IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 5, Issue 2, Ver. II (Mar. - Apr. 2015), PP 53-60
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

Speaker Identification Based On MFCC and IMFCC Using GMM-UBM

Anagha S. Bawaskar[1], Prabhakar N. Kota[2]

[1](*Department of Electronics and Tele-Communication, M.E.S.C.O.E College, Savitribai Phule University of Pune, India*)

[2](*Department of Electronics and Tele-Communication, M.E.S.C.O.E College, Savitribai Phule University of Pune, India*)

Abstract: *In Speaker identification (SI) systems long-lasting feature extraction unit is required. For the purpose of proper representation of this features, there is a speaker modeling scheme after extraction unit. Several feature sets are used for speaker related application, one of the standard feature set is MFCC (Mel Frequency Cepstral Coefficient) which are generally modeled on human auditory system. On the other hand complementary information present in the higher frequency range known as an IMFCC (Inverted Mel Frequency Cepstral Coefficient) is another feature set which is useful. This paper concentrates on Gaussian Mixture Model (GMM)along with the Universal Background Model (UBM) is being used for the modeling purpose. Instead of triangular filters here Gaussian shaped filters are being used. Here , in this paper the results are being verified by the standard database TIMIT. The accuracy for the individual set of features such as for MFCC is coming to be 96.6% for the 16 mixtures while for the IMFCC is 95.4% for the 16 mixtures respectively for first set of speakers used and on other side the accuracy for MFCC and IMFCC in the other set of speakers is 97.22% and 86.11 respectively.*

Keywords: *Fast Fourier Transform (FFT), Gaussian Mixture Model (GMM), Inverted Mel Frequency Cepstral Coefficients (IMFCC), Mel Frequency Cepstral Coefficients (MFCC), Universal Background Model (UBM).*

I. Introduction

Various measurements and signals have been preferred for investigation which is used in biometric recognition system. The many of biometric system are fingerprint recognition, face and voice. There are some advantages and disadvantages of above mentioned some biometric system. There are two main aspects which make voice a compulsory factor in biometric [**1**].

Speech processing can be used in various applications for voice identification in ordinary personal computers to biometric and forensic applications. Speech processing consists of two main techniques. One is speaker recognition and the other is speech recognition. This paper is concentrating more on speaker recognition. In this it uses the audio features that have found in many different individuals. The sound produced by an individual is different for different individuals as a shape of a vocal tract is different for different individual. The shape of vocal tract is generally important physiology distinguishable factor of speech. The audio pattern of the voice reflect the structure of vocal tract and learned behavioral patterns such as voice pitch, style of speaking, etc.The speaker recognition is further classified into two major categories namely speaker identification and speaker verification [2].For representation of extracted features, Speaker identification system have a front end and the back end, the feature extractor is used as a front end module while the back end is consisting of particular modeling technique, here in this case we have used GMM-UBM for the modeling purpose. The specific feature from the speaker is being extracted by using MFCC for finding out accurate speaker. Investigations by the researchers find out speaker specific complementary information relative to MFCC that are called as Inverse Mel Frequency Cepstral Coefficients (IMFCC) respectively. For combining score models, complementary information is being used along with the MFCC. These models are nothing but the mathematical representation of the particular system [3].The inverse bank method is being used for capturing this complementary information from high frequency part of energy spectrum. The information which is neglected by MFCC is captured by IMFCC. Thus; Gaussian Mixture Model is used for modeling the MFCC and IMFCC [2] .The diagram for typical speaker identification is shown in fig1.

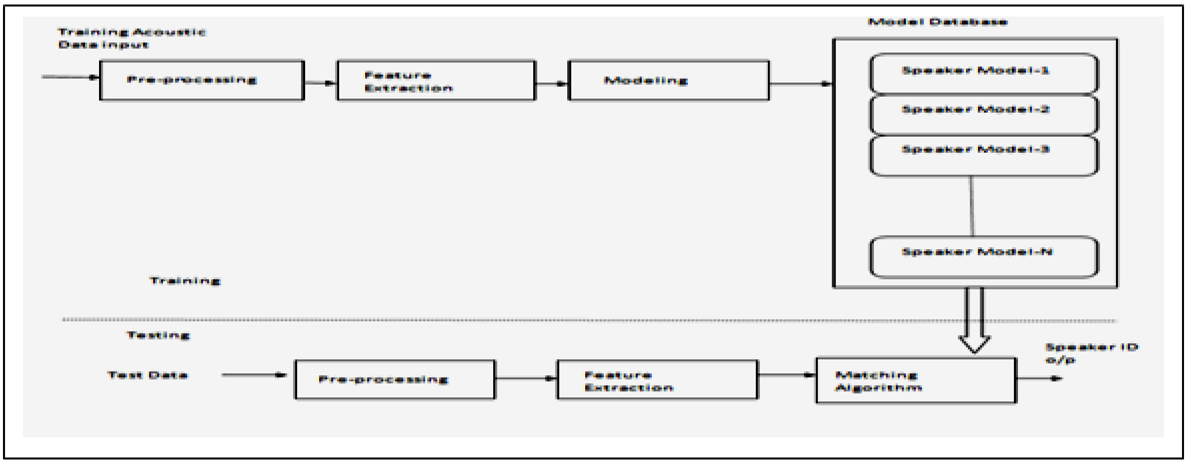

Fig 1. A Typical Speaker Identification System [5]

Generally, for standard purpose the triangular filters are being used for filtering the spectrum of speech signal that simulates the characteristics of human ear. But; as it is used everywhere still it got some disadvantage that it provides sharper or crisp partition in an energy spectrum. Due to these the information of the adjacent sub bands is lost. Paper particular focus on the Gaussian filters based on MFCC and IMFCC. The crisp and sharp transition in an energy spectrum is avoided if we used the Gaussian filters instead of triangular. This results in smoother adaptation from one sub band to other. This preserves most of the correlation between them. From the mid points located at the base of triangular filters as well as from the endpoints of the same, calculations of Gaussian filters is simple and are used for MFCC and IMFCC [5].

Better results are being shown by the Gaussian filters based on MFCC and IMFCC respectively instead of the triangular filters. The speaker identification is done here in the standard database TIMIT. For mathematically representation model.

The rest of the paper is assembled as follow:

Section II explains concept of the MFCC and its implementation using Gaussian filters. Section III explains the IMFCC and its implementation using Gaussian filters and Section IV explains the concept of Gaussian Mixture Model and Universal Background Model. Section V explains the experimental results. Section VI gives the experimental evaluation and VII concludes the paper.

II. Mel Frequency Cepstral Coefficients (MFCC) By Gaussian Filters

The most important step in speaker identification is feature extraction. The features if not chosen carefully we will not get good results even if the best classifier is chosen. The set of features which are used by the classifiers for training actually influence the percentage of accuracy of speaker recognition as well as for speaker identification. The original waveform is being divided into smaller set of information during feature extraction. In speaker recognition, the feature vector is an dimensional vector of features extracted from audio files to be used for further process in identification of speaker such as training and testing respectively. The MFCC is used for both speaker recognition as well as for speech recognition. In this paper it is used for speaker identification. MFCC is based on human hearing ability. They cannot perceive frequencies over 1Khz. In other words, MFCC can be defined as known variation of the human ear's critical bandwidth with frequency. Two types of filter are used in MFCC. One is spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz.

Mel Scale: The MFCC name is being given from the small unit called Mel Scale respectively. The Mel scale is said as an affective scale of pitches, this is being examined by listener to be equal in distance from one another. From the word melody the Mel scale is being defined that is we can say origin of the word Mel comes from the word melody. The scale is based on pitch comparisons. The note should be taken here that the points are at equal distance, apart in the Mel has higher resolution at low frequencies. These features are long lasting and dependable to the changes according to speaker and recording conditions.

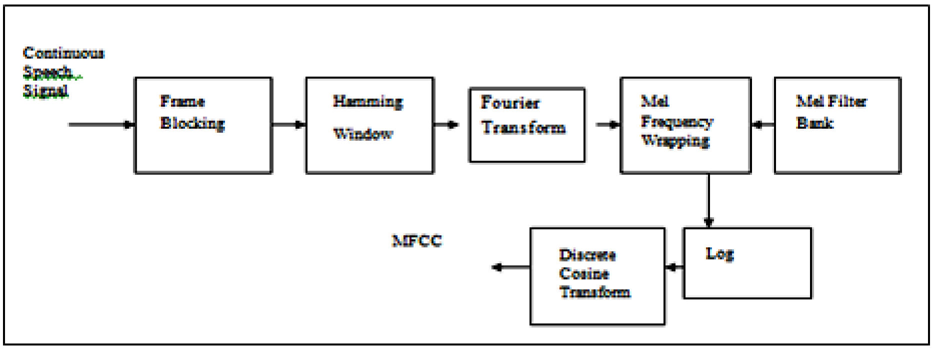

Fig2. Block Diagram for Inverted Mel Frequency Cepstral Coefficients [6]

From the block diagram in fig 2, we can say that, input is continuous speech signal. This signal is given as input to the frame blocking block. This block consists of frame, which are made up of speech signal of random number of samples depending on the user. Here; overlapping frames can also be used which smoothens transition from frame to frame. The next block is windowing block. The output of framing block serves as an input to this block. The technique which is generally used is hamming windowing technique. Hamming window eliminates the discontinuities at the edges. It is represented as w (n).This output of the windowing block is given as an input to FFT block. Each frames FFT is being calculated. The functioning of FFT is helpful for converting frequency domain to time domain. FFT also helps for speeding up the process [7].

Now; for the values which we get might be large in number, so needs to be compressed, for this purpose logarithmic Mel scale filter bank is applied to frame which is Fourier transformed frame. The importance of Mel scale is that, it is linear up to 1 KHz and logarithmic above 1 KHz. The relation between frequency of speech and Mel scale can be established as:

$$f_{mel} = 2595\log_{10}(1+ {}^{f}\!/_{700})......(1)$$

Where f_{mel} is the intuitive pitch in Mel's corresponding to f the actual frequency in Hz. This gave rise to the definition of MFCC, a basic acoustic or audio feature for speech as well as speaker recognition application.

2.1 MFCC features using Gaussian filter bank

For general purpose, MFCC uses triangular filter, but triangular filters are asymmetric tapered and also not provide any correlation between sub bands and its nearby spectral components. The information lost occurred there. Therefore, for avoiding this loss, Gaussian filters are used, which are tapering towards both the end and provide correlation between sub bands and its nearby spectral components. The mathematical equation for the Gaussian filter is written as [5].

$$\psi_i^{g_{MFCC}} = e^{-\frac{(k-k_{bi})^2}{2\sigma_i^2}} \text{.........................(2)}$$

Where; k_{bi} is a point between the i^{th} triangular filters boundary located at its base and considered as mean of i^{th} Gaussian filter while the σ_i is the standard deviation or square root variance of and can be defined as,

$$\sigma_i = \frac{k_{b_{i+1}} - k_{b_i}}{\alpha} \text{ .... (3)}$$

Where; α is the parameter that controls the variance.

In the figure 3, two plots are shown in single figure. One is for triangular filter and the other is for the Gaussian filter. This plot is made by considering a single value of sigma. Here in this case plot can be drawn by considering different values for sigma respective. Fig 4 and Fig 5 shows the individual response for triangular filter bank and the Gaussian filter bank respectively.

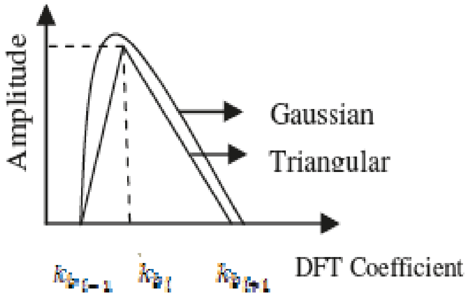

Fig3. Filter bank design [2]

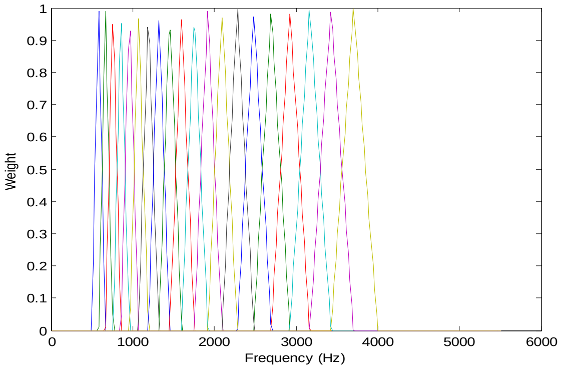

Fig 4 Mel Scale Triangular filter bank

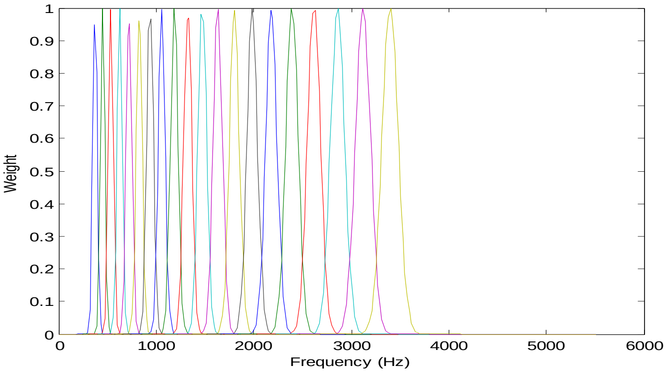

Fig 5.Mel Scale Gaussian filter bank

These filter banks are being forced on the energy spectrum obtained by taking Fast Fourier transform of the preprocessed signal as follow:

$$e^{g_{MFCC}}(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(k)|^2 \,.\, \hat{\psi}_i^{g_{MFCC}}(k) \text{ .... (4)}$$

Where; $\psi_i(k)$ is respective filter response and $Y(k)^2$ is the energy spectrum. Finally, DCT is taken on the log filter bank energies $\{\{\log[e(i)]\}_i^Q\}$ and the final MFCC coefficients can be written as-

$$C_m^{g_{MFCC}} = \sqrt{\frac{2}{Q}} \sum_{l=0}^{Q-1} \log[e^{g_{MFCC}}(i+1)] \cdot \cos[m \cdot (\frac{2l-1}{2}) \cdot \frac{\pi}{Q}] \text{ ….. (5)}$$

Where; $0 \le m \le R-1,\ R$ is the desired number of cepstral features.

III. Inverted Mel Frequency Cepstral Coefficients (IMFCC) By Gaussian Filters

From [8], IMFCC (Inverted Mel Frequency Cepstral Coefficient) is defined as one of the characteristics property of the audio system. As the name conveys the meaning, it follows the opposite path of evolution of human audio system. The basic idea is to capture the information which has missed by the original MFCC. The structure of the filter bank is been inverted by the researchers for performing various experiments. Thus By doing this, higher frequency range is average by more accurately spaced filters and small number of widely spaced filters used in lower frequency range. Such a feature set is named as Inverted Mel frequency Cepstral Coefficients (IMFCC). The procedure for IMFCC is somewhat same as that of MFCC but used reverse filter bank structure. The nature of IMFCC is reciprocal in nature as compared against MFCC. The figure 4 shows the block diagram for Inverted Mel Scale Cepstral Coefficient [4].

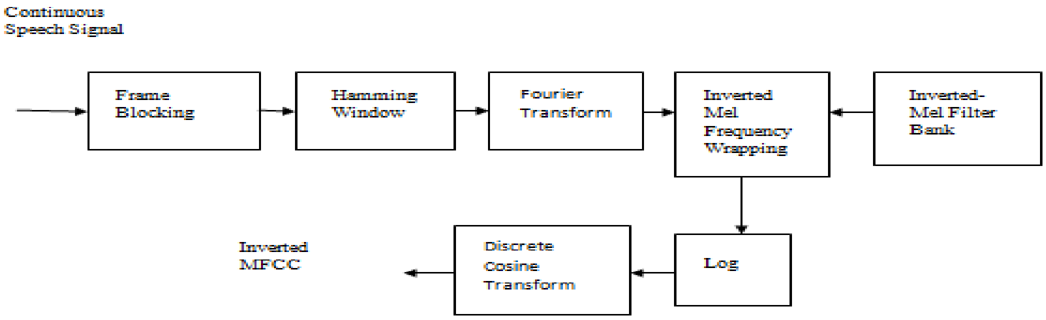

Fig 4. Block Diagram for Inverted Mel Frequency Cepstral Coefficients [4]

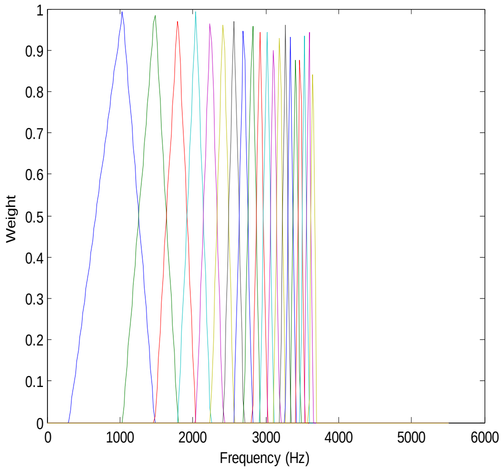

Fig. 5 Inverse Mel Scale Triangular Filter bank

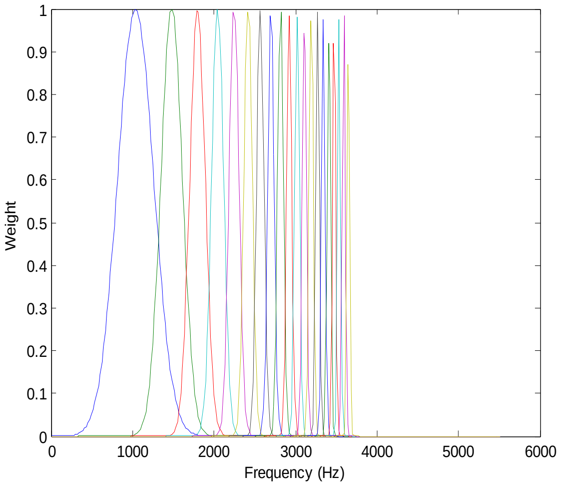

Fig 6. Inverse Mel Scale Gaussian Filter Bank

For increasing the frequency resolution in higher resolution range, Mel wrapping function and inverted Mel wrapping function are used. Here, Mel scale relation is converted to the linear frequency.
Mathematically;

$$f_{mel}^{-1}(f_{mel}) = 700(10^{fmel/2595}-1).......(6)$$

Where, $f_{mel}^{-1}(f_{mel})$ is the subjective pitch in the new scale corresponding to the f_{mel}, the actual frequency in Hz.

3.1 IMFCC features using Gaussian filters

IMFCC features are also modeled by using Gaussian filter bank. The Gaussian filter for IMFCC filter bank and corresponding cepstral parameters can be calculated as:

$$\hat{\psi}_i^{g_{IMFCC}} = e^{-\frac{(k-\hat{k}_{b_i})^2}{2\hat{\sigma}_i^2}} \qquad .... (7)$$

$$\hat{e}^{g_{IMFCC}}(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(k)|^2 .\hat{\psi}_i^{g_{IMFCC}}(k) \qquad .... (8)$$

And finally;

$$\hat{C}_m^{g_{MFCC}} = \sqrt{\frac{2}{Q}}\sum_{l=0}^{Q-1}\log[e^{g_{IMFCC}}(i+1)]\cdot\cos[m\cdot(\frac{2l-l}{2})\cdot\frac{\pi}{Q}]. \qquad .... (9)$$

The filter bank structure for IMFCC using triangular filters and IMFCC using Gaussian filters are shown above in the Fig. 5 and Fig. 6 respectively.

IV. Gaussian Mixture Model (GMM)-Universal Background Model (UBM).

A Gaussian Mixture Model (GMM) is one of the parametric models. In GMM we find out the probability of all the Gaussian components densities. Biometric system is one of the application area where GMM are used. GMM parameters are estimated from training data using Expectation-Maximization (EM) algorithm or Maximum a Posteriori (MAP) estimation from a well-trained prior model [9]. The expected maximization algorithm is iterative in natures

GMM are generally used for text independent speaker identification. The drawback of the previous systems is being overcome by using GMM-UBM. It overcomes on the cost of the mode; it is not as expensive that of the GMM. There is no need for the vocabulary database or big phoneme. GMM is more advantageous than HMM.

Capturing the general characteristics of a population and accordingly adapting it to individual speaker is the basic idea of UBM. In other words more briefly UBM is defined as the model which is used in many application area but one of them is biometric system which is used to compare the person's independent feature characteristics against person specific feature model during decision of acceptance or rejection. UBM is also said as GMM only with large set of speakers. Firstly likelihood score or ratio for an unknown speech sample is found after that the match score of speaker specific mode and universal background model is formed by using speaker specific GMM trained samples from a particular speaker. UBM is used as a prior model in MAP parameter estimation.. For reduction in the computational complexity of the system many of the researchers tried many methods. The name GMM for text independent speaker identification has created interest of many researchers in past few years. Application and accuracy are the two criteria for selection of proper system. The GMM shows best performance as compared to the other known method and explore many new ideas to improve the performance of the system.

A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation;

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i g(x \mid \mu_i, \Sigma_i) \text{ ……………….. (10)}$$

Where; x is a D-dimensional continuous-valued data vector (i.e. measurement or features),

$g(x \mid \mu_i, \Sigma_i)$, i = 1….. M are the mixture weights, and

w_i , i = 1… M is the component Gaussian densities. Each component density is a D-variate Gaussian function of the form;

$$g(x \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\} \text{ …… (11)}$$

With mean vector μ_i and covariance matrix Σ_i the mixture weights satisfy the constraint that $\Sigma_{i=1}^{M}\omega_i = 1$ The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation,

$$\lambda = \{\omega_i, \mu_i, \Sigma_i\} \text{ , i =1 … M ……… (12)}$$

For a sequence of T training vectors $X = \{x_1, ...x_T\}$.The GMM likelihood, assuming independence between the vectors, can be written as $p(X \mid \lambda) = \prod_{t=1}^{T} p(x_t \mid \lambda)$ …………….. (13)

For utterances with T frames, the log-likelihood of a speaker models is;

$$L_s(X) = \log p(X \mid \lambda_s) = \sum_{t=1}^{T} \log p(x_t \mid \lambda_s)\text{........(14)}$$

For speaker identification the value of $L_s(X)$ is computed for all speaker models λ_s enrolled in the system and the owner of the model that generates the highest value is the returned as the identified speaker. During training phase, Feature vectors are being trained using Expectation and Maximization (E&M) algorithm. An iterative update of each of the parameters in λ , with a consecutive increase in the log likelihood at each step.

V. Experimental Evaluation

5.1Database for Experimentation

TIMIT:

TIMIT corpus is one of the standard databases used by the many researchers for the purpose of speaker identification. This paper also concentrates on the TIMIT database. It comprises of the 16 speakers. The recordings are from 8 dialect regions. Each speaker has 10 utterances respectively Total 160 sentences recordings (10 recordings per speaker). The audio format is wav format, single channel, 16 kHz sampling, 16 bit sample, PCM encoding.

5.2Experimental Results

(1)For this part we have taken the 16 speakers in total for speaker identification. In that we have divided them as 5 speakers data for the universal background model and the remaining 11 speakers data for testing purpose. The value of the alpha that is filter constant is kept as 0.97 respectively. The following table shows the performance accuracy for both MFCC and IMFCC. We have tested according to the description stated above and got the results as follows:

Table I. Results for MFCC and IMFCC using Gaussian filter in TIMIT database.

(a)Score Threshold=0.6

No. of Mixtures	MFCC	IMFCC
4	95.0413%	92.5620%
8	95.0413%	93.3884%
16	96.6942%	95.0413%
32	93.3884%	95.0413%

(b) Score Threshold=0.77

No .of Mixtures	MFCC	IMFCC
4	92.5620%	91.7355%
8	93.3884%	92.5620%
16	95.0413%	93.3884%
32	95.8678%	94.2149%

(c)Score Threshold=0.8

No. of Mixtures	MFCC	IMFCC
4	92.5620%	91.7355%
8	93.3884%	92.5620%
16	93.3884%	93.3884%
32	95.0413%	94.2149%

(2) For the second part, we have taken out off 16 speakers 10 speakers as an universal background model and the remaining 6 speakers data for the testing purpose for finding the accurate speaker. The following table gives the accuracy percentage for the both MFCC and IMFCC respectively.

Table II. Results for MFCC and IMFCC using Gaussian filter in TIMIT database

(a)Score Threshold=0.6

No. of Mixtures	MFCC	IMFCC
4	86.1111%	83.3333%
8	86.1111%	83.3333%
16	94.4444%	86.1111%
32	97.2222%	86.1111%

(b) Score Threshold=0.77

No. of Mixtures	MFCC	IMFCC
4	86.1111%	83.3333%
8	86.1111%	83.3333%
16	88.8889%	83.3333%
32	88.8889%	83.3333%

(c)Score Threshold=0.8

No. of Mixtures	MFCC	IMFCC
4	83.3333%	83.3333%
8	83.3333%	83.3333%
16	86.1111%	83.3333%
32	86.1111%	83.3333%

From the TABLE I we can see that the various percentage for the MFCC and IMFCC. As we vary the threshold, the accuracy percentage also varies. But for the threshold value 0.6, we get better accuracy as compared to others. These percentages give us the comparative accuracy for both MFCC and its complementary information respectively. The individual performances of the IMFCC are good but not as that of the MFCC.The complementary information helps the MFCC to improve the performance of MFCC. From the table we can say that the MFCC percentage is better than the IMFCC. Overall the performance and accuracy is good if we used the Gaussian filters instead of triangular filters.

From the TABLE II we can see that the accuracy slight changes as the score threshold changes. From this we can say that by changing the training and testing number of samples we get different results for the three typical values of thresholds respectively. Better results are seen for the values of threshold 0.6 that is 97.22% for MFCC accuracy is measured and while for IMFC the accuracy percentage is 86.12%.

VI. Conclusion

In this paper main focus were given on the study of the MFCC and IMFCC features by using the Gaussian filters respectively. These vectors were modeled by using GMM. This provides good balance between filters coverage area and amount of correlation. The performance if we compare the both MFCC and IMFCC, the MFCC is better than IMFCC. But the IMFCC helps MFCC for improving the accuracy also. As we see the individual performance for both, are good. For the Future work the performance can be improved by combination of these two vectors MFCC and IMFCC respectively. These combining of the two features are called as Fused Mel Feature Set. In future experimentation we will get to know about if any further improvement in the accuracy or not.

References

[1]. Fr ´ed´ eric Bimbot,1 Jean-François Bonastre,2 Corinne Fredouille,2 Guillaume Gravier,1 Ivan Magrin-Chagnolleau and Douglas A. Reynolds6 "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal Processing 2004:4, 430–451c_ 2004 Hindawi Publishing Corporation.

[2]. R.Shantha Selva Kumari a, S. Selva Nidhyananthan ba*, Anand,"Fused Mel Feature sets based Text-Independent Speaker Identification using Gaussian Mixture Model".G c a,bDepartment of ECE, Mepco Schlenk Engineering College , Sivakasi -626005, INDIA cDepartment of ECE, PSRR College of Engineering for Eomen, Sevalpatti,INDIA. International Conference on Communication Technology and System Design 2011

[3]. J. Kittler, M. Hatef, R. Duin, J. Mataz, "On combining classifiers", IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 226-239

[4]. Zheng F., Zhang, G. and Song, Z., "Comparison of different implementations of MFCC", J. Computer Science & Technology, vol.16 no. 6, pp. 582-589, Sept. 2001.

[5]. International Journal of Information and Communication Engineering, Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter, Sandipan Chakroborty* and Goutam Saha

[6]. Ruchi Chaudhary, National Technical Research Organization, "Performance study of Text-independent Speaker identification system using MFCC & IMFCC for Telephone and Microphone and Speeches" (IJCBR) Vol 3, 2 May 2012.

[7]. Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speeh Recognition" G H Patel College of Engineering, Gujarat Technology University, INDIA, IJART.

[8]. Yegnanarayana B., Prasanna S.R.M., Zachariah J.M. and GuptaC.S.,"Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system", IEEE Trans.Speech and Audio Processing, Vol. 13, No. 4, pp. 575-582, July 2005

[9]. Signal & Image Processing: An International Journal (SIPIJ) Vol.4, No.4, August 2013DOI : 10.5121/sipij.2013.4409 109 ,A GAUSSIAN MIXTURE MODEL BASED SPEECH RECOGNITION SYSTEM USING MATLAB,Manan Vyas B.E Electronics, University of Mumbai mananvvyas@gmail.com