# Automatic Speaker Recognition using Fuzzy Vector Quantization

Suresh Kumar Chhetri, Subarna Shakya

*Department of Electronics and Computer Engineering, IOE, Central Campus, Pulchowk, Tribhuvan University, Nepal*

Corresponding Mail: ersuresh2012@gmail.com

**Abstract:** Speaker recognition (SR) is a dynamic biometric task. SR is a multidisplinary problem that encompasses many aspects of human speech, including speech recognition, language recognition, and speech accents. This technique makes it possible to use the speaker's voice to verify his/her identity and provide controlled access to services. The Mel-frequency extraction method is leading approach for speech feature extraction. In this thesis a new algorithm has been proposed which incorporates FVQ and DCT based MFCC feature extraction method. The proposed system will be improved the performance of SR through MFCC and FVQ method. The FVQ performance result will be compared with K means quantization in terms of EER.

**Keywords:** Speaker Recognition, speech feature extraction, Mel-frequency Cepstral Coefficients, K Means clustering, fuzzy C means clustering, Vector Quantization.

## 1. Introduction

SR is the identification of the person who is speaking by characteristics of their voices (voice biometrics), also called voice recognition. There is a difference between speaker recognition (recognizing who is speaking) and speech recognizing (recognizing what is being said). These two terms are frequently confused, and "voice recognition" can be used for both. In addition, there is a difference between the act of authentication and identification. Finally, there is a difference between speaker recognition (recognizing who is speaking) and speaker diarisation (recognizing when the same speaker is speaking). Recognizing the speaker can simplify the task of translating speech in systems that have been trained on specific person's voices or it can be used to authenticate or verify the identity of a speaker as part of a security process. [1]

Speaker recognition has a history dating back some four decades and uses the acoustic features of speech that have been found to differ between individuals. These acoustic patterns reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioral patterns (e.g., voice pitch, speaking style). Speaker verification has earned speaker recognition its classification as a "behavioral biometric.

## 2. System Modeling

ASR is the process used to identify or verify a person using speech features extracted from an utterance. A typical ASR system consists of a feature extractor followed by a robust speaker modeling technique for generalized representation of extracted features and a classification stage that verifies or identifies the feature vectors with linguistic classes. In the extraction stage of an ASR system, the input speech signal is converted into a series of low-dimensional vectors, the necessary temporal and spectral behavior of a short segment of the acoustical speech input is summarized by each vector.

Verification of an individual's identity is the key purpose of ASR. A subsequent outcome is the identification of commands or utterances that may be used to identify commands for an electro-mechanical or computing system to implement. The outcomes of ASR, recognition and device control, permits an individual to control access to services such as voice call dialing, banking by telephone, telephone shopping, telemedicine, database access services, information services, voice mail, security control for confidential information areas and many other activities. The benefit of ASR is to provide people with a mechanism to control electro-mechanical devices, machines and systems utilizing speech rather than through some mechanical action such as that achieved through the use of hand motions.

### *2.1 Speaker Identification*

Speaker identification is defined as the process of determining which speaker provides a given utterance. The speaker is registered into a database of speakers and utterances are added to the database that may be used at a later time during the speaker identification process. [4] The speaker identification process is shown in Figure 2.1. The steps shown in Figure 2.1 include feature extraction from the input speech, a measure of similarity from the available speaker utterances and a decision step that identifies the speaker identification based upon the closest match algorithm used in the previous step.

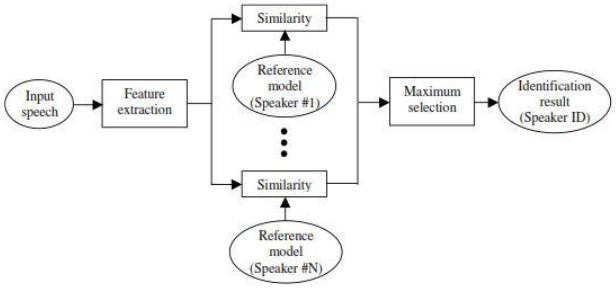


**Figure 2.1: Speaker identification.**

## *2.2 Speaker Verification*

The acceptance or rejection of an identity claimed by a speaker is known as Speaker Verification. [3] The speaker verification process is shown in Figure 2.2 and includes feature extraction from the source speech, comparison with speech utterances stored in the database from the speaker whose identity is now being claimed and a decision step that provides a positive or negative outcome.

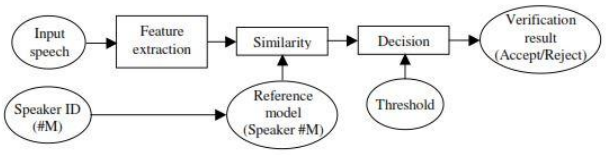


**Figure 2.2: Speaker Verification.**

## *2.3 System Overview*

The SR system used in the research is presented in Figure3.2. The training and testing steps are shown and how the classifier utilizes the trained data set. The training and testing steps include signal pre- processing to remove noise and clean up the signal prior to the next stage of training and testing. The next stage includes the classifier which involves the feature extraction and classification utilizing FVQ. The resulting vectors are models within the training steps and in the test steps the resulting vector is compared with the trained codebook to identify if there is a match or not.

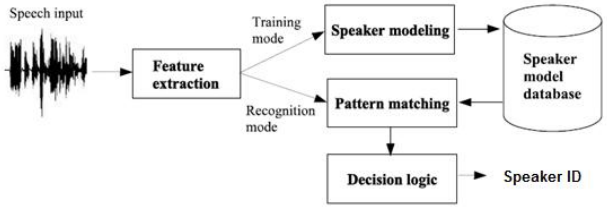


**Figure 2.3 (a) Schematic diagram of the closed-set speaker identification system.**

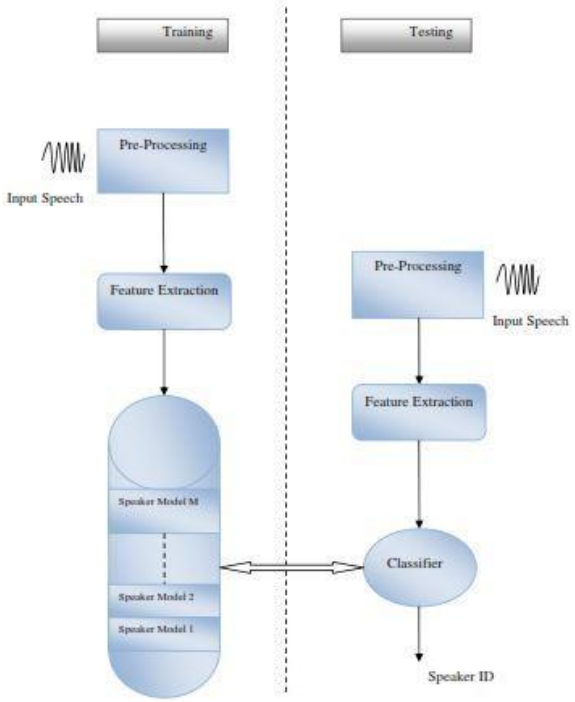


**Figure 2.3 (b) Speaker recognition system.**

## *2.4 Speech Parameterization Methods*

Parametric representation of speech waveforms is required (at a considerably lower information rate) for further analysis and processing as a step in the SR process. A wide range of parametric representation options exist that may be used to represent the speech signal parametrically for the speaker recognition process.

### *2.4.1 Mel-Frequency Cepstrum Coefficient Processor*

MFCC's are based on the Mel scale which is a heuristically derived perceptual scale. The Mel scale provides the relationship between perceived frequency or pitch, of a pure tone as a function of its acoustic frequency. In the Mel scale, to capture the phonetically important characteristics of speech of frequency *F* in Hz, a subjective pitch is measured in units known as *Mel*. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold; with a pitch of 1000 mels.Therefore the approximate formula shown in Equation 2.4.

$$\text{Fmel} = 2595 \log_{10}\left(1+\frac{F}{700}\right) \quad (2.4)$$

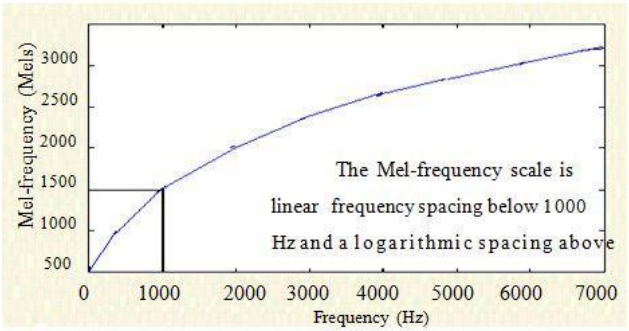


**Figure 2.4.1 (a) Frequency (linear) vs Mel frequency.**

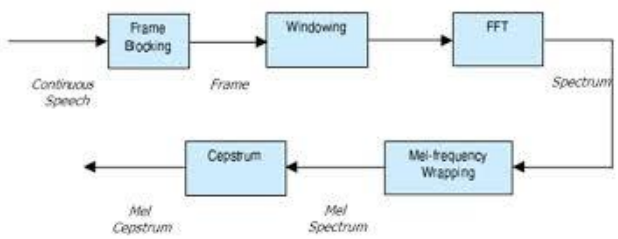


**Figure 2.4.1 (b) Block diagram of MFCC processor.**

***MFCC Algorithm:***

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the Mel frequencies.
4. Take the discrete cosine transform of the list of Mel log powers, as if it were a signal.

### *2.4.2 Pattern Recognition*

SR belongs to a much broader topic in scientific and engineering so called pattern recognition. The goal of pattern recognition is to classify the objects of interest into one of a number of categories or classes. The objects of interest are generically called patterns and in our case are sequences of acoustic vectors .The classification procedure in our case is applied on extracted features; it can be also referred to as feature matching.

### *2.4.3 Vector Quantization*

The VQ method is a classical signal processing technique which models the probability density functions by the prototype vector distributions. VQ was originally designed to be used as a data compression technique where a large set of points (vectors) in a multidimensional space could be replaced by a smaller set of representative points with distribution matching the distribution of the original data. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a *cluster* and can be represented by its center called a code word. The collection of all code words is called a codebook

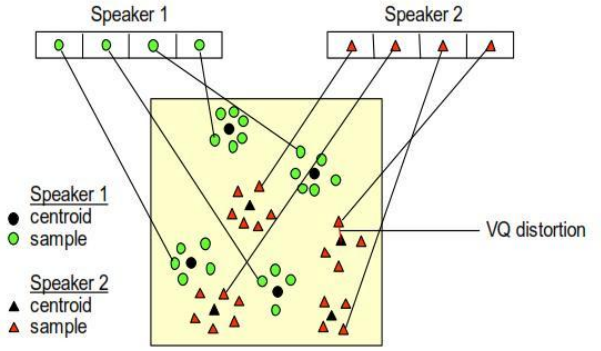


**Figure 2.4.3: Conceptual diagram illustrating vector quantization codebook information.**

### *2.4.4 Feature Matching*

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called pattern recognition. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching.

### *2.4.5 Clustering*

The objective of clustering is the classification of objects according to similarities among them, and organizing data into groups. Clustering techniques are among the unsupervised methods, they do not use prior class identifiers. The main potential of clustering is to detect the underlying structure in data, not only for classification and pattern recognition, but for model reduction and optimization.

#### *2.4.5.1 K MEANS CLUSTERING*

This is an algorithm to classify or to group data vectors based on attributes/features into K groups (or clusters). The K-means algorithm was developed for the VQ codebook generation. It represents each cluster by the mean of the cluster centroid vector. The grouping of data is done by minimizing the sum of squares of distances between the data vectors and the corresponding cluster's centroids.

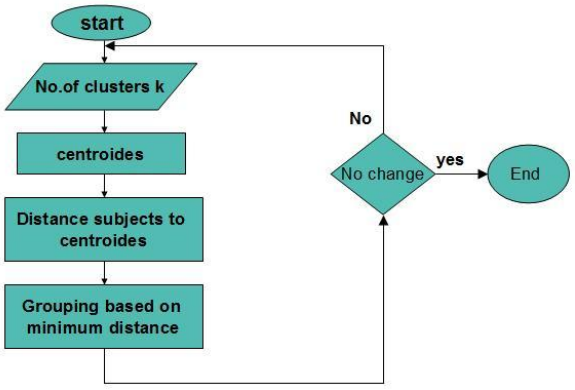


**Fig .2.4.5.1 Flow diagram of the K-means algorithm**

The K-means algorithm will do the three steps below until convergence:

Iterate until *stable* (= no object move group):
1. Determine the centroid coordinate.
2. Determine the distance of each object to the centroids.
3. Group the object based on minimum distance (find the closest centroid).

### *2.4.5.2 Fuzzy C means Clustering.*

Since clusters can formally be seen as subsets of the data set, one possible classification of clustering methods can be according to whether the subsets are fuzzy or crisp (hard). Hard clustering methods are based on classical set theory, and require that an object either does or does not belong to a cluster. Hard clustering of a data set *X* is the partitioning of the data into a specified number of mutually exclusive subsets of *X*. The number of subsets (clusters) is denoted by *c*. Fuzzy clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership.

The data set *X* is thus partitioned into *c* fuzzy subsets. In many real situations, fuzzy clustering is more natural than hard clustering, as objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial memberships. The discrete nature of hard partitioning also causes analytical and algorithmic intractability of algorithms based on analytic functional values, since these functional values are not differentiable.

## 3. Result and Discussion

### *3.1 Data Extraction and Preprocessing*

The first is data extraction that converts a wave data stored in audio wave format into a form that is suitable for further computer processing and analysis. The speech signal is a slowly timed varying signal (it is called quasi-stationary). An example of speech signal is shown in Figure 2.The pre-processing stage includes speech normalization, pre-emphasis filtering and removal of silence intervals. The dynamic range of the speech amplitude is mapped into the interval from -1 to +1. In this case, hamming window is used.

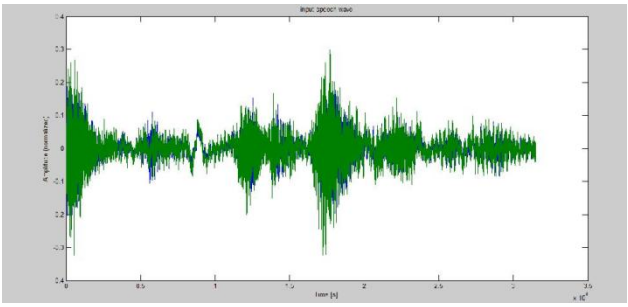

**Fig .3.1 Speech signal of Hello word.**

### *Framing*

In this step the continuous speech signal is blocked into frames of *N samples*, with adjacent frames being separated by *M* (*M* < *N*). The first frame consists of the first *N* samples. The second frame begins *M* samples after the first frame, and overlaps it by *N* - *M* samples and so on. This process continues until all the speech is accounted for within one or more frames. Typical values for *N* and *M* are *N* = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and *M* = 100.The new novel MFCC feature extraction method has new feature extraction algorithm forms part of the SR system presented in the research results. In this research, a DCT-II is used when computing the MFCC coefficients. The dynamic features were computed from the first and second order derivatives. This is a new and novel approach.

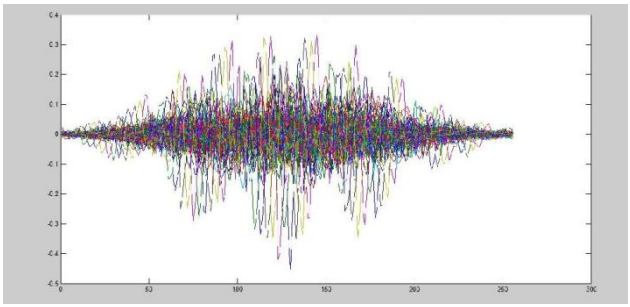

**Figure 3.2 Framing output.**

### *Hamming Window*

Hamming window is given by equation W (n) =0.54+0.46cos (2πn/N-1) where N is the length of the window.

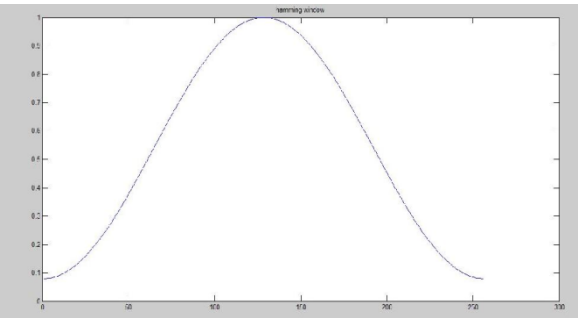

**Figure 3.3: hamming window**

### *Fast Fourier Transforms*

Fast Fourier Transform, converts a signal from the time domain into the frequency domain samples. The FFT is a fast algorithm to implement the Discrete Fourier transforms .

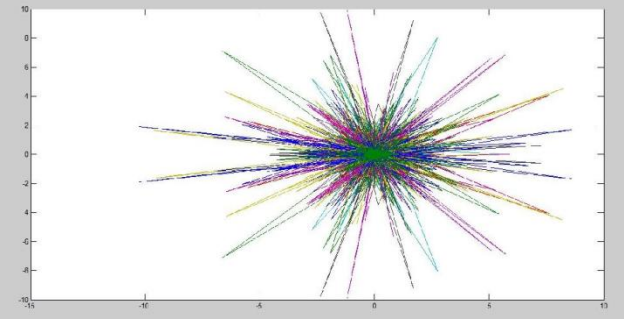

**Figure 3.4: Fast Fourier transforms**

### *Mel Scale Filter Banks*

- Mel-Frequency analysis of speech is based on human perception experiments**.**
- Human ears, for frequencies lower than 1 kHz, hears tones with a linear scale instead of logarithmic scale for the frequencies higher than 1 kHz.

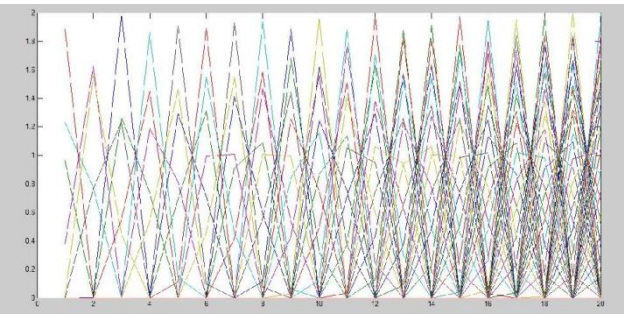

**Figure 3.5: Mel scale filter banks.**

### *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). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT).

### *Vector Quantization*

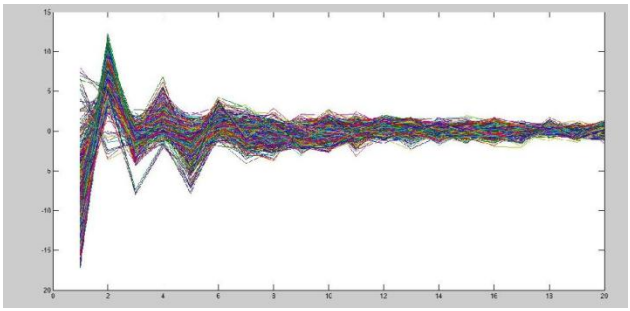

**Figure 3.7 (a) the vectors generated from training before VQ**

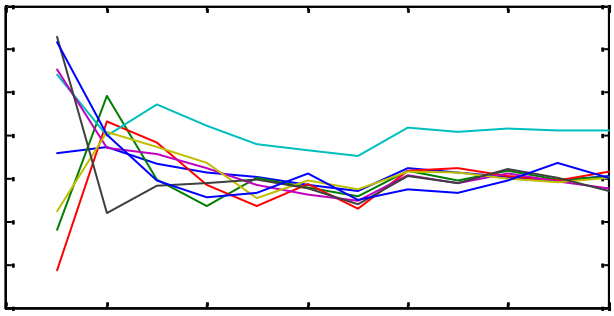

**Figure 3.7 (b) the representative feature vectors resulted After VQ**

### *K Means Clustering Technique*

The K-means algorithm partitions the T feature vectors into M centroids. The algorithm first randomly chooses M cluster-centroids among the T feature vectors. Then each feature vector is assigned to the nearest centroid, and the new centroids are calculated for the new clusters. This procedure is continued until a stopping criterion is met, that is the mean square error between the feature vectors and the cluster-centroids is below a certain threshold or there is no more c In other words, the objective of the K-means is to minimize total intra-cluster variance, V.

$$V = \sum_{i=1}^{k} \sum_{j \in S_i} |x_j - \mu_i|^2 \qquad (3.8)$$

Where there are *k* clusters *Si*, *i* = 1, 2... *k* and μ*i* is the centroid.

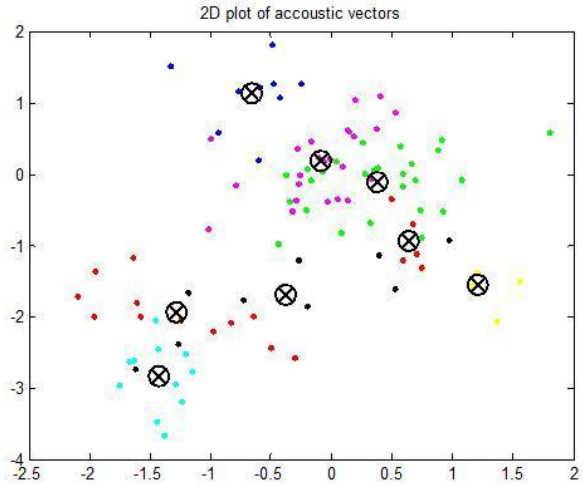


**Figure 3.8(a): Clusters in the K-means algorithm**

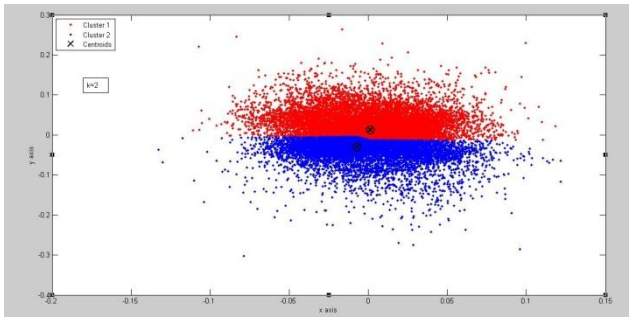

**Figure 3.8(b): Partitioning**

### *Fuzzy C Means Clustering*

It allows one piece of data to belong to two or more clusters. It is based on minimization of the following objective function:

$$J_{m=\sum_{i-1}^{N}\sum_{j-1}^{C} u_{ij}^{m}\left\|x_i-c_j\right\|^2} \qquad (3.9)$$

where m is any real number greater than 1, uij is the degree of membership of xi in the cluster j, xi is the ith of d-dimensional measured data, cj is the d-dimension center of the cluster, and || || is anynorm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership uij and the cluster centers cj.

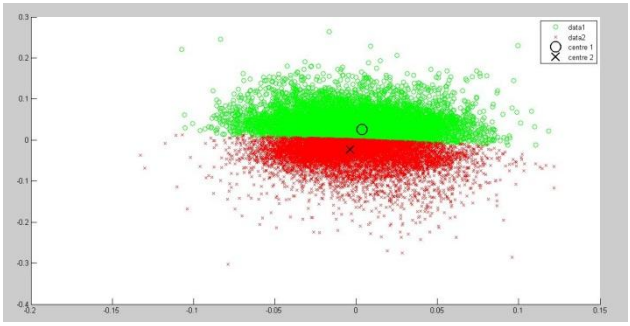

**Fig 3.9 Fuzzy Partitioning.**

### *Equal Error Rate*

The Equal Error Rate (EER) decision point is defined as the point where the false rejection and the false acceptance error probabilities are equal. However, in practice it is implemented as the point where the distance between the false rejection and the false acceptance errors is minimal. *False rejection* (FR) error occurs when the true target speaker is falsely rejected as being an impostor, and as a result, the system *misses* recognizing an attempt belonging to the true authorized user. A *false acceptance* (FA) Error occurs when a tryout from an impostor is accepted as if it came from the true authorized user.

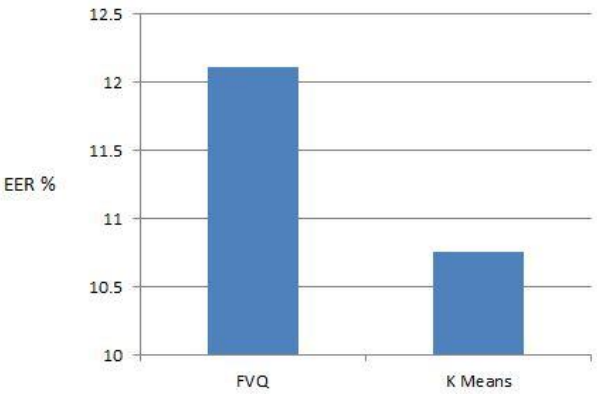


**Figure 3.10: Performance of the system.**

### *Comparison*

There is no clustering algorithm that can be universally used to solve all problems. The K-Means and Fuzzy C-Means are very well known clustering algorithm in the partition based algorithms and they have been used in many applications areas. It is very well known that each of the two algorithms has pros and cons. On because of its simplicity, the K-Means runs faster, but vulnerable to noises. Fuzzy C-Means is a little bit more complex and hence runs slower but stronger to noises. The computational time of K-Means algorithm is less than the FCM algorithm. Further, K-Means algorithm stamps its superiority in terms of its lesser execution time. Also ,the distribution of data points by K-Means

algorithm is even to all the data centres,but,it is not even by the FCM algorithm. This means that the data points are evenly distributed by K-Means algorithm. But, the FCM algorithm has more variations in the distribution.

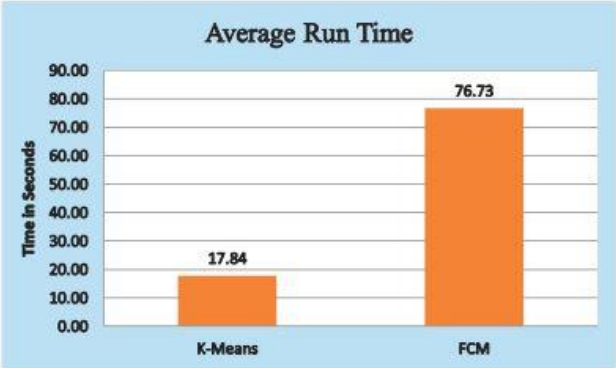


**Figure 4(a): Result Comparison**

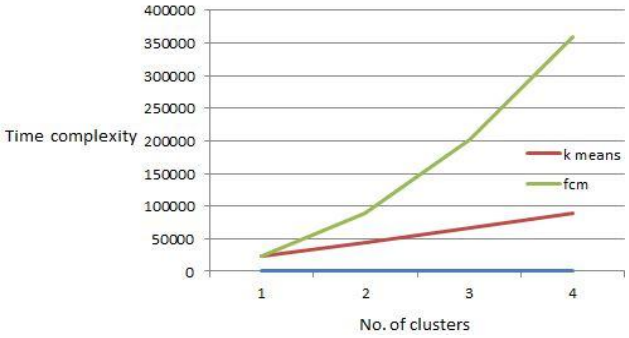


**Figure 4(b): Time Complexity**

## 4. Conclusion

The performance of proposed speaker verification test is better than any other system. The performance of the MFCC feature extraction method was improved by using DCT.DCT improved performance in terms of identification accuracy .The performance of the feature extraction methods and classifiers were measured based on Equal Error Rate (EER) values. In this thesis, fuzzy c-means clustering and k means clustering algorithm are used. The working principle of FVQ is different from K- means VQ, in the sense that the soft decision making process is used while designing the codebooks in FVQ, whereas in K-means VQ hard decision process is used. Moreover, in K-means VQ each feature vector has an association with only one of the clusters. Whereas in FVQ, each feature vector has an association with all the clusters with certain degrees of association dictated by the membership function. Since all the feature vectors are associated with all the clusters, there are relatively more number of feature vectors for each cluster and hence the representative vectors i.e., code vectors may be more reliable than the other VQ technique. Therefore, clustering is better in FVQ which lead to better performance compared to k means VQ. In this thesis, a new approach for speech feature extraction utilizing FVQ has been presented that improves speaker recognition performance.

## References


[1] Campbell Jr, Joseph P. "Speaker recognition: A tutorial." *Proceedings of the IEEE* 85.9 (1997): 1437-1462

[2] Kumar, Srinivasan, and P. Mallikarjuna Rao. "Design of An Automatic Speaker Recognition System Using MFCC, Vector Quantization And LBG Algorithm. "International Journal on Computer Science & Engineering 3.8 (2011).

[3] Karayiannis, Nikolaos B., and Pin-I. Pai. "Fuzzy vector quantization algorithms and their application in image compression." Image Processing, IEEE Transactions on 4.9 (1995):1193-1201.

[4] 4. Sheeraz Memon, Margaret Lech and Namunu Maddage, " Speaker verification based on different vector quantization techniques with gaussian mixture models," Third International Conference o n Network and System Security, 2009, pp. 403 – 408.

[5] ATAL, B. S. 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of Acoustical Society of America* Vol.55.No.6**,** 1304-1312.

[6] Hossan, Md Afzal, Sheeraz Memon, and Mark A. Gregory. "A novel approach for MFCC features extraction." Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on. IEEE, 2010.

[7] 7. Wang, Haipeng, et al. "A novel fuzzy-based automatic speaker clustering algorithm." Advances in Neural Networks–ISNN 2009. Springer Berlin Heidelberg, 2009. 639-646.

[8] 8. A. Hossan and M .A.Gregory, "Distributed DCT based dynamic feature for automatic speaker recognition(in press)," *IEEE International Symposium on Signal Processing and Information Technology*, Egypt, 2010