In a *text-independent system,* training and testing speech is completely unconstrained.[1, 2, 5, 7, 8] Beyond from a text-independent system, there is a *language-independent system* in which non-specified language is used.[3] In this paper we minimize misrecognitions errors for better identification by using two modes: training mode and identification mode.[7]

2. Feature Extraction

Feature extraction is a special form of dimensionality reduction, and here we need to do dimensionality reduction for the input speech for less complexity.[2] Features of speech waveform [2] should be-

- Easily measurable.
- Occur naturally and frequently in speech.
- Vary much as possible among the speakers, but be consistent within each speaker.
- Not be affected by background noise nor depend on the specific transmission medium.

Here, we use formant as a feature which carries the identity of the speech. The form and shape of the vocal and nasal tracts change continuously with time, creating an acoustic filter with time-varying frequency response.[9] The resonance frequencies of the vocal tract tube are called *formant frequencies* or simply *formants*, which depend on the shape and dimensions of the vocal tract.[1] The speed by which the cords open and close, is unique for each individual and define the feature and personality of the particular voice.[9] So, we want to extract these formants (the amplitudes of the speech wave form).

2.1. *MFCC algorithm*

This is the block diagram for the feature extraction processes applying MFCC algorithm:

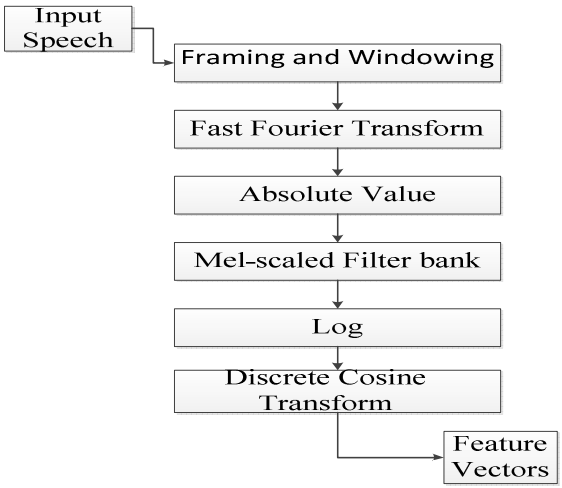

Fig. 2: process of computing MFCC from speech signal

The *mel-frequency* scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz.[7] It is based on perception of speech by human ears. Human ears hear the tones for frequencies lower than 1 kHz, with a linear scale instead of logarithmic scale for the frequencies higher than 1 kHz. We can use the following formula to compute the mels for a given frequency *f* in Hz:

$$mel(f) = 2595 * \log_{10}(1 + f / 700) \quad \text{——— (1)}$$

The information carried by low frequency components of the speech signal is more important compared to the high frequency components and mel filter-banks give more filters in the low frequency regions and less number of filters in high frequency regions.[6] So the signal is processed in such away like that of human ear response:

$$\tilde{S}(l) = \sum_{k=0}^{N/2} S(k) M_l(k) \quad \text{——————— (2)}$$

Where:
S(l): Mel spectrum; S(k): Original spectrum; M(k): Mel filter-bank; $l = 0, 1, \ldots\ldots\ldots\ldots, L-1$, Where l is the total number of mel filter-banks and N/2 = Half FFT size.
The cepstrum is a representation of the signal where these two components are resolved into two additive parts.

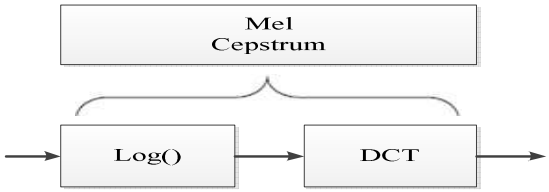

Fig. 3: Mel-Cepstrum

In the final step, the log mel spectrum has to be converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs).[3] The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis.

3. Classification and Feature Matching

There are two major types of models for classification stochastic models (parametric) and template models (non-parametric).
In stochastic models, the pattern matching is probabilistic (evaluating probabilities). The results in a measure of the likelihood, or conditional probability, of

the observation given the model.[1] Stochastic models provide more flexibility and better results. It includes: Gaussian mixture model (GMM), Hidden Markov Model (HMM) and Artificial Neural Network (ANN).
For template models, the pattern matching is deterministic (evaluating distances). Template models are considered to be the simplest ones. It includes: Dynamic Time Warping (DTW) and Vector Quantization (VQ) models.
In regard to the choice of the classification method, the kind of application of the speaker recognition system is crucial. For text dependent recognition and text independent recognition, template models are suitable [8] and for language- independent recognition the more advanced Gaussian mixture models (GMMs) are used most often.[4, 7, 10] In this paper, we chose to use GMM approach due to its accuracy and robustness.

3.1. *Gaussian Mixture Model (GMM)*

The pdf of the observed spectral features generated from a statistical speaker model is a Gaussian mixture model (GMM). For the identification each speaker is represented by his/her GMM, [7] which is parameterized by all component densities.
In terms of the parameters of an M-state statistical speaker model, the GMM pdf is

$$p(x \setminus \lambda) = \sum_{i=1}^{M} p_i b_i(x) \qquad \text{——————} \quad (3)$$

Where,

$$\lambda = (p_i, \mu_i, \sum_i), \text{ for i=1,…,M}$$

represents the parameters of the speaker model. Thus the probability of observing a feature vector coming from a speaker model with parameter A is the sum of the probabilities that was generated from each hidden state, weighted by the probability of being in each state. With this summed probability we can produce a quantitative value, or score, for the likelihood that an unknown feature vector was generated by a particular GMM speaker model.[6] Despite the apparent complexity of the GMM, model parameter estimates are obtained in an unsupervised manner by using the expectation-maximization (EM) algorithm.[10]

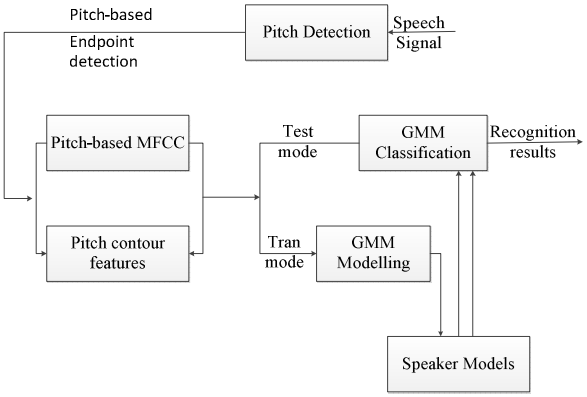

Fig. **4**: GMM-based speaker recognition system.

In GMM-based speaker recognition, the binary decision to accept or reject a claimed identity is based on the likelihood score. The likelihood ratio of the claimed speaker A is,

$$\lambda_A(Z) = \frac{p_A(Z \setminus H_1)}{p_A(Z \setminus H_0)} \qquad \text{——————} \quad (4)$$

If $\lambda_A(Z) \leq T$, choose H0, H1 otherwise. The variable T is the acceptance or rejection threshold. The decision threshold is located at the point where the probabilities of both the errors are equal.[1] In the identification phase, speaker with maximum likelihood is selected as the author of a speech sample.

3.2. *E.M. Algorithm for GMM parameters estimation*

An initial model can be obtained by estimating the parameters from the clustered feature vectors whereas proportions of vectors in each cluster can serve as mixture weights. After the estimation, the feature vectors can be reclustered using component densities (likelihoods) from the estimated mixture model and then model parameters are recalculated. This process is iterated until model parameters converge. This algorithm is called Expectation Maximization (EM).[10] This is a two-step process for finding optimal solutions.

E-step (Expectation): Use the current values of the parameters to evaluate the responsibilities.

M-step (Maximization): Re-estimate the distribution parameters (using the current responsibilities) to maximize the likelihood of the data.

By using E.M. algorithm, increase in likelihood function is guaranteed i.e. parameters are estimated more accurately.

4. Decision Process

The next step after computing of matching scores for every speaker model enrolled in the system is the process of assigning the exact classification mark for the input speech. This process depends on the selected matching and modeling algorithms.
This process is represented in the figure below.

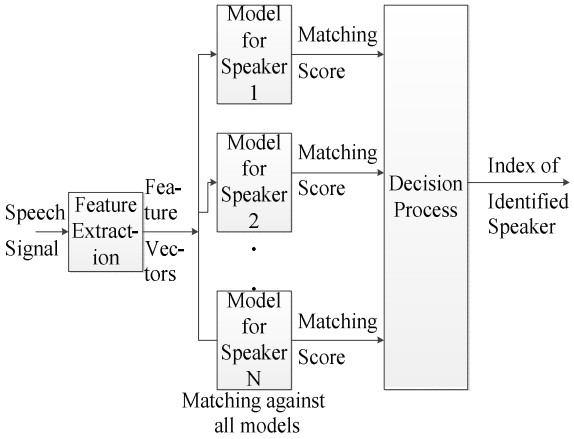

Fig. **5**: Decision Process of Speaker Identification.

In template matching, decision is based on the computed distances, whereas in stochastic matching it is based on the computed probabilities.
We will use Euclidean distance because it can be easily determined.[1] The formula used to calculate the Euclidean distance can be defined as following:
The Euclidean distance between two points
$P = (p_1, p_2, ..., p_n)$ and $Q = (q_1, q_2, ..., q_n)$ is,

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+.....+(p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad - (5)$$

If the distance of the feature vectors of authentic speaker to the feature vectors of that speaker which will be identified is small than the threshold then the speaker will be identified as authentic speaker otherwise identification rejected.

5. Speaker Database

The first step for identifying the speakers is to build a speaker-database, sub_database = { sub_1 , sub_2 , …. , sub_10}, in which we have collected 100 speech samples of 10 speakers. We made 10 round training sessions in which single speaker speaks 10 different utterances in two languages (US English and Hindi). These utterances are stored for extracting language independent features for speakers from their respective speech samples and is used in the testing phase of the speakers. By using these features of speech samples for each speaker, we can differentiate between two speakers and find similarity in the utterances of same speaker.

6. Results

In this section, we are showing the pitch of a male's speech waveform.

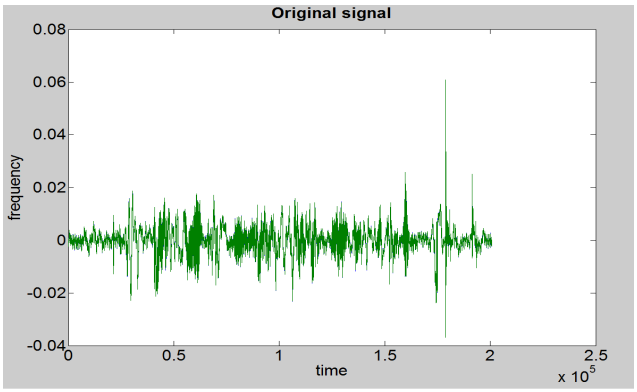

Fig. **6**: Original waveform of speech signal.

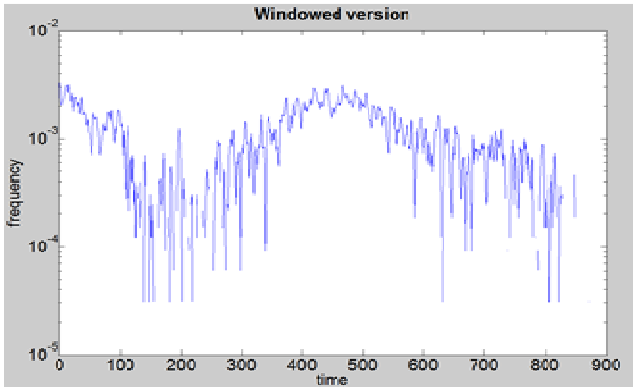

Fig. **7**: Windowed version of original signal

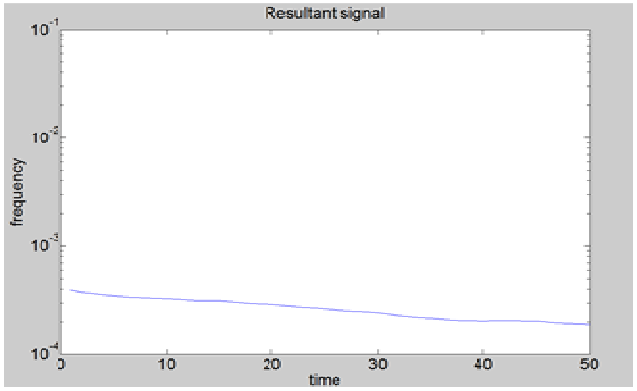

Fig. **8**: Pitch of resultant signal.

Since we can see that pitch of speech signal is almost constant so we can take it as a feature for human identification. But the constant nature of speech holds only for a short period of time. That is why, we are preferring MFCC as a reliable feature for human identification because MFCC is the cepstra of the frequency spectrum of speech on a mel-scale whose functioning is similar to the functioning of human ears.
Here we are presenting MFCC graphs of speech waveform.

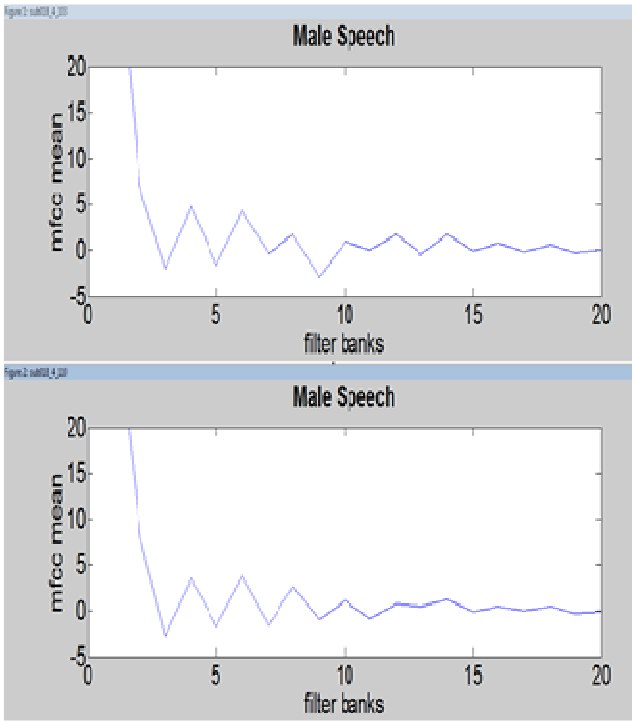

Fig. **9**: Different utterances speaking same male speaker.

Graph between mean error of MFCCs of same speaker and the number of mel-filter banks is given below.

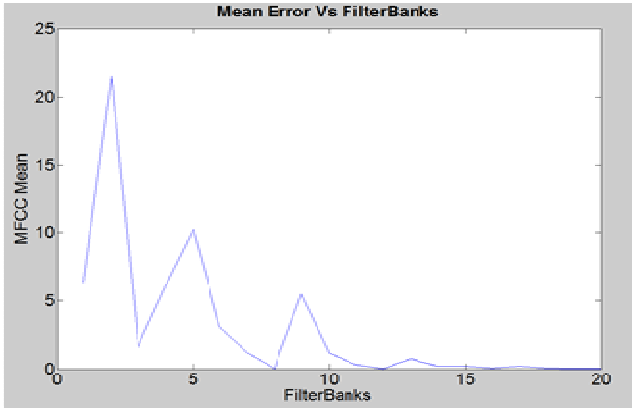

Fig. **10**: Error rate for the same speaker.

The graph between mean error and no. of mel-filter banks converges to the zero as number of mel-filter banks increase. We used a suitable number of mel-filter banks i.e. 20, to avoid computational complexity and ambiguity.

7. Conclusion

The goal of this paper was to minimize the error for a language-independent speaker identification system. The feature extraction is done using Mel Frequency Cepstral Coefficients {MFCC} and the speakers were modeled by using Gaussian Mixture Model. We find pitch and MFCC from the speech waveform.
Finally, we plot a graph between mean error and no. of mel-filter banks which describes that error becomes almost zero. Thus the system becomes able to identify the speaker by reducing the error between different utterances of the same speaker to zero level.

References

(1) Speaker Recognition: A Tutorial; Joseph P. Campbell, Jr.; *Proceedings of the IEEE*, Vol. 85, No. 9, September 1997.

(2) An Overview of Text-Independent Speaker Recognition: from Features to Super vectors; Tomi Kinnunen, Haizhou Li; *IEEE*/July, 2009.

(3) Speaker Recognition; Homayoon Beigi; Recognition Technologies, Inc.; U.S.A.; www.intechopen.com

(4) Automatic Speaker Recognition Using Gaussian Mixture Speaker Models; Douglas A. Reynolds; *The Lincoln Laboratory journal,* Volume B, 1995.

(5) Pitch in Speaker Recognition; Zhu Jian-wei, Sun Shui-fa, Liu Xiao-li, Lei Bang-jun; *IEEE*/2009.

(6) The MIT LL 2010 Speaker Recognition Evaluation System: Scalable Language-Independent Speaker Recognition; Douglas Sturim,William Campbell, Najim Dehak, Zahi Karam, AlanMcCree, Doug Reynolds, Fred Richardson, Pedro Torres-Carrasquillo, Stephen Shum; *IEEE*/2011.

(7) Speaker Recognition Using GMM; G. Suvarna Kumar, K.A.Prasad Raju, Dr.Mohan Rao CPVNJ, P.Satheesh; *IEST*/2010.

(8) Automatic Speaker Recognition Acoustics and Beyond; Douglas Reynolds; *MIT Lincoln Laboratory JHU CLSP*; 10 July 2002.

(9) Extended Average Magnitude Difference Function Based Pitch Detection; Ghulam Muhammad; *IAIT*/April, 2011.

(10) EM Algorithms of Gaussian Mixture Model and Hidden Markov Model; Guorong Xuan, Wei Zhang, Peiqi Chai; Department of Computer Science, Tongji University; Shanghai; *IEEE*/January,2001.