International Journal of Computing and Business Research (IJCBR)

ISSN (Online) : 2229-6166

Volume 3 Issue 2 May 2012

Performance study of Text-independent Speaker identification system using MFCC & IMFCC for Telephone and Microphone Speeches

Ruchi Chaudhary, National Technical Research Organization

Abstract: A 'state-of-the-art' Speaker Identification (SI) system requires a robust feature extraction unit, followed by a speaker classifier scheme. Over the years, Mel-Frequency Cepstral Coefficients (MFCC), modelled on the human auditory system, has been used as a standard acoustic feature set for speech related applications. Furthermore, it has been also shown that the Inverted Mel-Frequency Cepstral Coefficients (IMFCC) is also a useful feature set for SI, which contains information complementary to MFCC as, it covers high frequency region more closely. In this study, performance of speaker identification system is evaluated by generating Detection-error-trade-off (DET) curves, for both MFCC & IMFCC (in individual and fused mode, using two different kinds of databases (i.e. Microphone Speech, Telephone Speech). It is found, that IMFCC feature based classifier, produces improved accuracy, especially for telephone speech database and also, preferred mixing proportion of two streams (MFCC & IMFCC in combined model) are also obtained for both kind of database.

Key Words: Speaker Identification, MFCC, IMFCC, Fussed feature set

1. INTRODUCTION

Automatic Speaker Recognition is to verify a person's claimed identity from his voice. In text-independent speaker identification system, there is no constraint on the words which speakers are allowed to use. The reference (what is spoken in training) and the test utterances (what is uttered in actual use) may have completely different context. Feature extraction is method of obtaining the unique characteristic pattern of a speaker, known as features sets. A feature provides a more suitable, robust and compact representation of speaker's speech than the raw input signal. MFCC has been widely accepted as features input for a typical speaker recognition system because of its less vulnerability to noise perturbation, little session variability and, easiness to extract than other methods namely Line Spectral frequency (LSF), log Area Ratio (LAR), Perceptual log Area Ratio (PLAR), Perceptual Linear Prediction (PLP) etc. [1-2].

The computation of MFCC involves, averaging the low frequency region (upto 1 kHz) of the energy spectrum, by employment of closely spaced overlapping triangular filters. Smaller numbers of less closely spaced triangular filters are used to average the high

frequency zone. The figure 1 shows the block diagram for Mel frequency Cepstral coefficients.

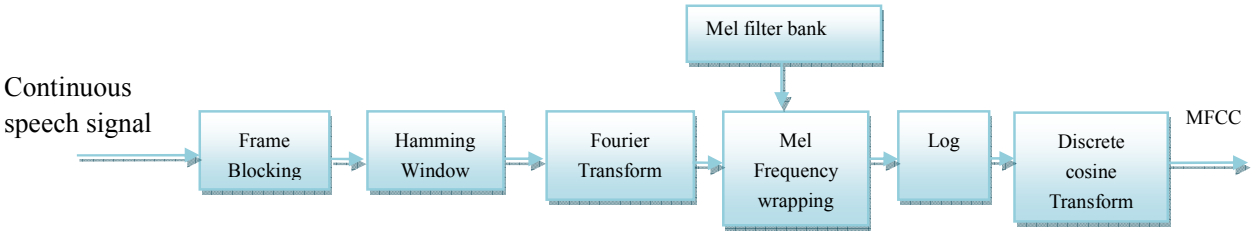

Figure 1: Block diagram for Mel frequency cepstral coefficients.

For MFCC feature extraction, Mel-scale frequency is related to linear frequency by empirical equation in (1), and the figure 2 shows the mel scale frequency relation to linear scale frequency.

$$f_{mel} = 2595 \log_{10}(1 + f/100) \quad (1)$$

the inverse of mel frequency wrapping function is given as (2)

$$f^{-1}_{mel}(f_{mel}) = 700\,(10^{f_{mel}/2595} - 1) \quad (2)$$

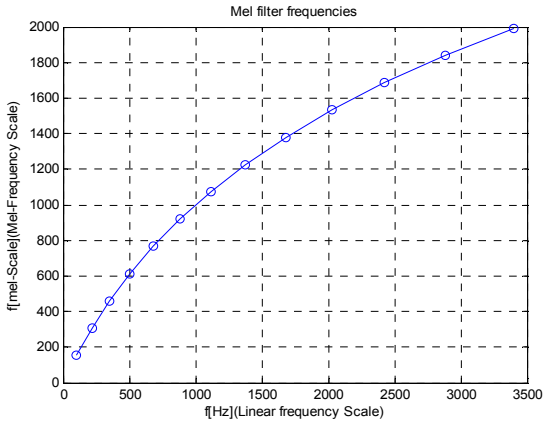

Figure 2: Mel scale Frequency related to linear scale frequency.

MFCC, thus, represents the low frequency region more accurately than the high frequency region and hence, can capture formants efficiently, which lie in the low frequency range and which characterize the vocal tract resonances. However, other formants that lie above 1 kHz are not effectively captured by the larger spacing of filters in the higher frequency range as shown in the figure 3.

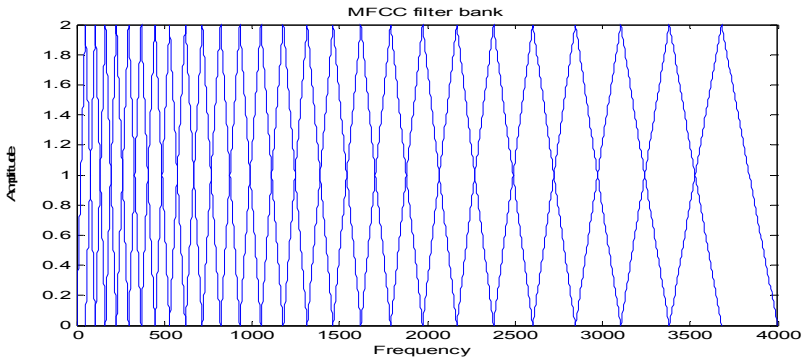

Figure 3: Mel scale filter bank structure.

The, authors in [2-5], have conducted the experiments by inverting the entire filter bank structure; such that the higher frequency range is averaged by more accurately spaced filters and a smaller number of widely spaced filters are used in the lower frequency range. This feature set named as Inverted Mel Frequency Cepstral Coefficients (IMFCC), follows the same procedure as MFCC but use reversed filter bank structure that is complementary in nature to the human vocal tract characteristics described by MFCC. The figure 4 shows the block diagram for Inverted Mel Scale Cepstral Coefficient.

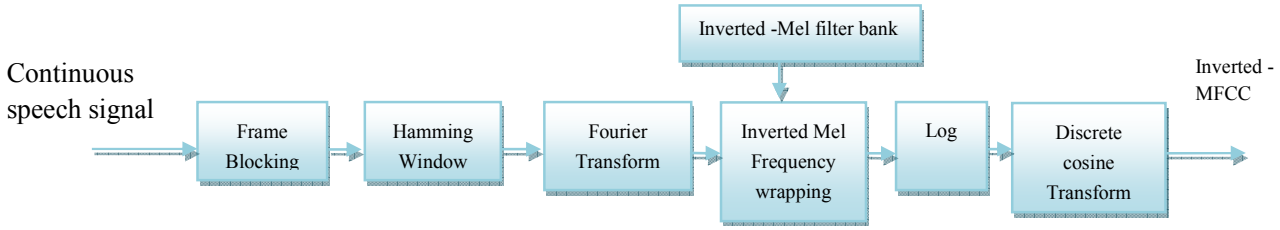

Figure 4: Block diagram for Inverted- Mel frequency cepstral coefficients.

To increase the frequency resolution in the high frequency range, the Mel wrapping function and the inverted Mel wrapping function (for sampling frequency of 8 kHz) the empirical relation (3) & (4) have been used and the inverted mel scale relationship to linear frequency is presented in figure 5 and the inverted mel scale filter bank structure is depicted in figure 6 below.

$$f_{invertedmel} = 2146.1 - 2595 \log_{10}(1+(4000-f)/700) \quad (3)$$

$$f_{invertedmel}^{-1}(f_{invertedmel}) = 2195.2860 - 2595 \log_{10}(1+ 4031.25 - f/700) \quad (4)$$

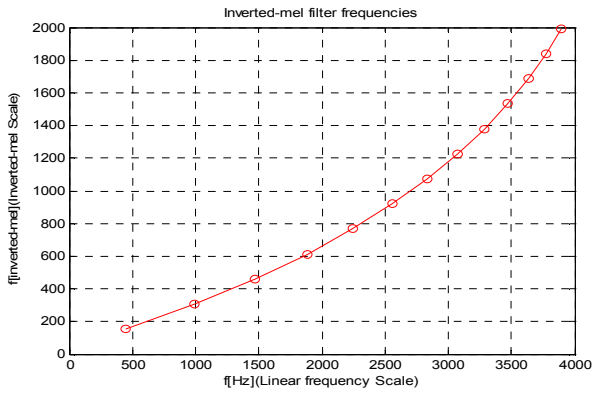

Figure 5: Inverted Mel Scale frequency wrapping.

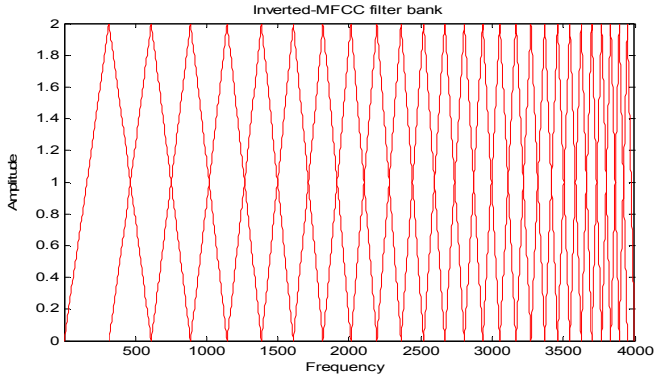

Figure 6: Mel scale filter bank structure.

In usual frequency scale, filters are placed densely in the high frequency range and sparsely in the low frequency range. The figure 7 shows filter bank for (a) Mel scale (b)

Inverted Mel scale, in time domain. Cepstral coefficients are calculated using the inverted Mel filter bank in place of the Mel filter bank. The detailed procedure is given in publication [2-5].

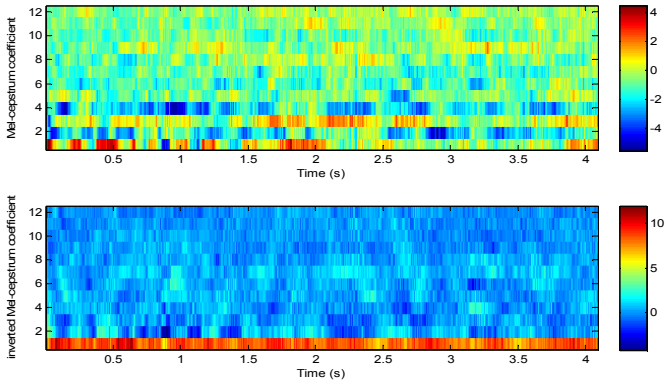

Figure 7: (a) Mel filter bank (b) Inverted Mel Filter bank, in time domain.

The combination of two or more classifiers performs better if they were supplied with information that is complementary in nature [6-8]. MFCC and IMFCC feature vectors, which are complementary in information content, can be fused in order to obtain improved identification accuracy. Number of possible combination schemes such a product, sum, minimum, maximum, median, average etc., can be utilized, but sum rule outperforms the other combination schemes and it is most resilient to estimation errors [6-8].

2. Databases used for Experiments:

Two kind of database were used namely Telephone and Microphone recorded speech for the experiment. The descriptions of the database are as under:-

(i) **Telephone Speech:** The Centre for Spoken Language Understanding (CSLU) speaker Recognition corpus (Release 1.1) was collected from web site: http://cslu.cse.ogi.edu, consists of telephone speech. Each participant has recorded speech in twelve sessions. Each participant calls a toll free telephone number and answers a few question. These files were sampled at 8 kHz, 8-bit. There are 4 speakers (2 males and 2 females); for each speaker, there are 96 utterances. In this work, the 360 (4 X 90 utterances) speeches are used for developing

the speaker model in training mode and 24(6 X 4 utterances) utterances are put under test to evaluate the identification accuracies.

(ii) **Microphone Recorded Speech**: This database is obtained, from the internet, through the speech recording of 5 speakers at 16 kHz sampling rate using Microphone. Further, all speech samples were down-sampled to 8 kHz frequency. For each speaker there are 20 utterances (total 5 x 20 utterances) all are of speech length of approx. 2 to 5 seconds. For this database also, 75 (15 X 5 utterances) speeches are used for developing the speaker model in training mode and 25 (5 x 5 utterances) speeches are put under test to evaluate the identification accuracies.

3. Experiment Setup

The experiment has been set, as shown in the figure 8, to obtain performance of fused MFCC-IMFCC based speaker identification system (for two kind of database as mention above) and for evaluation of system using Detection-Error-Trade off (DET) plots. MFCC, IMFCC and MFCC-IMFCC, based GMM parallel fused classifier were created in Matlab. A Gaussian Mixture Model (GMM) based classifier is used which provides an unsupervised clustering technique to model the speakers. For Each speech, 12 numbers of Gaussian mixture features set has been generated and the scores (obtained from MFCC and IMFCC based SI System) are fused, using sum rule. For the i^{th} speech, the combined score S^i_{com} can be expressed as (5).

$$S^i_{com} = wS^i_{MFCC} + (1\text{-}w)\, S^i_{IMFCC} \qquad (5)$$

Where S^i_{MFCC} and S^i_{IMFCC} are the scores generated by the two models, MFCC and IMFCC, respectively and where **w** is the fusion coefficient.

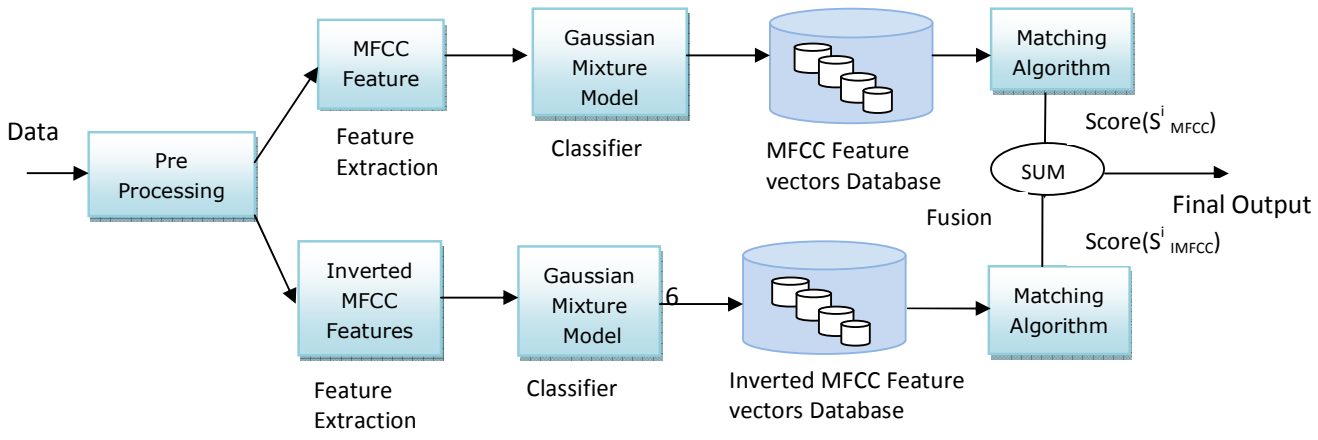

Figure 8: MFCC-IMFCC fused Speaker identification System.

The performance of the fused system has been obtained for both the databases. Thereafter, the performance of fused speaker identification system, for two different kind of speech corpus, for analysing the effect of fusion coefficient for MFCC and IMFCC features is evaluated using DET plots.

4. Results & Discussion

DET performance curve has been obtained for MFCC, IMFCC and fused MFCC-IMFCC for both the database, as mentioned above. The figure 6(a) shows the speaker detection performance for MFCC, IMFCC and MFCC-IMFCC (with fusion coefficient 0.5) obtained using telephone speech. The figure 6(b) shows the speaker detection performance for MFCC, IMFCC and MFCC-IMFCC (with fusion coefficient 0.5) obtained using microphone Speech.

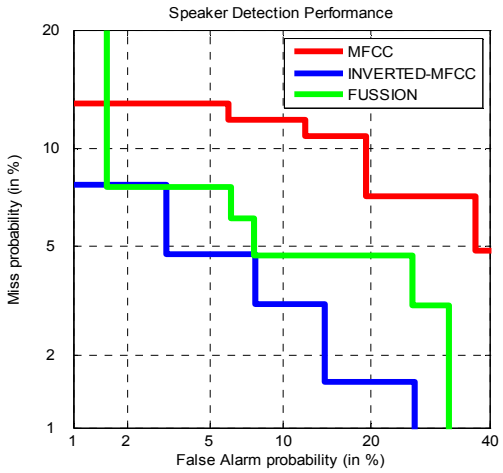

Figure 6(a): DET curve for MFCC, IMFCC and fused MFCC-IMFCC (with fusion coefficient 0.5) for Telephonic speech database.

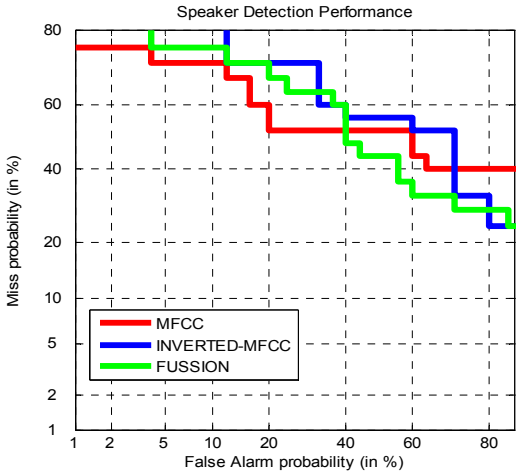

Figure 6 (b): DET curve for MFCC, IMFCC and fused MFCC-IMFCC for Microphone speech database.

Table 1: Equal Error Rate for MFCC, IMFCC and Fused Speaker Detection System.

Database	MFCC System (% EER)	IMFCC System (% EER)	MFCC-IMFCC Fused System (% EER)
Telephone Speech	19%	7.9%	7.9%
Microphone Speech	55%	60%	48%

Speaker identification system performance results, using MFCC, IMFCC and fused MFCC-IMFCC fusion based features set, equal error rate parameter, are summarized in Table 1, for both databases. It may be seen that the combined scheme shows significant improvements in SI system over MFCC based system alone, for both Microphone Database and Telephone speech. Especially for telephone speech database, the independent performance of the IMFCC based classifier is comparatively better to that of the MFCC based classifier.

The figure 7(a) shows the performance for the fusion of MFCC-IMFCC using various fusion coefficients, obtained using telephone speech and figure 7(b) shows the performance for MFCC-IMFCC based classifier using various fusion coefficients, obtained using microphone Speech. The DET plot shows the miss probability against the false alarm

probability: Tables 2 below gives the comparative performance based on different combination of fusion.

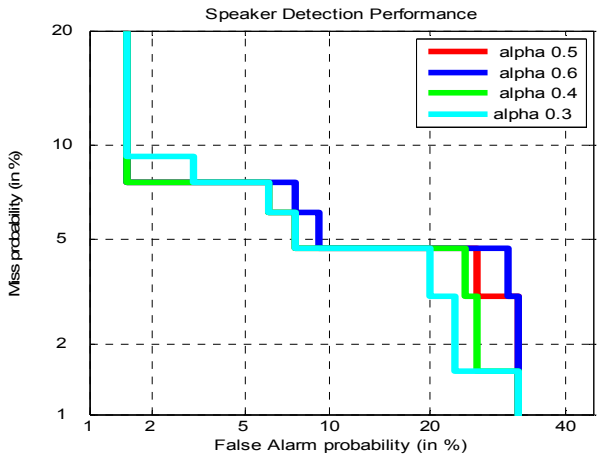

Figure 7(a): DET curve for Telephonic speech database, with various fusion coefficients.

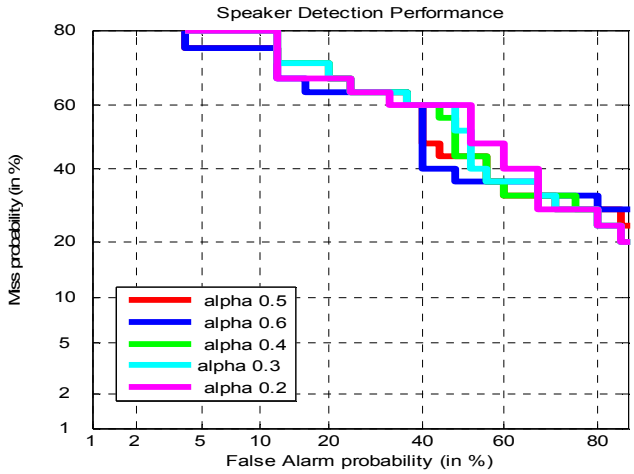

Figure 7(b): DET curve for Microphone speech database, with various fusion coefficients.

Table 2: Equal Error Rate for MFCC-IMFCC fusion with various fusion coefficients.

Database	w=0.5 (% EER)	w=0.6 (% EER)	w=0.4 (% EER)	w=0.3 (% EER)	w=0.2 (% EER)
Telephone Speech	7.9%	9%	8.1%	7.8%	21%

Microphone Speech	48%	40%	49%	51%	47%

Individual MFCC, IMFCC and fused MFCC-IMFCC with different fusion coefficient were used for both databases. It may be seen that for the used telephone speech database, the fusion coefficient 0.3 outperforms the speaker identification system and for used Microphone speech database fusion coefficient 0.6 has given enhanced the system performance. Same can also be established from the DET plots obtained through fusion using equal contribution of MFCC and IMFCC.

5. CONCLUSION

The IMFCC feature based classifier can provide improved accuracy for telephone speech database, by proper choice of mixing proportion of two streams in combined model. The study reveals that in order to improve the performance of the speaker identification system, for telephonic speech database the contribution of IMFCC should be more as comparable to MFCC. This is because of the fact that bandwidth in telephone channel is limited. On the other hand, for Microphone speech the contribution of MFCC should be large. The appropriate selection of the fusion coefficient, in order to improve the accuracy of the system, can be used by the DET plots for any kind of database.

6. REFERENCES

1. *J. Campbell*, "Speaker recognition: a tutorial", Proceedings of the IEEE VOL. 85, NO. 9, pp. 1437–1462, September 1997.
2. *J. Kittler*, "Combining Classifiers: A Theoretical Framework", Pattern Analysis & Applied Springer-Verlag London Limited, Issue 1, pp.18-27, 1998.
3. *J. Kittler, M. Hatef, R. Duin, and J.Matas*. "On combining classifiers". IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 20, issue 3, pp 226–239, March 1998.
4. *J. Kittler, F.M. Alkoot*, "Sum Versus Vote Fusion in Multiple Classifier Systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 25, Issue 1, pp.110 – 115, January 2003.

5. *Sandipan Chakroborty, Anindya Roy and Goutam Saha*, "Improved Closed Set Text-Independent Speaker Identification by combining MFCC with Evidence from Flipped Filter Banks" , International Journal of Information and Communication Engineering volume 4, issue 2 , 2008.
6. *Sandipan Chakroborty, Goutam Saha, "*Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter" International Journal of Signal Processing Volume 5 issue 1, 2009.
7. *Tomi Kinnunen, Haizhou Li, "*An overview of text-independent speaker recognition: from features to supervectors", Speech Communication volume 52, pp 12-40, 2010.
8. *Nirmalya Sen, Tapan Basu, Sandipan Chakroborty*, "Comparison of Features extracted Using Time-Frequency and Frequency-Time Analysis Approach for Text-Independent Speaker Identification", IEEE National conference on Communication, pp.1 - 5 , 30 Jan. 2011.
9. *Satyanand Singh, Dr. E.G. Rajan*, "Vector Quantization approach for Speaker Recognition using MFCC and Inverted MFCC", International Journal of Computer Applications Volume 17, issue 1, pp. 0975- 8887, March 2011.

AUTHOR

Ruchi Chaudhary, received M.Tech degree in the year 2009 from Guru Govind Singh Indraprasth University, Kashmiri Gate, Delhi, and in 2002, B.Tech Degree in Electronics & Communication Engineering from CJSM Kanpur University. In 2003, she joined Defence Research & Organisation as Junior Research Fellow, and in 2004, she joined Guru Prem Sukh Memorial College of Engineering as a Lecturer in the Department of Electronics & Communication and subsequently became Head of Department of ECE in the same Institution in 2007. She is presently working as a Scientist in Government Organization and pursuing PhD from Guru Govind Singh Indraprastha University. Her interest includes Speech Processing and Soft Computing Techniques. She has also contributed in Research paper of International Journal of Sensors & Actuated in 2004 on Pattern Recognition Techniques.