A Comparative Analysis of Speaker Identification on English and Hindi Database

Anjali Jain
M.Tech Scholar
Department of Electronics and Communication Engineering
Poornima College of Engineering, Jaipur

O.P. Sharma
Professor
Department of Electronics and Communication Engineering
Poornima College of Engineering, Jaipur

ABSTRACT
In this paper a text-dependent speaker recognition method is presented by combining Mel frequency cepstrum coefficients (MFCC) and Euclidean distance. The robustness of this speaker identification method for different speaking language is analyzed in this paper. The speaker identification algorithm using English and Hindi Indian voice database (IVD) which contains sentences of data spoken is accomplished. An improvement in recognition rate is observed by using different windows and increasing the number of training voice samples. Accuracy upto 100% can be obtained for text-dependent speaker identification for different windows by using a short training and testing utterance about 4 seconds.

Keywords
Speaker Identification, MFCC, Euclidean distance classifier, Feature extraction and database.

1. INTRODUCTION
Speech recognition is a topic that is very useful in many applications and environments in our day to day routine. It is a branch of biometric authentication where it is one of the fast gaining popularity as means of security measures due to its unique physical characteristics and identification of individuals [1-4]. Although retinal scans and fingerprints are more unfailing means of identification, but the speech has an advantage over them as it has been seen that voice/speech is a non-evasive biometric that can be collected with or without the person's knowledge or even transmitted over long distances via telephone and the voice of an individual cannot be stolen or misplaced like another forms of recognition, such as passwords or keys [5].

Speaker recognition has been a research topic for many years and various types of speaker models have been studied. Speaker-Recognition system [6] is a system specialized for speaker identification and authentication in which different users are distinguished by their unique voices. The general field of speaker recognition includes two fundamental tasks: speaker identification and speaker verification [4, 7–10]. Speaker identification involves classifying a voice sample as belonging to (that is, having been spoken by) one of a set N of reference speakers (N possible outcomes), whereas speaker verification involves deciding whether or not a voice sample belongs to a specific reference speaker (two possible outcomes—the sample is either accepted as belonging to the reference speaker or rejected as belonging to an impostor). The goal of work in speaker recognition is to find measurable quantities that minimize within-speaker variability and simultaneously maximize between-speaker variability [11].

The rest of the paper is organised as follows: Section II explains the system description. Section III introduces the self created English and Hindi IVD. The experimental analysis and results are discussed in Section IV and finally, Section V concludes the paper.

2. SYSTEM DESCRIPTION
The structure of the speaker identification system which works in two phases training and testing is shown in figure 1. Mel Frequency Cepsrtal Coefficient (MFCC) is used for feature extraction followed by Euclidean distance classifier for matching. In the training phase system, input training speech is analyzed and transformed into a feature vector sequence by MFCC block following the basic steps of pre-processing, framing, windowing, creation of mel-filter bank, discrete cosine transform explained in [12].

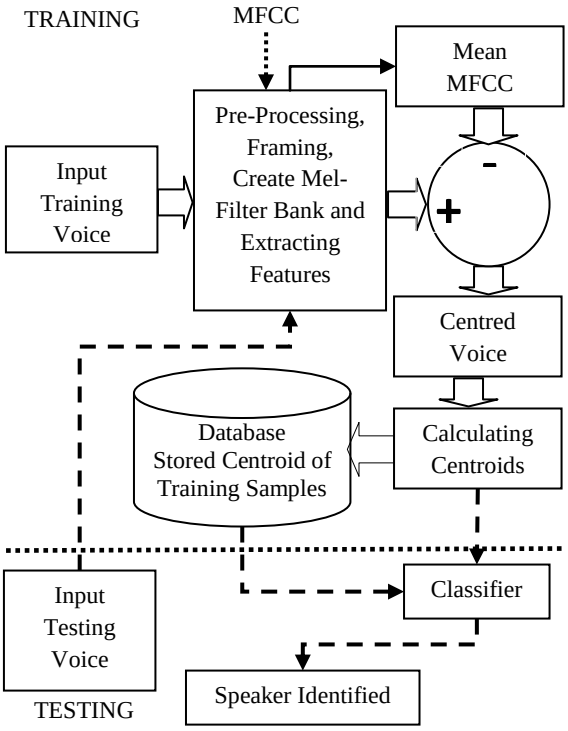

Fig 1: System architecture

Next the mean of MFCC is calculated and after that difference between this mean and MFCC is computed. The difference value obtained is referred to as the centred voice. After this centroid of each row and each column of the centred voice sample is calculated and matrix of the centroid obtained is stored in database.

The testing voice samples is fed to the MFCC block following the same procedure of training phase shown by dotted arrows

in figure 1, the centroids of testing voice sample are derived which are compared against with those stored in database using Classifier (Euclidean distance). The difference obtained using Euclidean distance is compared with the range of optimal threshold and the difference which satisfies this optimal range is the considered as the minimum distance. And finally the minimum distance gives the best match or can be said speaker is identified.

3. DATABASE OF VOICE SAMPLES

3.1 English and Hindi IVD

English and Hindi IVD corpus of read speech has been designed to provide speech data for the development and evaluation of speaker identification system. English and Hindi IVD corpus design was a joint effort of my own and my colleagues and bachelor students from department of Electronics and Electrical Engineering, Rajasthan Technical University. For English IVD, the sentence spoken is "I love my country" and for Hindi IVD the same sentence is translated to Hindi i.e. "Mujhe mere desh se pyar hai". It is a single-session database including 22 Indians speakers (15M/ 7F), each of which is repeated 9 times and the age covered from 20 to 28 and was recorded in a noise free environment with a fixed microphone. The recording work has been carried out in a closed room with all fans off on I floor. The equipment for recording is portable simple microphone. For database, voice messages were recorded into the most commonly used file type-.wav. The sampling frequency is chosen 8 kHz with a bit rate of 128kbps.

4. EXPERIMENTAL ANALYSIS AND RESULTS

For experimental analysis centroids are calculated by using MFCC feature extraction technique and algorithm is performed by varying the windows and the number of training samples. The algorithm is developed in MATLAB 10.0. In the experimental set-up, initially both for English and Hindi IVD, 5 voice samples per person is used for training i.e. 22*5=110 and remaining 4 voice samples per person for testing i.e.22*4=88. After that number of training voice samples per person was increased in step of 1, i.e. 6 samples per person are taken for training than varied to 7, 8 and finally 9 and the testing samples are fixed to 4 per person. The case containing 9 training sample per person is the one in which all test voice samples are already stored in training. The experimental result shows that as the number of training database increases, accuracy of the system increases for both the database.

The accuracy of speaker identification was measured by Euclidean distance between the test samples and all train voice samples. The result of the analysis on English and Hindi voice database has been shown in tabulated form in Table 1 and 2 also corresponding graphs are plotted in next figures no. 2 to 6 respectively.

Table 1. Comparison of speaker Identification on various windows on English IVD

Accuracy(%) on English IVD					
Voice samples	**Hamming**	**Hanning**	**Blackmann**	**Kaiser**	**Rectangular**
5	85.227	76.136	68.181	63.636	61.363
6	89.772	84.090	80.681	77.272	76.136
7	95.454	92.045	88.636	80.681	79.545
8	98.863	98.863	96.590	88.636	87.500
9	100	100	100	96.590	94.318

Table 2. Comparison of speaker Identification on various windows on Hindi IVD

Accuracy(%) on Hindi IVD					
Voice samples	**Hamming**	**Hanning**	**Blackmann**	**Kaiser**	**Rectangular**
5	48.863	46.591	43.182	40.909	38.636
6	65.909	63.636	56.818	55.682	53.409
7	79.545	76.136	72.727	64.773	62.500
8	93.182	89.773	88.636	71.591	69.318
9	100	100	100	79.545	77.273

It is also found that as the number of training voice samples increases from 5 to 9 per person the gap between the recognition rates for English and Hindi reduces and finally it can be seen from figure 6 that same recognition rate (100%) can be achieved for certain windows for both the database.

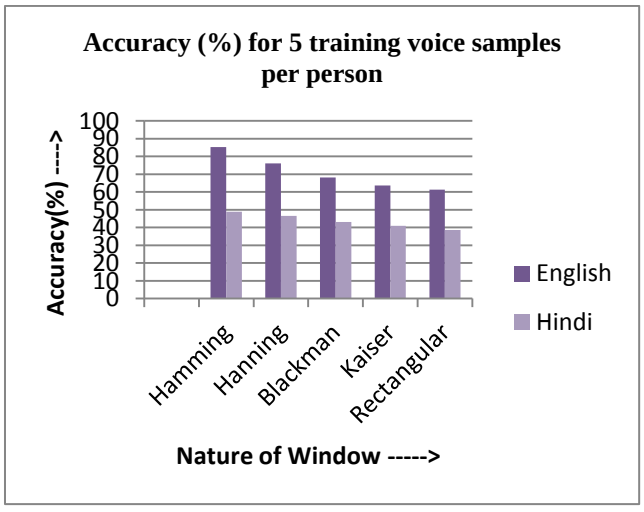

Fig 2: Accuracy v/s windows on training 5 voice samples per person

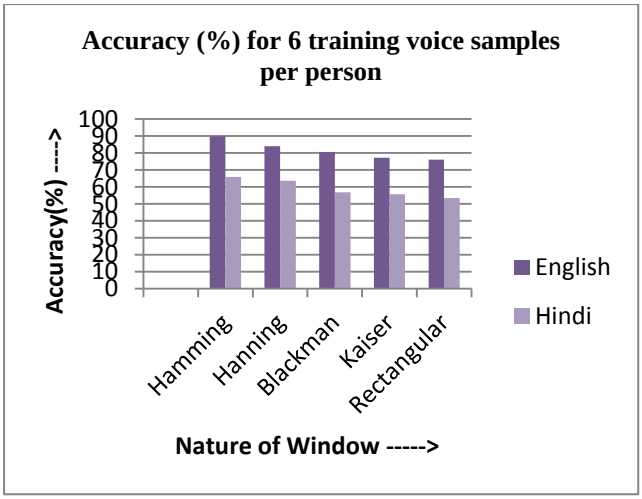

Fig 3: Accuracy v/s windows on training 6 voice samples per person

Fig 4: Accuracy v/s windows on training 7 voice samples per person

Fig 5: Accuracy v/s windows on training 8 voice samples per person

Fig 6: Accuracy v/s windows on training 9 voice samples per person

5. CONCLUSION

The goal of this paper was to implement a text-dependent speaker identification system. The proposed approach shows that the identification rate of the system increases as the number of training voice sample increases. In order to train and test the developed speaker identification system, IVD is created for English and Hindi language, which contains recordings of 22 Persons (15 males and 7 females) each of which is repeated 9 times. This paper also investigates the effect of various windows on the overall performance of the system and shows that out of all the windows hamming window gives the best result.

It is further verified that English language based database gives better result in terms of efficiency as compared to Hindi language database. The results also show that 100% recognition rate can be achieved by hamming, hanning and blackmann window in case if training voice samples already contain all the test voice samples for both the database.

6. REFERENCES

[1] P. Kartik, S.R.H. Prasanna and R.V.S.S.V. Prasad, "Multimodal Biometric Person Authentication System using Speech and Signature Feature", Technical Conference of IEEE Region 10 TENCON, Hyderabad, India, pp. 1-6, 2008.

[2] S Furui, "50 years of progress in speech and speaker recognition research", ECTI Transactions on Computer and Information Technology, Vol. 1, No.2, November 2005.

[3] D. A. Reynolds, "An overview of automatic speaker recognition technology", Proceeding IEEE International Conference on Acoustics, Speech, Signal Process, pp. IV-4072-IV-4075, 2002.

[4] Joseph P. Campbell, "Speaker Recognition: A Tutorial", Proceedings of the IEEE, Vol. 85, No.9, pp. 1437-1462, September 1997.

[5] H. Gish and M. Schmidt, "Text Independent Speaker Identification", IEEE Signal Processing Magazine, Vol. 11, No. 4, pp. 18-32, 1994.

[6] K. R. Farrell, R. J. Mammone and K. T, Assaleh, "Speaker Recognition Using Neural Network and Conventional Classifiers", IEEE Transaction speech and Audio processing, Vol.2, No.1, PART II, Jan., 1994.

[7] A. E. Rosenberg, "Automatic speaker verification: A review," Proceedings IEEE, Vol. 64, No. 4, pp. 475–487, Apr. 1976.

[8] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Commun., Vol. 17, No. 1–2, pp. 91–108, 1995.

[9] S. Nakagawa, W. Zhang, and M. Takahashi, "Text-independent speaker recognition by combining speaker specific GMM with speaker adapted syllable-based HMM" in Proceeding ICASSP, Vol. 1, pp. 81–84, 2004.

[10] F. Bimbot et al., "A tutorial on text-independent speaker verification", EURASIP Journal of Appl. Signal Processing, pp. 430–451, 2004.

[11] T. B. Alderman, "Forensic Speaker Identification, A Likelihood Ratio- Based Approach Using Vowel Formants", ser. Lincom Studies in Phonetics. Munich, Germany: LINCOM, 2005.

[12] Anjali jain, O.P. Sharma, "A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coefficient (MFCC): A Review", International Journal of Electronics & Communication Technology (IJECT) Vol. 4, Issue Spl - 4, pp. 26-29, April - June 2013.

APPENDIX

For hamming window

$$W_{ham}(n) = 0.54 - 0.46\cos\frac{2\pi n}{N_{samp}-1}$$
$$0 \le n \le N_{samp}-1 \qquad (3)$$

For hanning window

$$W_{han}(n) = 0.5 - 0.5\cos\frac{2\pi n}{N_{samp}-1}$$
$$0 \le n \le N_{samp}-1 \qquad (4)$$

For blackmann window

$$W_{black}(n) = 0.42 - 0.5\cos\frac{2\pi n}{N_{samp}-1} + 0.8\cos\frac{4\pi n}{N_{samp}-1},$$
$$0 \le n \le N_{samp}-1 \qquad (5)$$

For Kaiser window

$$W_{kai}(n) = \frac{I_0\left\{\beta\left[1-\left(\frac{n-\alpha}{\alpha}\right)^2\right]^{\frac{1}{2}}\right\}}{I_0(\beta)}$$
$$0 \le n \le N_{samp}-1 \qquad (6)$$

where I_o is the 0^{th} order modified Bessel function of the first kind, α is phase delay and β is shape parameter.

For rectangular window

$$W_{rect}(n) = 1$$
$$0 \le n \le N_{samp}-1 \qquad (7)$$

where N_{samp} is the number of samples in each frame.

Recognition rate or accuracy

Accuracy of the Voice System = [Number of Correctly Identified Voice Samples/Total Number of Tested Voice Samples] * 100