known speakers	Euclidean distance
			0.114	0.162	KS1	0.580
			1	1	KS2	0.640
			0	0	KS3	0.774
US4	0.551	0.543	0.858	0.805	KS4	0.403
			0.542	0.567	**KS5**	**0.025**

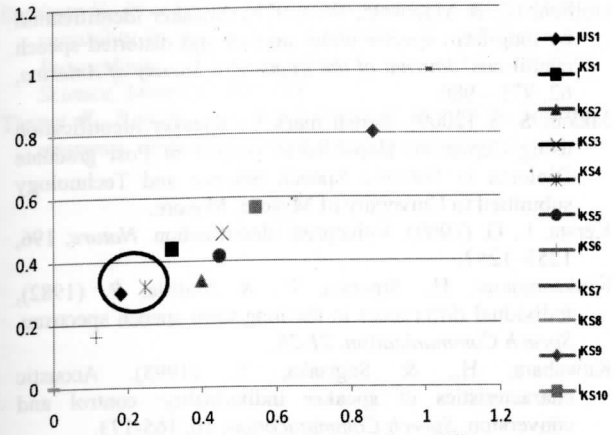

Figure 2. *False identification of US1 as KS4 in a group of ten speakers.*

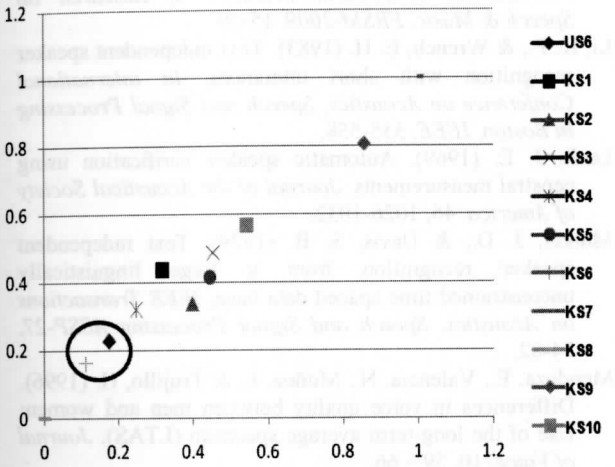

Figure 3. *Correct identification of US6 as KS6 in a group of ten speakers.*

Results

Condition I: The data showed high variations in skewness and kurtosis. Subject 8 had normalized skewness and kurtosis of '0' and subjects 7 had normalized skewness and kurtosis of '1". Table 4 shows the mean and SD of normalized skewness and kurtosis in ten subjects across trials. Tables 5 and 6 and figures 2 and 3 show the Euclidian distances and correct/ false identification, respectively. The overall percentage of the correct responses was found to be only 30%.

Condition II: The results indicated variability among subjects. Mean kurtosis was higher than mean skewness. Table 7 shows the mean (M) and Standard Deviation (SD) of skewness and kurtosis in groups of 5 subjects. Table 8 and 9 shows the Euclidian distances and figures 4 and 5 show an example of correct and false identification. The percentage of the correct identification was 90%. To summarize, in the first

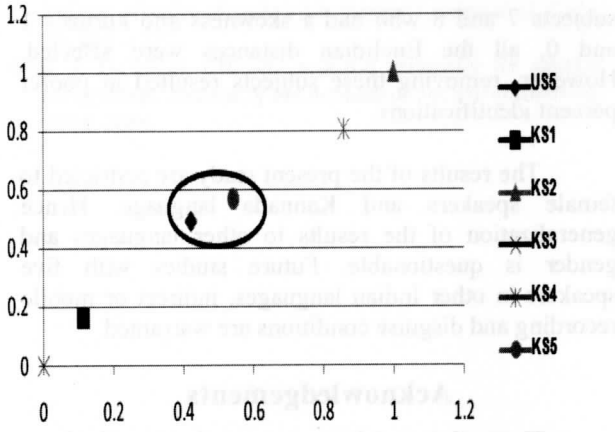

Figure 4. *Correct identification of US5 as KS5 in a group of five speakers.*

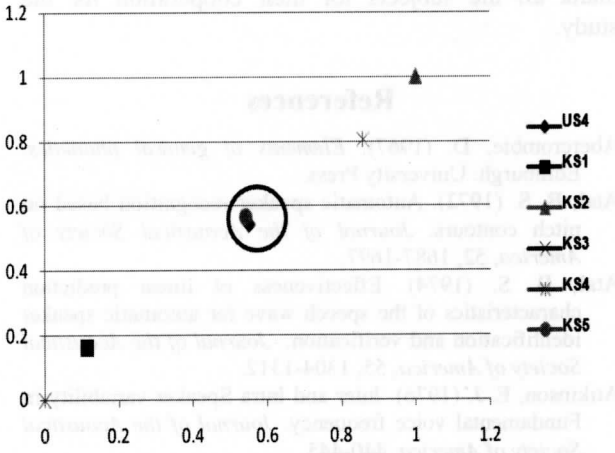

Figure 5. *False identification of US4 as KS5 in a group of five speakers.*

condition, the correct identification was 30% and in the second condition it was 90%.

Discussion

The results supports the earlier studies in that the percent correct identification reduced with increase in the number of subjects. Hollien and Majewski (1977) found 100% and 88% identification using LTAS in normal speech in full band and limited band conditions. However, Hollien and Majewski (1977) used the power spectra but the present study used skewness and kurtosis extracted from LTAS. Skewness and kurtosis appear to be robust when the number of subjects is limited to 5. Its efficiency drops when the number of subjects increased. However, ***a 90% benchmarking was obtained for a group of 5 speakers.***

It appeared that some speakers were very distinct (subjects 7, 8) and others were not. Because of

subjects 7 and 8 who had a skewness and kurtosis 1 and 0, all the Euclidian distances were affected. However, removing these subjects resulted in poorer percent identifications.

The results of the present study are restricted to female speakers and Kannada language. Hence generalization of the results to other languages and gender is questionable. Future studies with five speakers in other Indian languages, indirect or mobile recording and disguise conditions are warranted.

Acknowledgements

The authors wish to express their gratitude to Dr. Vijayalakshmi Basavaraj, Director, AIISH for granting permission to carry out this study. They also thank all the subjects for their cooperation for the study.

References

Abercrombie, D. (1967). *Elements of general phonetics.* Edinburgh: University Press.

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 of the Acoustical Society of America*, 55, 1304-1312.

Atikinson, E. J. (1976). Inter and Intra Speaker variability in Fundamental voice frequency. *Journal of the Acoustical Society of America*, 440-445.

Bricker, P. D., & Pruzansky, S. (1976). Speaker recognition. In N. J. Lass (Eds.), *Contemporary issues in experimental phonetics.* (pp.295-326). New York: Academic Press.

Che, C., & Lin, Q. (1995). Speaker recognition using HMM with experiments on the YOHO database. *In Eurospeech*, 625-628.

Doherty, E. T., & Hollien, H. (1978). Multiple-factor speaker identification of normal and distorted speech. *Journal of Phonetics*, 6, 1 – 8.

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

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and signal Processing,* 29, 254-272.

Hecker, M. H. L. (1971). Speaker recognition: basic considerations and methodology. *Journal of Acoustical Society of America,* 49, 138.

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.

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

Hollien, H. (2002). Forensic voice identification. San Diego, CA: Academic.

Hollien, H., & Majewski, W. (1977). Speaker identification by long-term spectra under normal and distorted speech conditions. *Journal of the Acoustical Society of America*, 62, 975 – 980.

Jakkar, 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, Mysore.

Kersta, L. G. (1960). Voiceprint identification. *Nature*, 196, 1253–1257.

Kiukaanniemi, H., Siponen, P., & Mattila, P. (1982), Individual differences in the long term speech spectrum. *Speech Communication, 21-28.*

Kuwabara, H., & Sagisaka, Y. (1995). Acoustic characteristics of speaker individuality: control and conversion. *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,* 15-19.

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.

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

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.

Mendoza, E., Valencia, N., Muñoz, J., & Trujillo, H. (1996). Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *Journal of Voice*, 10, 59 – 66.

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

Pittam, J., & Rintel, E. S. (1996). The acoustics of voice and ethnic identity. In P. McCormack & A. Russell, (Eds.), *Proceedings of the sixth Australian International Conference on Speech Science and Technology* (pp. 115 – 120). Adelaide, Australia: Australian Speech Science and Technology Association.

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

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

Reynold, D.A. (1995). Speaker identification and verification using Gaussian mixture speaker models, *Speech Communication, 17,* 91-108.

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.

Stevens, K. N. (1968). Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material. *Journal of the Acoustical Society of America,* 44, 1596–1607.

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.

Tanner, K., Roy, N., Ash, A., & Buder, E. H. (2005). Spectral moments of the long-term average spectrum: Sensitive indices of voice change after therapy? *Journal of Voice*, 19, 211 – 222.

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