Speaker Recognition and Performance Comparison based on Machine Learning

Rajeev Ranjan
rajeevranjan1134@gmail.com
Department of Electronics & Communication Engineering,
Chandigarh University, Mohali, Punjab

Abstract- The speaker recognition is one of the most emerging field in the area of digital signal processing. In this paper, analyzed the performance comparison of voice recognition using Mel frequency Cepstrum coefficient (MFCC) with two methods based on vector quantization (VQ) techniques by Linde buzo gray (LBG) algorithm and improved weighted VQ method. In the first phase, extract the minor amount of data points from the voice signal that is used subsequent to represent every speaker is known as MFCC and in the second phase, for feature matching there are two approaches which have been proposed here based on VQ techniques for recognition purpose and also comparison of both algorithms are being done with different time length speech samples to get better recognition rate and improving in efficiency of the system. The VQ is used as feature matching with traditional LBG Algorithm and improved weighted VQ algorithm for better recognition. The weight of vector in VQ algorithm to get the weighted distortion and after compare between traditional and improved weighted VQ algorithm. When the test time '1s' for traditional VQ algorithm 81.25% whereas improved weighted VQ 87.5 % and if test time '2s' for traditional VQ algorithm 87 % whereas, improved weighted VQ 93.7 %. That is improved weighted VQ algorithm gives the better speaker recognition than VQ with LBG algorithm.

Keywords: MFCC, Weighted Vector Quantization, Vector Quantization, LBG algorithm.

1.Introduction

The speaker recognition is the operation of identifying a speaker who is speaking on the base of entity information through audio waves. In this speaker recognition making a certain algorithms which is designed for machine having knowledge of speech signal to authenticate the unknown speaker through the database. Speech communication is process of sending and receiving information or messages among individuals. Speech signal waveform has phonetic data, vocal characteristics of speaker [1, 2]. The different types of classification of speaker recognition systems are classified as-

I. Open set VS closed set

There are two type of system which is classified into two categories (a) Open set framework can be keep several numbers of speakers constantly more than one. (b) A closed set framework has just a limited number of speakers enrolled to the framework.

II. Identification VS verification

It is performed in two parts: Identification and Verification [4]. The verification and identification of speaker are generally considered as the most economical and natural methods to avoid unapproved access of computer systems or physical locations. Identification of speaker is the procedure of recognizing which enlisted speaker gives a given expression by comparing his essentiality with each instruction stored in database. Verification of speaker is the procedure of acceptance or rejection of a speaker, the identity claim by particular person's stored database.

III. Text-dependent VS text-independent

It is based on the description of text utterance by the speaker concerning the process of identification. For text-dependent type, the test utterance is same as the text apply in training stage and test speaker has earlier perception of the system. For text-independent type, the test utterance is not same as the text apply in training stage and does not have earlier perception about the information of the training stage and Speaker can speak everything. Here text independent speaker recognition method is used [5]. The speaker recognition system is basically divided into the two parts also shown by block diagram in Fig. 2 as Speech feature extraction and Feature matching algorithm [3].

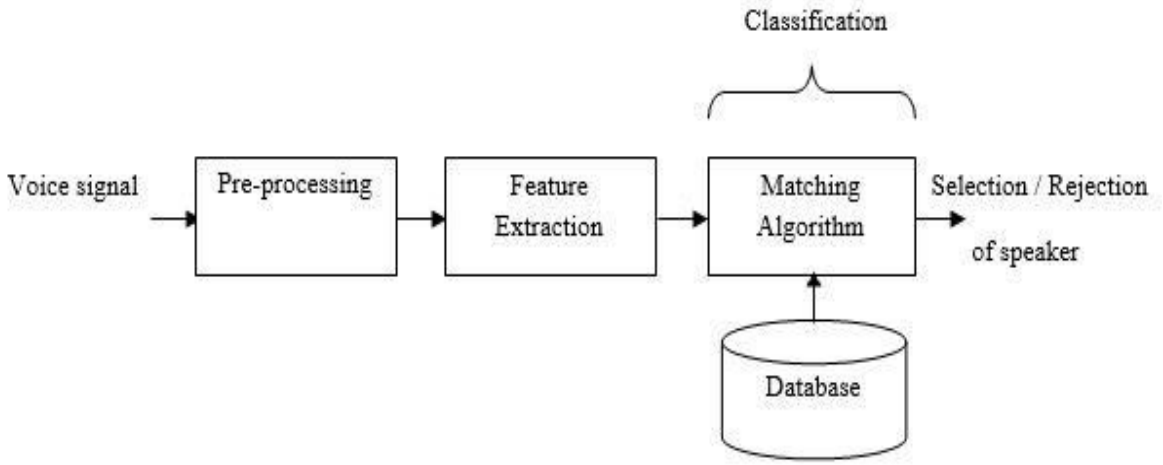

Fig. 1: Block diagram of speaker recognition system

In the first part of this recognition system finding with MFCC for speech signal. Here used MFCC because it is more consistent with human hearing. In second part, classification model or matching algorithm which is used for recognition system. So used VQ as feature matching with traditional LBG Algorithm and improved weighted VQ algorithm for better recognition. The weight of vector in VQ algorithm to get the weighted distortion and after compare between traditional and improved weighted VQ algorithm. The improved weighted VQ algorithm gives the better speaker recognition than VQ with LBG algorithm.

The outline of this paper is arranged as follow: Section-2 gives feature extraction techniques and section-3 gives classification using VQ algorithm and weighted VQ algorithm methods. Section-4 describes the simulation results and discussion, finally section 5 is the conclusion.

2.Feature Extraction

A wide range of techniques which are used for the feature extraction. It can be analyzed into different classification. One technique is using with filter bank coefficient and other is predictive coefficient by all pole model. Here the MFCC is used for feature extraction [12,13]. The MFCC block diagram is shown in Fig.2. Firstly we take a voice signal and sampled at 16 kHz. Voice signal is proceed through pre-emphasis which is used as noise reduction system to improve the amplitude of input data which signal to noise ratio is low and boost the signal spectrum approximately 20 dB/decade and the spectrum of speaker sounds that

have a sharp roll-off in high frequency zone. To perform pre-emphasis filter, we choose value of α is 0.95 and represented by equation below:

$$H(z) = 1 - \alpha z^{-1} \quad (1)$$

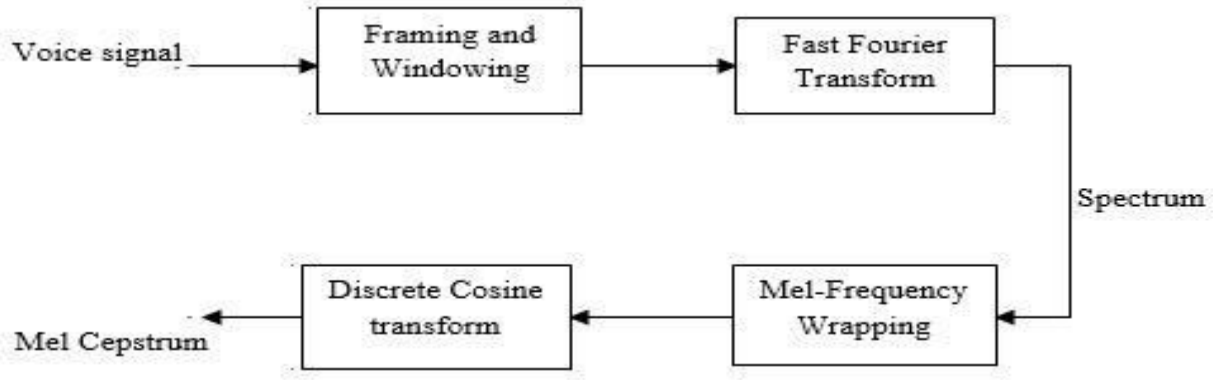

Fig. 2: Block diagram of MFCC

After that, the voice signal has been break into short frames. Here we take N samples in each frame with beside frames being partition by M (M< N). Then windowing every individual frame for minimal signal discontinuities. Typically Hamming window is used. Then the Fast Fourier transform (FFT), they converts each frame of N samples time domain to frequency domain. It is fast algorithm to implement the Discrete Fourier Transform (DFT).

$$P(k) = \sum_{n=0}^{N-1} p(n) e^{-\left(\frac{j2\pi kn}{N}\right)} \quad ; \qquad k = 0,1,2 \ldots N-1 \quad (2)$$

After that to find periodgram based power spectral for each frame i.e. called power spectrum. Then the absolute of FT and its square. Compute the Mel-spaced Filter bank. This is a set of triangular filters banks that are apply for the periodgram power estimate of previous step. The filter bank taken as vector form and every vector is almost zeros, but certain section of spectrum is non-zero. Mel scale filter banks helps to windows equally and implementation easier. The formula for relation between hertz scale and Mel scale given by equation as:

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{100}\right) \quad (3)$$

Then calculate the energies of filter bank here each filter bank is multiply with power spectrum and the coefficient is add. After that take log of energies from above previous step. They give us with log filter bank energies. Then discrete cosine transform (DCT) of log filter bank energies to give cepstral coefficients [5-9].

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \; ; \; N-1 \leq k \leq 0 \quad (4)$$

.The resulting features 12 coefficients for each frame are called as MFCC.

3. Feature Classification

The Fig.3 show the VQ based recognition system and Fig.4 is VQ models of two speakers. In this section, we discussed on classification model or matching algorithm which is used for recognition purpose system. After analyzing the theory of VQ model with LBG and improved weighted VQ method algorithm and try to implementing with mathematical analysis. In VQ show there is an extensive set of feature vectors are being isolated into bunches having around a similar number of points nearest to them. Here each group of section is represented by its centroid. VQ is characterized as a mapping capacity that maps into k-dimensional vector space to a limited set $C_B = \{C_1, C_2, C_3 \ldots C_N\}$. The codebook is a finite set C_B. They consists of N numbers of code vectors and every code vector $C_i = \{c_{i1}, c_{i2}, c_{i3} \ldots c_{ik}\}$.Each code vector is dimension of k. The LBG algorithm is commonly used for codebook generation. Feature vectors are extricated from input voice signal and the Euclidean separation between input voice signal and each code vector is measured. The input vector has belongs to the accumulation of the code vector that makes the shortest distance [7,8].

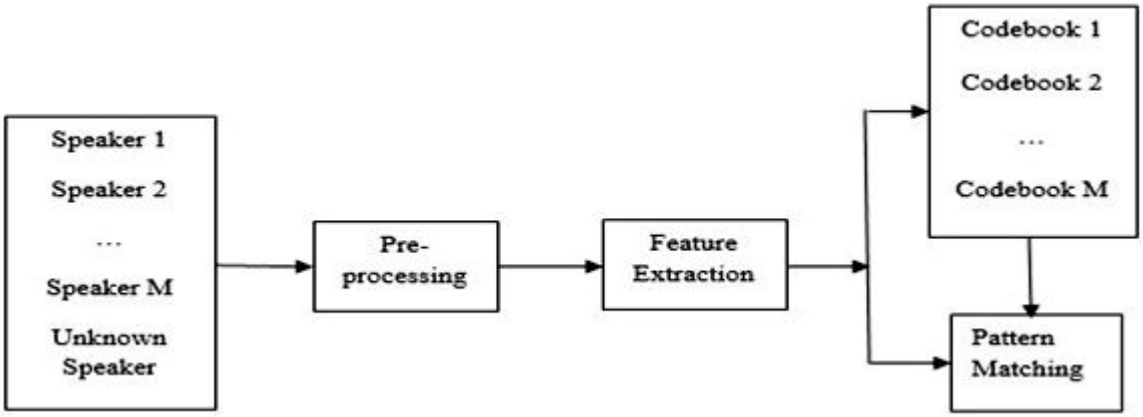

Fig. 3: VQ based speaker recognition system

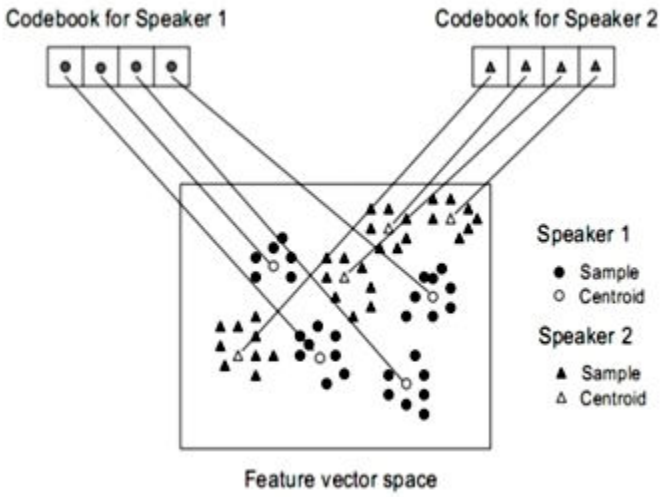

Fig. 4: Vector Quantization Model for two Speakers

3.1. Linde-Buzo-Gray Algorithm

The Fig. 5 show the flow cart of LBG algorithms. In this section for grouping an arrangement of L training vectors into an arrangement of M codebook vectors. The calculation is formally actualized by the accompanying recursive methodology as-first outline a 1-vector codebook that are the centroid of the whole arrangement of preparing vectors. Twofold the extent the codebook by part every current codebook Yn indicated as -

$$Y_{+n} = Yn\ (1+\varepsilon) \quad (5)$$

$$Y_{-n} = Yn\ (1-\varepsilon) \quad (6)$$

where, n lies in between 1 to the present size of the codebook and splitting parameter is ε (ε=0.01). Closest neighbor search for each training vector, discover the code word in the current codebook and flow that is nearest in terms of likeness estimation, and assign that vector related with the nearest code word. Centroid refresh the code word in every cell. Iteration 1: repeat stages 3 and 4 until the point that the normal separation less than preset limit. Iteration 2: repeat stages 2, 3 and 4 until a codebook size of M is planned [8,11].

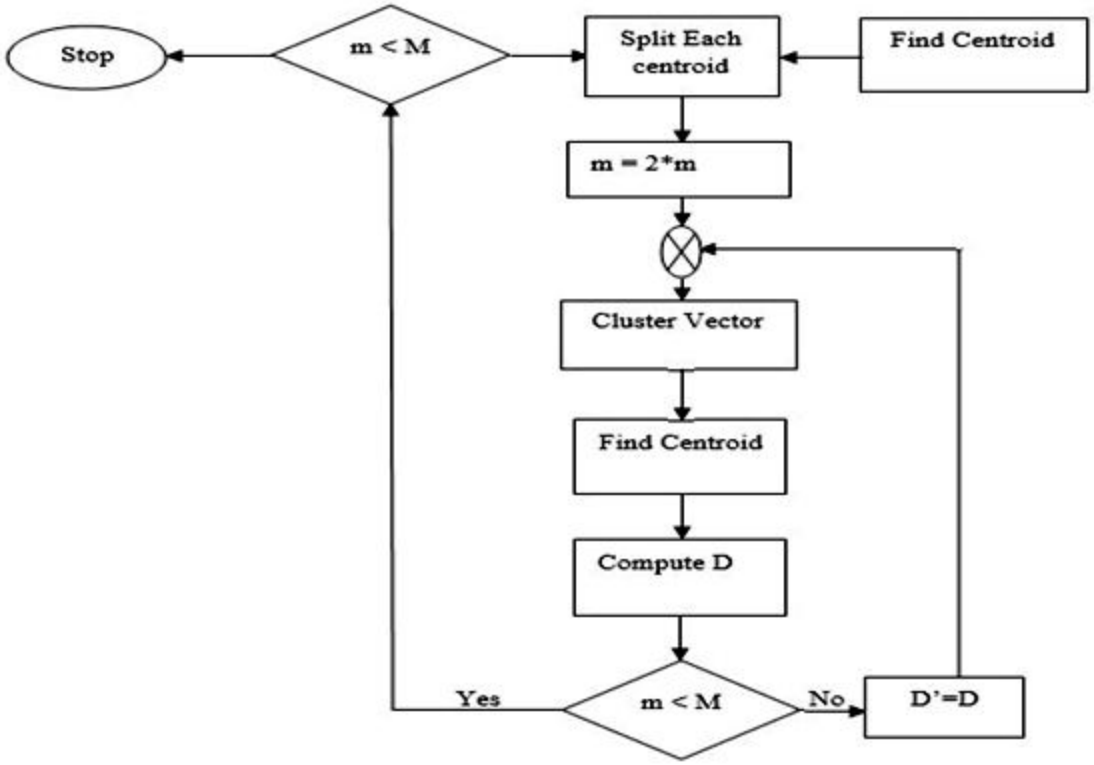

Fig. 5: Flow chart of LBG Algorithm

In the pattern matching scenario, VQ is utilized as a part of assessing distortion D among feature vectors $Z = \{z1, z2, z3, \ldots \ldots zK\}$ of an unknown speakers and all codebooks $\{C_1\ , C_2\ \ldots C_N\}$ represented in database. By utilizing straightforward choice standards needs to choose the speaker S that have minimum distortion

$$S=\arg\min D\,[Z, C(i)]; \quad \text{where } i=1, 2\ldots N. \tag{7}$$

Average distortion is defined as the equation which is given below:

$$D(Z,C) = \frac{1}{K}\sum_{z\in Z}^{K} d\left(z, c_{NN[z]}\right), \tag{8}$$

where, NN[z] = closest code vector index for z.

3.2. Improved weighted vector quantization algorithm

3.2.1. Assigning the weights

For allotting the weights of code vector it must be rely on upon the minimum separations from different code vectors. So the weight is allotted as follows:

$$w = \frac{1}{\sum_{k\neq x} 1/d(c,c_{NN}^{(k)})} \tag{9}$$

The above equation in which the reciprocal of the total amount of inverse distances is evaluated using closest code vector of all classes.

3.2.2.Weighted distortion measure:

Weighted distortion measure is defined by equation below:

$$D(Z,C) = \frac{1}{K}\sum_{z\in Z}^{K} f\left(w_{NN[z]}\right) d\left(z, c_{NN[z]}\right) \tag{10}$$

where, $w_{NN[z]}$ the weight connected with the closest code vector to the x in codebook and f is non-increasing argument function. If code vectors have better discrimination or heavy weight move forward to reduce the distance and if have non-discrimination or small weight code vectors move forward to enhance the distance. This above function result is found to be viewed like an administrator which pulls in vectors z, very near to codebook so it also can say that the weight of code vectors w is large so it decreases over all distortion and it can be seen as an operator which repels those vectors z, very large distance and it is quantized with low weight which increases over all distortion [10].

4.Results and Discussion

In this experimental setup implementation for speaker recognition purpose in this model, firstly recorded 16 speech signal through audacity software via windows. For this take sampling rate at 16 kHz. After this each speech signal has been blocked into short frames duration. We take 16ms interval short frame duration with respect to N=256 samples in each frame and adjacent frames is approximate 6.5ms duration with respect to M=100. After this windowing process has been done for each frame then evaluated FFT and absolute value of complex Fourier transform to get periodgram or power spectrum and taken set of 20

triangular filters Mel-spaced filter bank that apply for periodgram power estimate. Take the log of each of 20 energies from filter bank. This gives us with 20 log filter bank energies. After this taking the DCT of the 20 log filter bank energies to give 20 Cepstrum coefficients. For speaker recognition only lower 12-13 coefficients of 20 coefficients are kept. After that create database in Matlab 2013a containing each speech signal within train folder and testing folder also and then loaded each signal with evaluating MFCC and matching with vector quantization Model using LBG and weighted distortion method algorithms one by one from databases. Also evaluating Euclidean distance or average distortion with each signal from all other signal which is present in database for LBG and also weighted distortion measure with each signal from all other signal and finding with minimum chosen value between each code vector and input voice signal is measured. The simulation of speech signal and its MFCC VQ before and after VQ is shown by Fig.5 and Fig. 6 is given below:

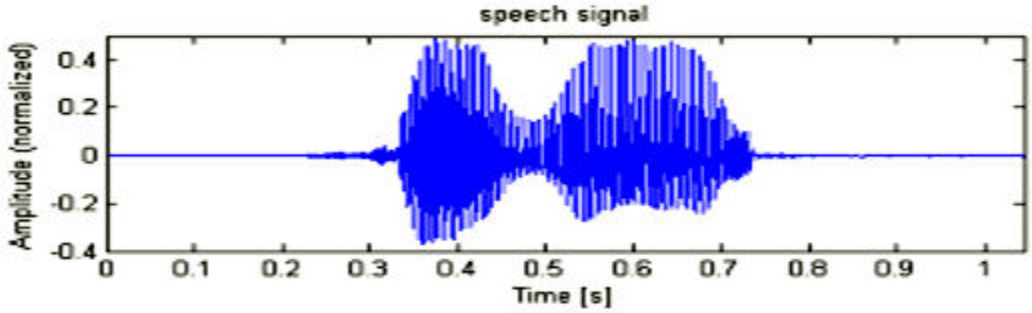

Fig. 6: Input voice signal

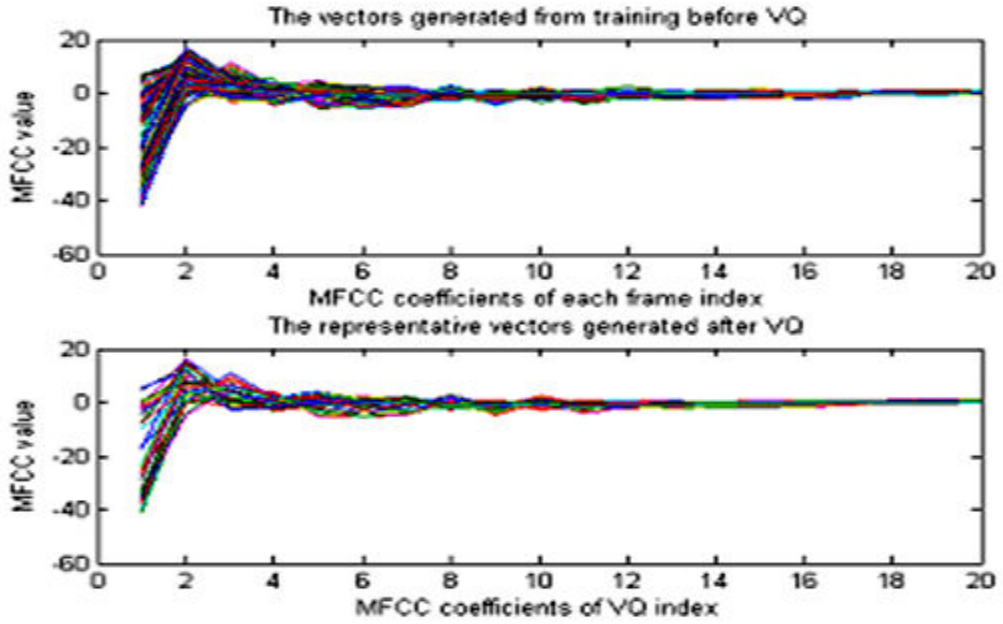

Fig. 7: MFCC vector representation before VQ and after VQ

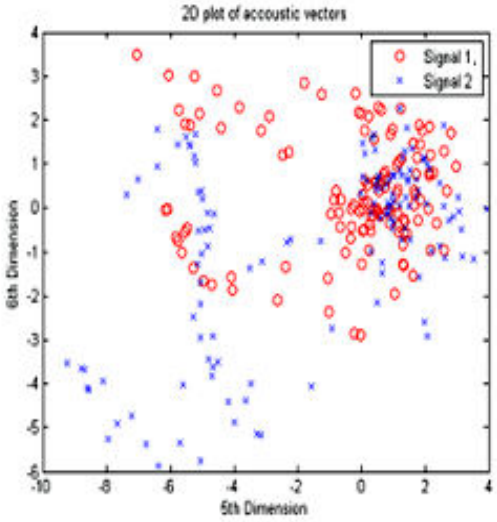

Fig. 8: 2D plot of acoustic vectors of signal 1 and 2

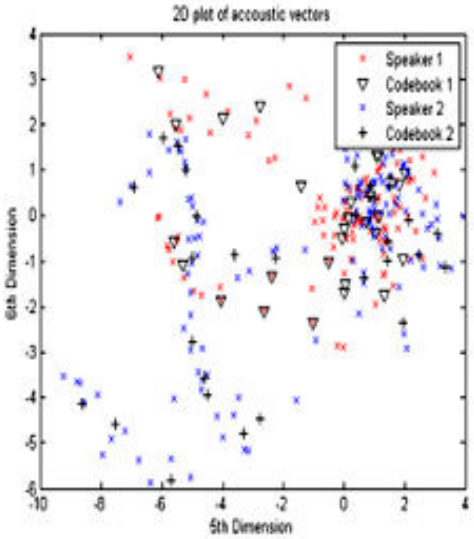

Fig. 9: VQ code book generation for speaker and 2

If we take different time length of speech sample of speaker like 1s, 2s…so on then the recognition rate will be different for different speech sample. Experimental results shows the improving recognition rate which is given by the following table 1.

Table 1: Recognition rate of traditional VQ and Improved weighted VQ algorithm for fixed codebook size (32)

Test time	Traditional VQ algorithm (%)	Improved weighted VQ algorithm (%)
1s	81.25	87.5
2s	87	93.7

After seeing result for using above two methods traditional VQ with LBG and improved weight VQ algorithm , for the fixed codebook size K (32) , recognition rate is found to be 87.5 % in improved weighted algorithm as compared to traditional VQ with LBG which is 81.25% so we find better recognition rate in improved weighted VQ algorithm. If we take different code book size from K=8 to 128 then we get different recognition rate for different test time speech samples for the traditional VQ with LBG and the improved weighted VQ algorithm. Below corresponding valued chart which can be plotted in the figure 9 and 10 below as follows:

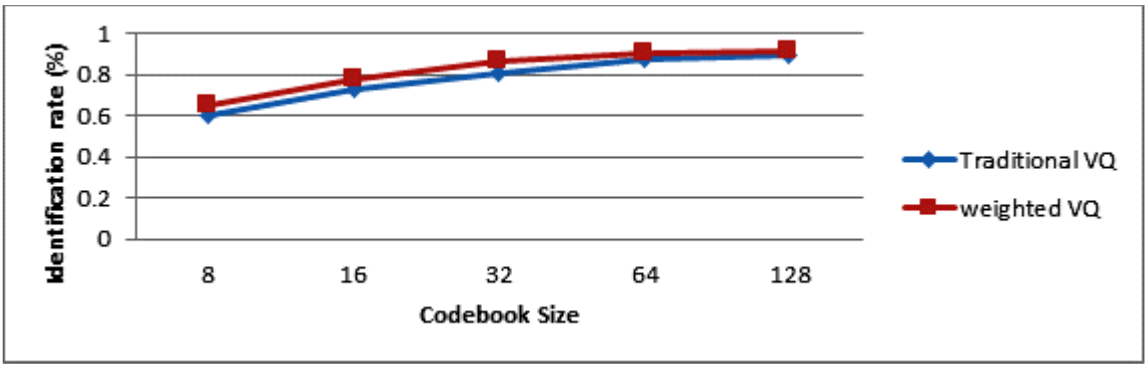

Fig. 10: Performance evaluation chart for 1s speech sample

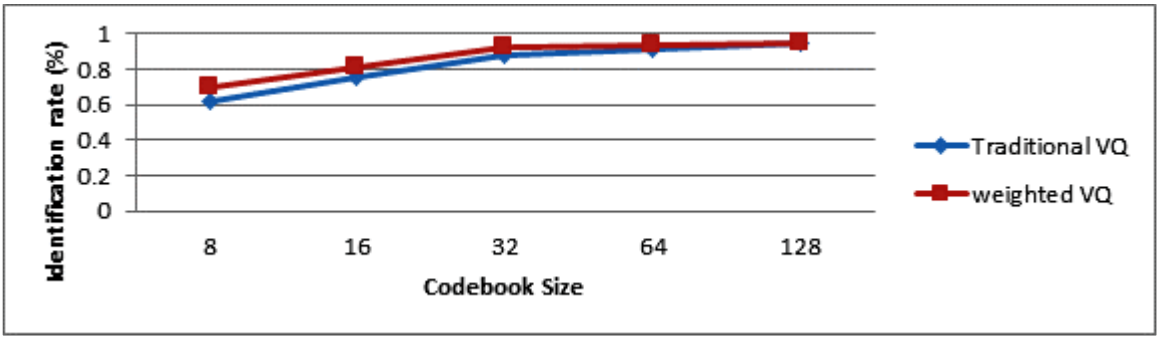

Fig. 11: Performance evaluation chart for 2 sec speech sample

5. Conclusion

As seen in above corresponding valued chart for plot Fig. 8 and 9, we find the different identification rate versus code book size with respect to speech sample 1s and 2s. In our conclusion, we find that as we increase code cook size from K=8 to 128 then the recognition rate is increased with corresponding to the values for traditional VQ with LBG algorithm and Improved weighted VQ also. We find that recognition rate is achieved better in improved weighted VQ method as compared to the traditional VQ with LBG. In this recognition process utilizing the technique for weight classes gives us a utilizing the technique distinction between the categories. The simulation results of this improved algorithm gives higher recognition rate than the other algorithms. Generally, it shows in result that using very short test samples with appropriate outline speaker recognition system of VQ can acquire better recognition rate model having lower sophisticated problems, it is more robust method than reference method. VQ is also preferred for real time systems because less memory required and robust in nature. In this work, the study also reveals that as increase the size of codebook, so the efficiency of the system is increased and also enhances the recognition rate of the system.

References

[1] Rabiner, L.R. and Schafer, R.W.: Introduction to speech Processing. Foundations and Trends in Signal Processing, vol. 1, no. 1-2, Ch. 2-4, (2007) 17-44.

[2] Campbell, J.P.: Speaker recognition: A Tutorial, Proceedings of the IEEE, vol. 85, issue -9, (1997) 1437-1462.

[3] Kaur, K. and Jain, N.: Feature extraction and classification for automatic speaker recognition system-mA review, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, issue-1, (2015) 1-6.

[4] Doddington, G.R.: Speaker recognition – Identifying people by their voices, Proceedings of the IEEE, vol. 73, issue-11, (1985) 1651-1664.

[5] Singh, A.K., Singh R. and Dwivedi, A.: Mel frequency cepstral coefficients based text independent automatic speaker recognition using matlab, International Conference on Reliability, Optimization and Information Technology (ICROIT), (2014) 524-527.

[6] Farah, S and Shamim, A.: Speaker recognition system using mel-frequency cepstrum coefficients, Linear prediction coding and vector quantization, 3rd International Conference on computer, Control & Communication (IC4), (2013) 1-5.

[7] Mishra, P. and Agrawal, S.: Recognition of speaker using mel-frequency cepstrum coefficient and vector quantization, International Journal of Science, Engineering and Technology Research, vol. 1, issue-6, (2012) 12-17.

[8] Singh, S. and Rajan, E.G.: Vector quantization approach for speaker recognition using mel- frequency cepstrum coefficient and Inverted MFCC, International Journal of Computer Applications, vol. 17, issue-1, (2011) 1-7.

[9] Luo, X.T., Ji, L.X. and Li, S.M.: Weighted distortion measure on standard deviation for VQ based speaker identification, 2nd International Conference on E- business and information system security, (2010) 1-4.

[10] Soong, F.K., Rosenberg, A.E., Rabiner, L.R. and Juang, B.H.: A Vector quantization approach to speaker recognition, in Proc. IEEE International Conference on Acoustic and Speech Signal Processing, vol. 10, (1985) 387-90.

[11] Kumar, C.R. and Rao, P.M.: Design of an automatic speaker recognition system using MFCC, Vector quantization and LBG algorithm, International Journal on Computer Science and Engineering, vol. 3, issue-8, (2011) 2942-2954.

[12] R. Ranjan and A. Thakur, "Analysis of feature extraction techniques for speech recognition system," International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, vol. 8, Issue-7C2, May 2019.

[13] R. Ranjan and R. K. Dubey, "Isolated word recognition using HMM for Maithili dialect," IEEE International conference on signal processing and communication, pp. 324-328, 2016.