Benchmark for Speaker Identification Using Glottal Source Parameters in Hindi Speakers

Aparna V. S.[1] & S. R. Savithri[2]

Abstract

*Studies concerned with establishing parameters for speaker verification are important because of the legal ramifications and because of the forensic involvements associated with the application of these studies. Success of identifying the speaker depends on extracting speaker-dependent features from speech signals that can effectively distinguish one speaker from another. It is not known as to what percent matching would indicate similarity/dissimilarity of speaker, or benchmarking of various features is not established. In this context the aim of the present study was to determine the benchmark for speaker identification using glottal source parameters in direct recording condition. Ten normal Hindi speaking male subjects in the age range of 21-38 years participated in the study. The material used was nine commonly occurring, meaningful Hindi words containing the long vowels /a:/, /i:/, and /u:/ in the word-medial position embedded in sentences. The vowels were displayed as waveform and were acoustically zoomed to extract the source and filter parameters using Acophone I and SSL software (Voice and Speech Systems, Bangalore). Glottal source parameters open quotient (OQ), leakage quotient (LQ) and speed quotient (SQ) were extracted in 10 steady state point of each of the vowels. The results of the present study showed that the glottal source doesn't remain the same even in normal mode of speaking. Hence these parameters don't serve as a good measure for speaker identification. In general it could be concluded that OQ*LQ, OQ*SQ, and LQ*SQ cannot be considered as an efficient parameter for speaker identification in field conditions in Hindi speakers.*

Key words: open quotient, leakage quotient, speed quotient, euclidian distance

Human beings have many characteristics that make it possible to distinguish one individual from another. Some individual characteristics can be perceived very rapidly such as facial features vocal quality and behaviors. Voice is the very emblem of the speaker, indelibly woven into the fabric of speech. In this sense each of our utterances of spoken language carries, not only its own message, but through ones accent, tones of voice and habitual voice quality it is also an audible declaration of our membership of a particular regional group, of our individual physical and psychological identity, and of our momentary mood. Thus the voice of an individual is said to be having its own characteristics and distinct distinguishable quality.

Speaker recognition is any decision making process that uses speaker dependent features of the speech signal (Hecker, 1971). Atal (1976) suggests that speaker recognition is any decision making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance. Nolan (1983) identified two classes of speaker recognition - speaker identification and speaker verification. Speaker recognition includes two sub-fields (a) naive speaker recognition and (b) technical speaker recognition. Technical speaker recognition is usually called as "Speaker Identification by expert" which uses specialized techniques (Nolan, 1983). Hecker (1971) and Bricker and Pruzansky (1976) identified three methods of speaker recognition (a) by listening (b) by visual inspection of spectrograms, and (c) by machine.

In speaker verification 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). An utterance from an unknown speaker has to be attributed, or not, to one of a population of known speakers for whom reference samples are available. Here only two types of decision are possible, either the unknown sample is correctly identified or it is not. The goal of speaker identification is to determine which one of a group of known speakers' best matches the test speech sample. Speaker identification can be constrained to a known phrase (text-dependent) or totally unconstrained (text-independent).

Of the three methods speaker identification by machine has received greater interest in recent past. Extracting speaker dependent parameter from signals and analyzing them by machines is an objective method which is classified into automatic and semiautomatic

[1]e-mail: aparna.sasi@gmail.com; [2]Professor of Speech Sciences, AIISH, Mysore, savitri2k@gmail.com

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. A combination of subjective and objective method is usually used. In the past pitch, intensity, phonemic voicing patterns (Hecker, 1971), long-term speech spectra (Hollien & Majewski, 1977), fundamental frequency (Abberton & Fourcin, 1978), cepstral parameterization (Plumpe, Quatieri & Reynolds, 1999), fundamental frequency, the third and fourth formants, and the closing phase of the glottal wave (Lavner, Gath & Rosenhouse, 2001), four formants (F1, F2, F3, F4), the amount of periodic and aperiodic energy in the speech signal, the spectral slope of the signal and the difference between the strength of the first and second harmonics (Carol Epsy-Wilson, Sandeep & Vishnubhotla, 2006), first three formants, word duration, closure duration, transition duration in disguised speech (Savithri, 2008) have been used. Sharma, Jain and Sharma (2009) in their study found that other supralaryngeal parameters like formant frequencies may shift during disguise but the open quotient and glottal leakage were found to occur in certain range for normal and disguised modes as the degree of glottal opening remains similar in normal mode but varies appreciably for disguised mode. Pamela (2002) studied the reliability of voice prints. Within the preview of her study, it was suggested that two samples can be considered to be from different speakers when more than 67% of measurements are different in natural speaking condition. But the validity of this method is still in question. Jakhar (2009) used quefrency for benchmarking and result obtained was a mean percentage of 88.33 (5 speakers), 81.67 (10 speakers) and 60 (20 speakers) in live v/s live condition, 81.67 (5 speakers), 68.33 (10 speakers) and 50 (20 speakers) in mobile v/s mobile, and 78.33 (5 speakers), 68.33 (10 speakers) and 43.33 (20 speakers) in live v/s mobile condition. The results indicate that speaker identification was higher when mode of recording was same and when the number of the speakers was less in the group. Lakshmi (2009) used formants F1- F2 for benchmarking and obtained benchmark of 70% for vowel /i:/, 65% for vowel/a:/ and benchmarking for other vowels were below chance level when 5 speakers were considered, and below chance level for ten and twenty speakers for all three vowels.

However, the question regarding the most appropriate speech parameter for semi automatic/automatic speaker identification in real forensic condition are still far from being answered. To prove that the suspect is a criminal, it needs to be verified beyond reasonable doubt that the voice of the criminal and voice of the suspect are the same. Success in this task depends on extracting speaker-dependent features from the speech signal that can effectively distinguish one speaker from another. Ideally, the features chosen for speaker recognition must satisfy the following criteria (Wolf, 1972): have lower within-speaker (within source) variability and relatively higher between speaker (between sources) variability, be stable over time, be difficult to disguise or mimic, be robust to transmission and noise, be relatively easy to extract and measure, and should occur frequently in the speech samples.

The glottal source is an important component of voice as it can be considered as the excitation signal to the voice apparatus. The use of the glottal source for pathology detection or the biometric characterization of the speaker is an important objective in the acoustic study of the voice now a days. The likely shape of the vocal tract can be approximately estimated from the analysis of the spectral shape of the voice signal. In automatic speaker recognition, coefficients representing the sounds, taking into consideration the vocal tract shape and excitation, are parameterized and used as features. It is not known as to what percent matching would indicate similarity/dissimilarity of speaker or benchmarking of various features is not established. In this context, the present study evolved a benchmark for speaker identification using glottal source parameters, specifically open quotient, leakage quotient and speed quotient extracted from glottal source in direct recording condition.

Method

Participants: Ten normal Hindi speaking male subjects in the age range of 21-38 years with no history of neurological or psychological illness participated in the study.

Material and Procedure: The material used was nine commonly occurring, meaningful Hindi words containing the long vowels /a:/, /i:/ and /u:/ in the word-medial position embedded in sentences. Direct (live) recording of the four repetitions of these sentences by the participants as done by Jakhar (2009) was taken for the present study. The words and in turn the target vowels were truncated from the samples and stored in folders D1, D2, D3 and D4. The vowels were displayed as waveform and were acoustically zoomed to extract the source and filter parameters using Acophone I and SSL software (Ananthapadmanabha, 2008; Voice and Speech Systems, Bangalore). Glottal source parameters open quotient (OQ), leakage quotient (LQ) and speed quotient (SQ) were extracted in 10 steady state point of each of the vowels.

For each vowel thirty values (3 * 10 observations) were obtained in a single recording. Each recording was considered as that of unknown speaker and subsequent recording as that of the known speaker and percentage correct identification was noted for five speakers for vector OQ*LQ. Percentage correct identification for 10 speakers were calculated taking the average of D1, D2 and D3, D4. The study was extended to find out percent correct identification for vectors OQ*SQ and LQ*SQ in five speakers.

The OQ*LQ was plotted with OQ on horizontal axis and LQ on vertical axis for a group of known speaker versus one unknown speaker. Euclidian distance (ED) was calculated as the distance between unknown speaker (reference sample) and the known speakers (test sample) using the following formula:

In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2)

$$D(p, q) = \sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2)}.$$

Where p=reference subject, and q= test subject

X and Y belong to one of the parameters (OQ, LQ and SQ). If the distance between reference sample and test sample is least it is considered as correct identification. Percentage of correct identification was calculated using the following formula:

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

Results

Inter-speaker identification

OQ-LQ: The OQ-LQ was plotted with OQ on horizontal axis and LQ on vertical axis for a group of known speakers versus one unknown speaker. A total of 210 figures (3 vowels X 2 group of speakers X 6 combinations of 4 recordings X 10 speakers) were plotted for speaker identification. The Euclidian distances between selected unknown speakers and the corresponding five known speaker were calculated. The lowest Euclidian distance value is highlighted in the table. If the lowest value and the Euclidian distance value corresponding to the actual speaker are the same it was considered as correctly identified. That is if the distance between the unknown and corresponding known speaker was the lowest, then speaker was deemed to be correctly identified. If unknown speaker is closer to some other known speaker in terms of Euclidian distance it as deemed to as wrong identification. That is if the distance between unknown speaker and corresponding known speaker is more, then the speaker was deemed to be not identified.

The results indicated percent correct identification of 48.3, 34.6, and 33.3 for vowels /a:/, /i:/, and /u:/, respectively when five subjects were considered. The percent correct identification reduced drastically when 10 subjects were considered. Table 1 show the percent correct identification for three vowels when five and ten subjects are considered. Figure 1 shows the benchmarking for three vowels.

Table 1. *Percent correct identification of vowels /a:/, /i:/ and /u:/ for two groups of speaker for OQ-LQ*

Groups	% correct identification		
	/a:/	/i:/	/u:/
5 speakers (A)	48.3%	34.6%	33.3%
10 speakers (B)	10%	20%	30%

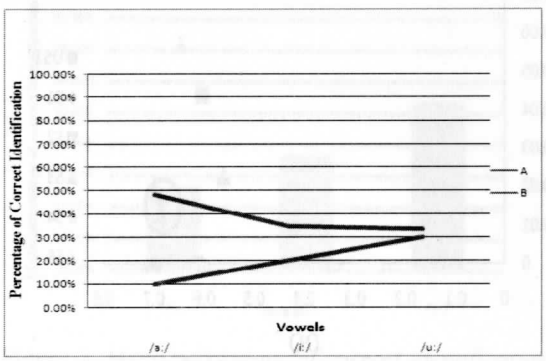

Figure 1. *Percent correct identification of vowels /a:/, /i:/ and /u:/ for two groups of speaker for OQ-LQ.*

Tables 2-3 shows the Euclidian distance of five speakers when the recordings D1 V/S D2 were considered for vowel /a/, and figures 2-3 represent the correct / false identifications.

Table 2. *Correct identification in a group of 5 speakers for OQ-LQ on vowel /a:/ when reference speaker was US1 (Lowest Euclidian distance is in bold)*

Reference speaker	Reference sample		Test sample			ED
	OQ	LQ		OQ	LQ	
US1	0.687	0.011	S1	0.682	0.016	**0.007**
			S2	0.604	0.044	0.089
			S3	0.559	0.023	0.128
			S4	0.605	0	0.082
			S5	0.645	0.057	0.062

For 10 speakers, average of two recordings D1 and D2 (D1) was considered as the reference and average of recordings D3 and D4 (D2) was considered as the test sample. As reported earlier the percent correct identification was less than chance level. Table

4-5 shows the correct/ false identification in 10 subjects and figures 4-5 shows the identification of subjects.

Table 3. *False identification in a group of 5 speakers for OQ-LQ on vowel /a:/ when reference speaker was US2 (lowest Euclidian distance is in bold)*

Reference Speaker	Reference sample		Test sample			ED
	OQ	LQ		OQ	LQ	
			S1	0.682	0.016	0.235
US2	0.447	0.018	S2	0.604	0.044	0.160
			S3	0.559	0.023	**0.113**
			S4	0.605	0	0.160
			S5	0.645	0.057	0.203

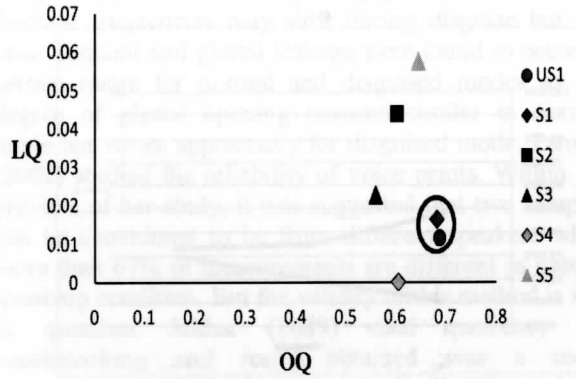

Figure 2. *Correct identification among 5 speakers for OQ-LQ on vowel /a:/.*

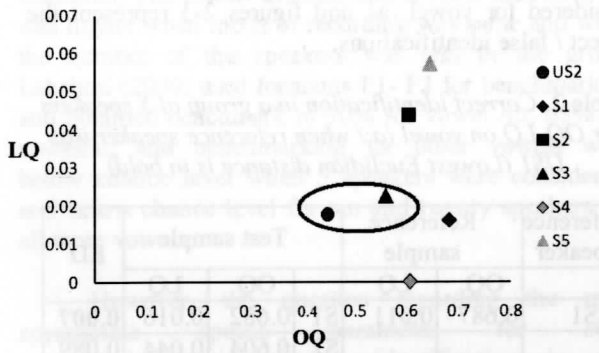

Figure 3. *False identification among 5 speakers for OQ-LQ on vowel /a:/ (US2 as S3).*

Table 4. *Correct identification in a group of 10 speakers for OQ- LQ on vowel /a:/ when reference speaker was US1 (US1 Least Euclidian distance is in bold)*

Reference speaker	Reference sample		Test sample			ED
	OQ	LQ		OQ	LQ	
US1	0.682	0.016	S1	0.685	0.012	**0.019**
			S2	0.625	0.039	0.080
			S3	0.503	0.020	0.199
			S4	0.653	0.009	0.050
			S5	0.618	0.060	0.094
			S6	0.689	0.071	0.055
			S7	0.609	0.093	0.120
			S8	0.631	0.040	0.075
			S9	0.652	0.05	0.060
			S10	0.714	0.144	0.127

Table 5. *False identification in a group of 10 speakers for OQ- LQ on vowel /a:/ when the reference speaker was US3 (least Euclidian distance is in bold)*

Reference Speaker	Reference Sample			Test Sample		ED
	OQ	LQ		OQ	LQ	
			S1	0.685	0.014	0.068
			S2	0.625	0.039	**0.010**
US3	0.619	0.031	S3	0.503	0.020	0.116
			S4	0.653	0.009	0.041
			S5	0.618	0.060	0.030
			S6	0.689	0.079	0.081
			S7	0.609	0.093	0.062
			S8	0.631	0.040	0.015
			S9	0.652	0.05	0.038
			S10	0.714	0.144	0.147

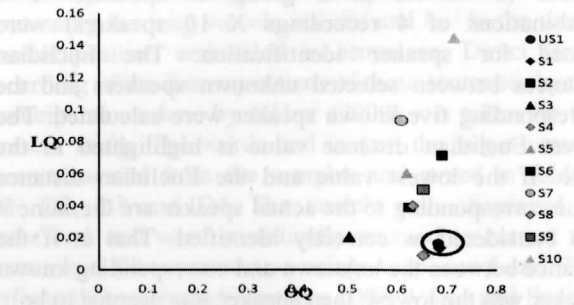

Figure 4. *Correct identification among 10 speakers for OQ- LQ on vowel /a:/ (US1 as S1).*

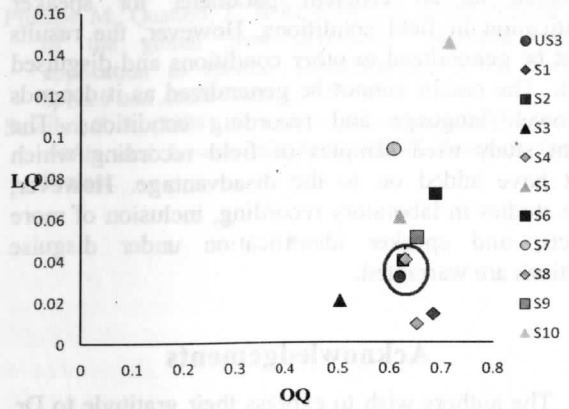

Figure 5. *False identification among 10 speakers for OQ- LQ on vowel /a:/ (US3 as S2).*

OQ-SQ and LQ-SQ: The OQ-SQ vector was plotted with OQ on horizontal axis and SQ on vertical axis and LQ –SQ vector was plotted with LQ on horizontal axis and SQ on vertical axis for a group of 5 known speakers versus one unknown speaker. A total of 180 figures (3 vowels X 5 speakers X 6 combinations of 4 recordings) were plotted together for OQ-SQ and LQ-SQ. The Euclidian distance between selected unknown speakers and the corresponding 5 known speaker were calculated. Results showed below chance level identification. Vowels /i:/ and /u:/ had better percent identification on OQ-SQ and LQ-SQ, respectively. Table 6 shows percent correct identification of three vowels for 5 speakers for both vectors. Figures 6 and 7 show the percent correct identification for OQ-LQ and LQ-SQ, respectively.

Table 6. *Percent correct identification of vowels /a:/, /i:/ and /u:/ for 5 speakers for vectors OQ-SQ and LQ-SQ*

Vectors	Group	% correct identification		
		/a:/	/i:/	/u:/
OQ-SQ	5 speakers	33.3%	46.6%	40%
LQ-SQ	5 Speakers	26.6%	36.6%	53.3%

Discussion

The present study investigated speaker identification using glottal source parameters in Hindi language in field conditions. The aim of the study was to determine benchmarking for source parameters. Specifically, open quotient, leakage quotient, and speed quotients were used to derive the benchmarking. The result throws light into several points of interest.

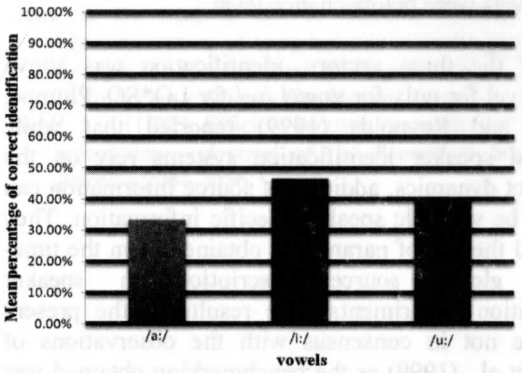

Figure 6. *Mean percentage of correct identification of vowels /a:/, /i:/ and /u:/ for 5 speakers using OQ- LQ vector.*

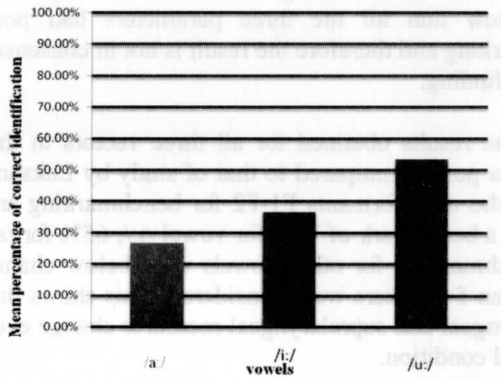

Figure 7. *Mean percentage of correct identification of vowels /a:/, /i:/ and /u:/ for 5 speakers using LQ- SQ vector.*

First of all in field condition (*OQ*LQ) the percent correct identification was better when 5 speakers were considered and the identification deteriorated when 10 speakers were considered.* The OQ*LQ was 48.3%, 34.6%, and 36.6% for vowel /a:/ /i:/ and /u:/, respectively when 5 speakers were considered. Benchmarking using vectors OQ * LQ were at below chance level for all vowels when ten speakers were considered.

Secondly, *vowel /u:/ had better percent identification* (LQ*SQ) compared to other vowels in most of the conditions. Thirdly, all three vowels had a very poor benchmark of below chance level for five speakers when vector OQ*LQ and OQ*SQ were used which indicates that these vectors are not useful for forensic speaker identification. Fourthly, vowel /u:/ obtained a benchmark of 53.3% for the vector LQ*SQ

when 5 speakers were considered and benchmarking for other vowels were below chance level.

Of the three vectors, identification was above chance level for only for vowel /u:/ for LQ*SQ. Plumpe, Quatieri and Reynolds (1999) reported that while traditional speaker identification systems rely on the vocal tract dynamics, addition of source information can prove to be valuable speaker-specific information. They suggested the use of parameters obtained from the time-domain glottal source description in speaker identification experiments. The results of the present study are not in consensus with the observations of Plumpe et al., (1999) as the benchmarking obtained was poor. Sharma et al., (2009) found that laryngeal measures (open quotient, leakage quotient) are less subjected to change compared to the supralaryngeal measures in disguise condition. So these parameters are supposed to provide better benchmarking. The results in the present study show that all the three parameters had poor benchmarking and therefore the result is not in consensus with the finding.

The results obtained for all three vectors in this study was poorer compared to that of study by Lakshmi (2009) who used formants F1-F2 for benchmarking and obtained a benchmark of 70% for vowel /i:/, 65% for /a:/ and benchmarking for other vowels were below chance level when 5 speakers were considered. This shows that both laryngeal and supralaryngeal measures change even in normal condition.

Jakhar (2009) used quefrency for benchmarking in Hindi speakers. Benchmarking obtained was 88.33 (5 speakers), 81.67 (10 speakers) and 60 (20 speakers) in live v/s live condition, 81.67 (5 speakers), 68.33 (10 speakers) and 50 (20 speakers) in mobile v/s mobile, and 78.33 (5 speakers), 68.33 (10 speakers) and 43.33 (20 speakers) in live v/s mob condition. With respect to the number of speakers, the percent correct identification was higher when the number of the speakers was less in the group. The present study showed very poor benchmarking on all three vectors than Jakhar for 5 speakers and 10 speakers. However, it is in consensus with Jakhar on the finding that percent correct identification decreased as the number of speakers increased. The results of the present study show that the glottal source doesn't remain the same even in normal mode of speaking. Hence these parameters don't serve as a good measure for speaker identification.

Conclusions

The result of the present study has contributed to the field of speaker identification. In general, it could be concluded that OQ*LQ, OQ*SQ and LQ*SQ cannot be considered as an efficient parameter for speaker identification in field conditions. However, the results cannot be generalized to other conditions and disguised speech. The results cannot be generalized as it depends on vowel, language and recording conditions. The present study used samples of field recording which might have added on to the disadvantage. However, future studies in laboratory recording, inclusion of more subjects, and speaker identification under disguise conditions are warranted.

Acknowledgements

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

References

Abberton, E., & Fourcin, A. J. (1978). Intonation and speaker identification. *Language and Speech,* 21, 305-318.

Ananthapadmanabha, T.V. (2008). Voice and Speech Systems, Bangalore.

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

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

Carol, Epsy-Wilson, Sandeep, M., & Vishnubhotla, S.A. (2006). *New Set of Features for Text-Independent Speaker Identification.* Institute for Systems Research and Dept. of Electrical & Computer Engineering, University of Maryland, College Park, MD, USA.

Hecker, M. H. L. (1971). Speaker Recognition: Basic considerations and methodology. *Journal of the Acoustical Society of America, 49,138-138.*

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.

Jakhar, 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.

Lakshmi, P. (2009). *Benchmark for speaker Identification using Vector F1 ~ F2.* Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to University of Mysore, Mysore.

Lavner, Y., Gath, I., & Rosenhouse, J. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, 4, 63-74.

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

Pamela, S. (2002). *Reliability of voice prints*, Unpublished dissertation submitted to the University of Mysore, Mysore.

Plumpe, M., Quatieri, T., & Reynolds, D. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Speech and Audio Proc*, 1, 569-586.

Savithri, S. R. (2008). *Acoustic similarities and differences within and between speakers*. AIISH Research fund Project.

Sharma, S., Jain, S. K., & Sharma, R. M. (2009). Characterization of temporal and acoustic parameters for speaker identification in disguise speech. *XX All India Forensic Science Conference, Jaipur.*

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