Mel frequency Cepstrum Coefficients and Enhanced LBG algorithm for Speaker Recognition

Hussein Lafta Attiya, Ali Yakoob Yousif

Babylon University College of Science - Computer Dept.
wsci.husein.attia@uobabylon.edu.iq, wsci.ali.yakoob@uobabylon.edu.iq

Abstract: In this paper, an improved strategy for automated text dependent speaker recognition system has been proposed in noisy environment. The preprocessing of speaker signal started with eliminate the background noise. The next step is signal filtering and features extraction using cepstrum coefficients method, this extracted features can be used to by the enhanced LBG for vector quantization algorithm for speaker recognition, such that the specified speaker can be determined by matching the speaker to be tested with in stored codebook in database. And finally select correct speaker that have the lesser Euclidean distance. The speech feature extraction was based on a dataset of 175 different samples collected from 25 different speakers The results of the proposed system approved with good recognition ratio of speaker identification with maximum accuracy about 96.2% for database with close set of selected words contains the most used phonemes. Also the results of experiments show that recognition accuracy increased with frames overlapping.

[Hussein Lafta Attiya, Ali Yakoob Yousif. **Mel frequency Cepstrum Coefficients and Enhanced LBG algorithm for Speaker Recognition.** *Researcher* 2015;7(1):19-25]. (ISSN: 1553-9865). http://www.sciencepub.net/researcher. 4

Key words: Speaker recognition, Speech Signal Pre-processing, Mel frequency cepstrum Algorithm. Vector quantization

1. Introduction

During the previous decades, a large number of speech processing techniques have been proposed and implemented, and a number of significant advances have been witted in this field during the last decades. Speaker recognition means the ability of program to recognize or identify spoken words and carry out voice. The spoken words are digitized into sequence of numbers, and matched against coded dictionaries so as to identify the words (Feng, Ling, 2004). Speaker identification (SI) refer to the searching for identity of an speaker by matchingthe tested voice with all others speakers in the database. It's a one-to-many comparison (J. A. Markowitz and colleagues,2003).

Speaker identification has been used in a variety of criminal cases, including murder, rape, extortion, drug smuggling, wagering-gambling investigations, political corruption, money-laundering, tax evasion, burglary, bomb threats, terrorist activities and organized crime activities. Forensic acoustic analysis also involves tape filtering and enhancement, tape authentication, gunshot acoustics, reconstruction of conversations and the analysis of any other questioned acoustic event (C. Srividya1, S.R. Savithri, 2011).

A speaker verification process containing two phases, the first is training phase and second is test phase. The first and foremost module is the feature

extraction module conveying speaker information extracted from the speech. This is the pedestal module, where the entire system performance relies. The next module is speaker modeling module, represent that speaker's voice and acoustic features.

The most commonly used acoustic vectors are Mel Frequency Cepstral Coefficients(MFCC), Linear Prediction Cepstral Coefficients(LPCC) and Perceptual Linear Prediction Cepstral(PLPC) Coefficients and zero crossing coefficients (Yegnanarayana et al., 2005; Vogt et al., 2005). These methods use the spectral information extracted from a short time windowed segment of input signal. The difference focus on the representation details of power spectrum. In this work mel frequency cepstral coefficients (mfcc) are used to extract voiced signal feature.

Vector quantization VQ (A. Gersho, R. M. Gray, 1992, A. Gersho, 1982) is an important and powerful technique for data compression. Speech and image signals compression are the usual applications of vector quantization. Vector quantization and more generally clustering techniques use a set called codebook for the reference vectors called codeword which can be derived from input data, named training set. Codeword are obtain in order to minimize a goal function representing the quantization error. Using the codebook, each single codeword represent vector of input data.

The performance for many vector quantization algorithms depends essentially of the choice of the initial codebook (A. Gersho, R. M. Gray, 1992).

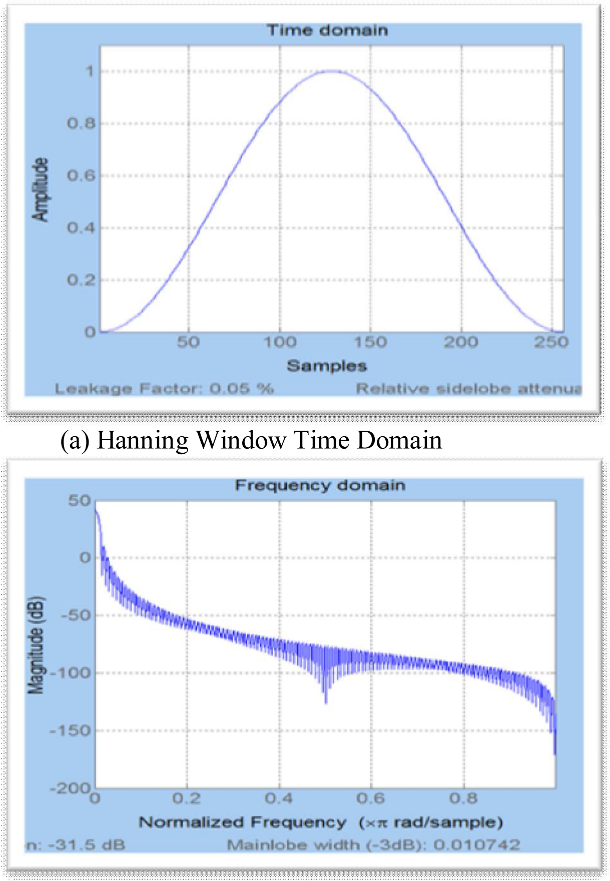
2. The Proposed system :

The first stage of the proposed system is speaker signal acquisition by recording the specific words by different speaker. The database used in this paper, containing recording of seven specific word by 25 speakers (13 men and 12 women), Table 1 show database details. After getting speech data, we starting with preprocessing of speaker data after it converted to digital form. Generally the most of recorded voice samples corrupted by amount of noise. Also, the recorded voice contain a durations at the beginning of and end of voice data, which called silent or unvoiced signal. These noise and silent part would decrease the quality of results, Therefore, we can use two parts of preprocessing for our data before extract the important feature from voice data, the first part called silent removal and filtering process.

2.1 Eliminate Silent from speech signal

In order to eliminate silent part from the recorded voice, we use the algorithm of energy detection that compute short-time energy and spectral centroid of the signal. We use the algorithm recursively for each frame of data contain 256 sample which equal approximately 23ms and eliminate each frames that have energy level lesser than the noise floor using energy threshold.

2.2 Filtering with Hanning Window



(a) Hanning Window Time Domain

(b) Hanning Window frequency Domain

Figure 1. Hanning window for a frame contain 256 sample

The noise considered to be composed of high frequencies of the digital signal, therefore we select the low-pass filter of the voice signal in order to enhance S/N ratio. In the proposed system we select Hanning window to perform filtering process as shown in the following figures 1(a) for Time Domain and 1(b) frequency Domain.

2.3Feature Extraction Using Mel-frequency cepstrum coefficients

After the preprocessing process complete, we get voice data ready to extract the important information from the signal but with less amount of data for further analysis. This process generally called the *signal processing front end.*

MFCC consider important method because it based on the variation of critical bandwidths for human ear with frequencies.

Figure 2 shows mel frequency method steps. We recorded the input signal at a sampling frequency more than 10 Kilo Hz. This sampling rate is selected to decrease *aliasing* problems, when the analog signal is converted to digital one. The sampled signals has the ability to overcome most of frequencies till 5 KHz, that will cover most energy of sounds that human can be generated.

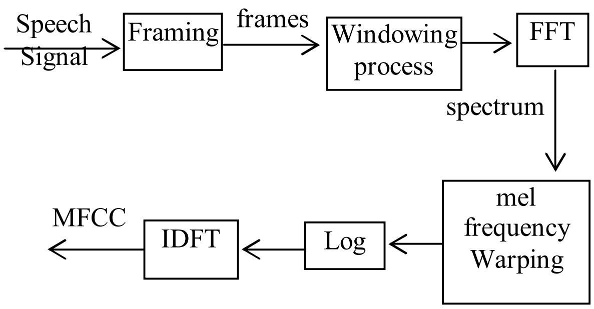


Figure 2. Steps of the MFCC

2.3.1 Framing:

Framing process convert the continuous input signal into frames, each one containing specific no. N of samples. Also it is useful to use frame overlapping. In such method we overlap first frame of N sample with next frame that begin with M such that M<N and then overlap frames with N-M samples and so on. We select *N=256* and *M* = 128.

2.3.2 Windowing Process :

Windowing process is performed by passing a window for each frame of the voice signal in order to decrease the discontinuities of input signal at the beginning and end of each frame. The means decrease the spectral distortion by approximating the beginning and end of each frame to zero. In the proposed system we use *Hanning* window as shown in the following equation:

$$w(i+1) = 0.5\left(1 - \cos\frac{2\pi i}{255}\right), \quad i = 0,..,255 \qquad (1)$$

2.3.3 Fast Fourier Transform (FFT):

Fast Fourier Transform, perform converting the frames (256 samples) from the spatial domain into the frequency domain. The Fast Fourier Transform is applied in our system which is applied on the set of N=256 samples f_n, as shown:

$$F_k = \sum_{n=0}^{N-1} f_n e^{-j2\pi kn/N}, \qquad k = 0,..,N-1 \qquad (2)$$

Generally F_k is a complex numbers and their absolute values frequency magnitudes are considered. After applying FFT on each frame we got signal spectrum.

2.3.4 Mel-frequency Wrapping:

In the mef frequency wrapping process a scale value is used to measure each pitch called the 'mel' scale. This scalar is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

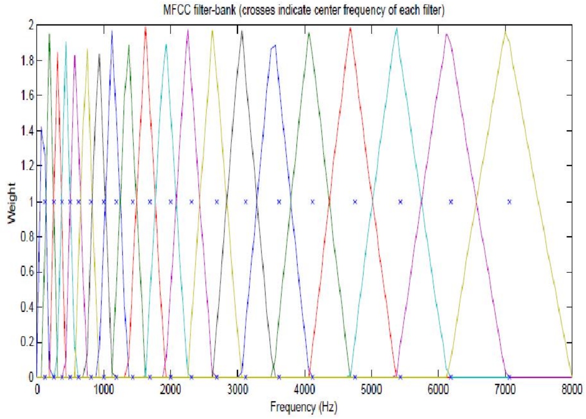


Figure 3. Mel-space filterbank

Filter bank is used to simulating the used subjective spectrum. This bank with a triangular bandpass frequency response. The selected no. of mel spectrum coefficients is mfc No= 20.

2.3.5 Cepstrum:

In the last stage the log mel spectrum converted to time domain again. The results of this step is the final mel frequency cepstrum coefficients which represent the speech spectrum. We use discrete Cosine transform DCT in order to convert the spectrum coefficients into time domain, because the result of applying the logarithm for mel spectrum coefficients are real values.

To get final result of mfcc we can use the following equation:

$$C_n = \sum_{k=1}^{K} (\log S_k)\cos\left[n(k-0.5)\frac{\pi}{K}\right], \quad n = 0,..,k-1 \qquad (3)$$

S_k : represent mel spectrum coefficients
$n=0,..,K-1$

2.4 Vectorquantization

In general, vector quantization VQ algorithm consider the perfect clustering tool to identify the correct speaker by clustering and matching feature vectors. We use vector quantization because it is not logically to represent each feature vector in feature space that generated from the specified speaker training utterance. Also VQ algorithm can saves the time in the testing phase because it depend on few feature vectors instead of large space of feature for specific speaker. In VQ algorithm every vector is called codeword (centroid) and these to fall codeword referred as codebook. Therefore VQ algorithm can generate codebooks for all speakers in the training phase. For all codeword there was nearest neighbor area as associated with it called Voronoi region (Du, Qiang, Vance Faber, and Max Gunzburger, 1999), and is defined by:

$V_i=\{ x \in R^k : ||x-y_i|| \le ||x-y_j||$, for all $i \ne j$
R^k: vector space
$Y=y_i$: finite set of vectors

The set of Voronoi areas partition the entire feature space of a given user is-

$$\bigcap_{i=1}^{N} V_i = R^k \qquad (4)$$

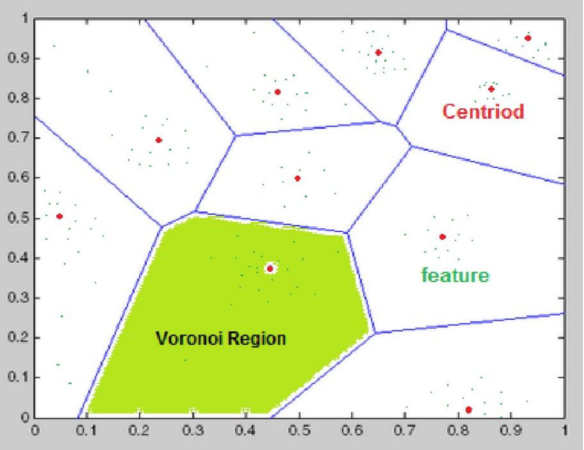
$$\bigcap_{i=1}^{N} V_i = \phi \quad \text{for all } i \ne j \qquad (5)$$

Generally, Voronoi region example can be shown in figure 4, in this 2D figure, the feature vectors contain two feature coefficients, therefore we represent each one in 2Dfeaturespace. Also we can see that, all feature vectors that associated with nearest neighbor using vector quantization algorithm and produce all centroids. All centroidsexist in own area named Voronoi area . Each area is bounded with virtual lines as shown in figure 4. For input vector, we got the Euclidian distance between each centroid and point with smallest distance that nearest for that vector. The Voronoi area associated with the given centroid is cluster area for the vector. The following equation represent the Euclidean distance:

$$dist(x, y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2} \qquad (6)$$

x_j is the jth element to input vector,
y_{ij} is the jth element of centroid y_i.

Example of Voronoi areas bounded by blue lines, Red points represent centroids, Green points represent features vector



In speech recognition, the computation component mostly depend on determine the spectrals similarity between two vectors. In the following figure illustrate the stages of speaker recognition using VQ:

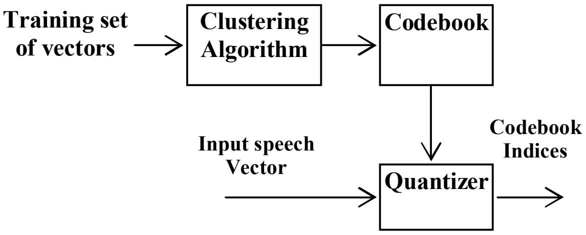


Figure 5. VQ model for speaker recognition

The VQ module perform speaker recognition process by generate the codebook for each speaker as shown in figure 6:

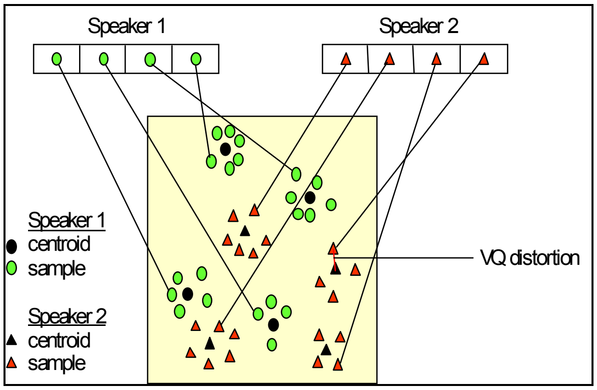


Figure 6. VQ module of generating the codebooks

In figure 6, there are two feature vectors extracted from speakers and then applying VQ algorithm. Each speaker generate four centroid. The set of all centroid represent the codebook for single speaker, therefore two code books are generated. In the testing phase, feature vectors of test speaker are mapping with the feature space, and then calculate Euclidian distance for all features vector and find the nearest centroid.

The resulted shortest distance referred as VQ distortion, and all VQ distortions are calculated for the rest feature vectors. The same process is also applied for the second speaker. The desired speaker founded by the smallest summation of the VQ distortions.

By applying the previous procedure, for all frames of 23ms with overlap, and then compute a set of mfcc.

2.4.1FeatureMatching

In the proposed system the classification process is applied on the features that extracted from acoustic voice therefore, it can be referred as feature matching.

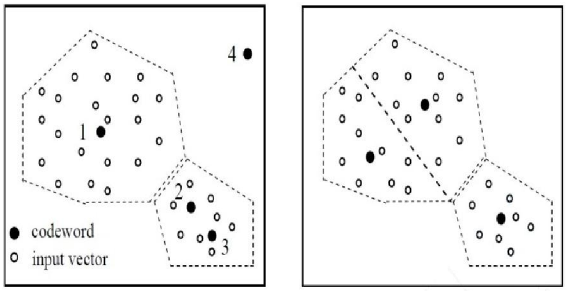
Our database is split into two sets: Training Set that used in the training phase. Testing Set: used in testing phase. In this paper, we enhance new classification method based on LBG for vector quantization algorithm. Figure 6 illustrate the process of speaker recognition. There was two speakers, the green circles represent speaker 1vectors, and the red triangles represent speaker 2 vectors. using the clustering algorithm, to generate the speaker-specific VQ codebook all speaker in the training phase. figure 7 shows the results of centroids by black circles and black triangles for speaker 1 and 2, respectively. The distance of a vector to the nearest centroid of a codebook is referred as VQ distortion. In the testing phase, an unknown speaker input is quantized using codebook with VQ distortion.

VQ codebook of the speaker corresponding to one have lesser distortion is identified .

2.4.2 Clustering the Training Vectors

After extracting feature vectors for each input speaker, we prepare a set of training vectors for that speaker. And then construct speaker-specific codebook for each speaker using training vectors. In the proposed system we enhance common algorithm LBG algorithm (Memon*et al.*, 2009; Alsulaiman*et al.*, 2010), for clustering a set of training vectors into a set of codebook vectors.

There was two disadvantages for the classical LBG algorithm, first it always generate an empty cluster because all the input vectors are nearer to the other codewords and classical LBG method cannot move and represent an input element. Second disadvantages can be shown in figure7, where two clusters and three codewords, one codeword have large cluster and the 2 others with smallest cluster. The large cluster show bad approximation with single codeword, while the other cluster show good approximation.



(a) (b)

Figure 7. Cluster and codewords

2.4.3Enhanced LBG algorithm

In order to avoid the limitation of classical LBG algorithm, we introduce an improvement based on the previous improvements describes in (B. Fritzke, 1997) and (G. Patane, M. Russo, 2001).First, on compute the mean quantization distortion over all cells:

$$D_m = \frac{1}{N}\sum_{i=1}^{N} D_i \qquad (7)$$

and define, for each cell, a utility measure Ui, as the value of distortion Di of the *i-th* cell, normalized with respect to the mean value *Dm*:

$$U_i = \frac{D_i}{D_m} \qquad (8)$$

For an optimum VQ, according to Gersho theorem (A. Gersho, 1979) each cell have the same contribution to the total distortion and consequently the utility index for all cell is equal to unity.

The equalization of cells utility is obtained by joining each cell with a low utility index with a cell adjacent to it, trying to obtain a bigger cell with high utility. At the same time, each high utility cell will be split in two smaller cells, equivalent to moving one codeword from the low utility cell inside the high utility cell, as in figure 7-b.

The split operation, illustrated in figure8, suppose that a cell is contained in a k - dimensional hyperbox defined by:

$I = [X_{1m}, X_{1M}] \times [X_{2m}, X_{2M}] \times .... \times [X_{km}, X_{kM}]$

where $X_{im}$and X_{iM} are the minimum and the maximum value by the *i* dimension of the cell. The two new codeword are placed on the principal diagonal of the hyper box *I* at equal distance of the hyperbox center. The improvement procedure selects each cell with a lower utility, in a sequential manner, and as improvement of previous algorithm we propose to try to minimize the overall distortion by using each cell with a utility higher that 1:

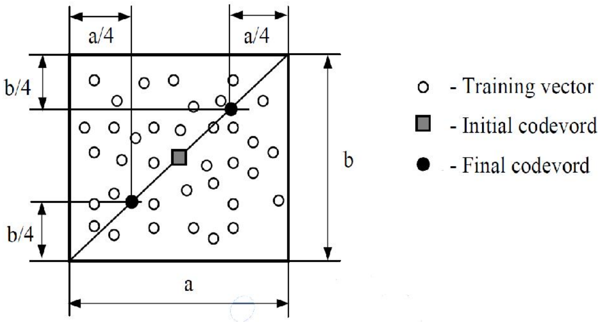


Figure 8. Splitting a codeword

For this, we compute the old distortion, before the codeword movement:

$D_{old} = D(\{Y, C\})$

and the new distortion corresponding to the new partition Y':

$D_{new} = D(\{Y', P(Y')\})$for all the moving possibility. The new situation with the minimum D_{new} is retained.

The steps of proposed algorithm performed by the following procedure:

Step 1 : construct the codebook vector containing the centroid for all of training vectors **.**

Step 2: increase codebook size through split each current codebook C_n according to following rule:

$$\mathbf{c}_n^+ = \mathbf{c}_n(1+\alpha)$$

$$\mathbf{c}_n^- = \mathbf{c}_n(1-\alpha)$$

where *n* takes values between 1 and the current size of the codebook, and the splitting parameter α= 0.01.

Step 3: Search for Nearest-Neighbor: In the current codebook, search for the nearest codeword for each training vector, and then assign this vector to the corresponding cluster.

Step 4: Update the Centroid: update the codeword in each cluster through the training phase .

Steps 3 and step 4 are repeated continuously until average distance exceeds the threshold and construct codebook with *M* size.

The enhanced LBG algorithm build codebook *with M*-vector, It firstly design 1-vector codebook, then split the code words to start the search for a two vectors code book, and continues the splitting until it reached *M*-vector codebook. Figure 8 shows the steps of the LBG algorithm. "Cluster vectors" is the nearest-neighbor search procedure which assigns the training vector to a cluster associated with the nearest centroid. "Find centroids" is the procedure that will update the centroid. "Compute Distortion D" sums the distances of all training vectors in the nearest-neighbor search in order to determine the convergence of procedure.

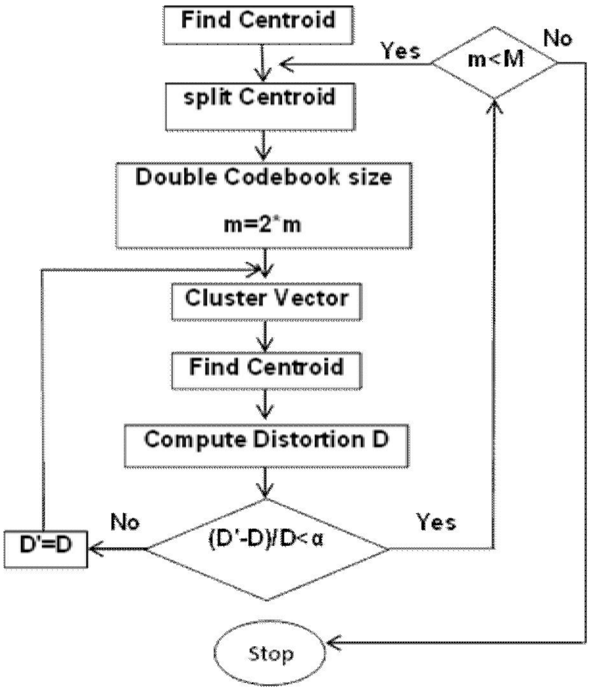


Figure 9. Steps of the LBG algorithm

3 Result and Discussion

Training and testing data are taken for each speaker in two sessions. Seven different texts are recorded so that text-dependent speaker identification can be done.

LBG algorithm is used for the generation of 25 code vector for the 175 dimensional vector space. The vectors are used in the training phase.

Database *is used to store* the feature vectors of all the reference speech samples of the training phase. the test sample that is to be identified is taken in the matching phase, and similarly processed as in the training phase to form the feature vector. The stored feature vector that gives a minimum Euclidean distance with the input sample feature vector is chosen as the speaker identified.

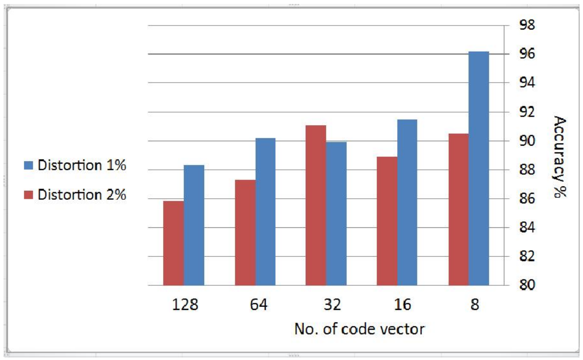


Figure 10. Variation in the number of code vectors with overlap

The curves obtained for text-dependent system by varying the number of feature vectors (code vectors) with overlap for a sample set of 25 speakers is shown in Figure10. As seen from the figure, for text-dependent samples, maximum accuracy is achieved with eight feature vectors. The maximum accuracy is 96.22% for a distortion of 0.01 and 90.51%for a distortion of 0.02. The accuracy decreases when the number of code vectors are increased.

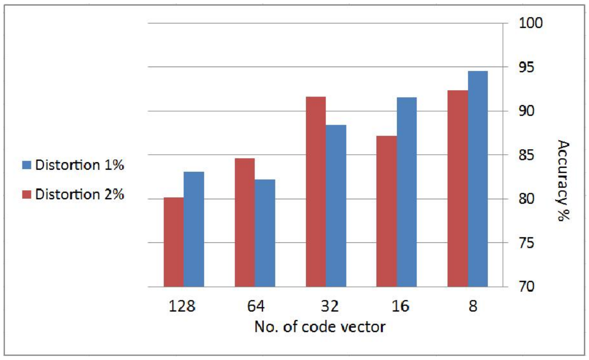


Figure 11. Variation in the number of code vectors without overlap

The curves obtained for text-dependent identification for a sample set of 25 speakers without overlap are shown in Figure11. The maximum accuracy is 94.61% for 8 code vectors for a distortion of 0.01 and 92.33% for distortion of 0.02.Again here also the accuracy decreases as the number of code vectors increases.

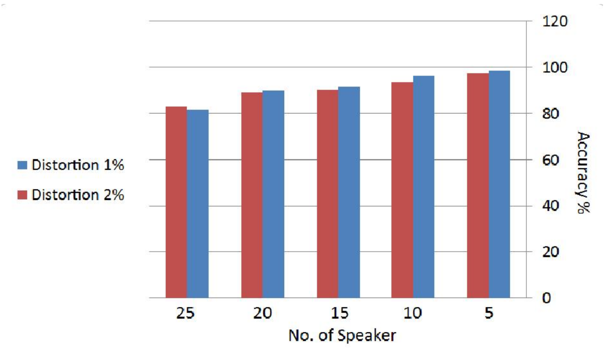


Figure 12. Variation of accuracy with the number of speakers with overlap for a codebook size of 8

Figure 12 shows the variation inaccuracy with the number of speakers with overlap. As can be seen the accuracy decreases as the number of speakers increases.

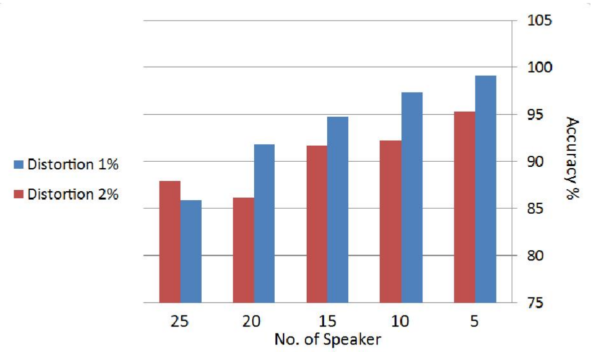


Figure 13. Variation of accuracy with the number of speakers without overlap for a codebook size of 8

Figure 13 shows the variation in accuracy with the number of speakers without overlap. Here also the accuracy decreases with the increase in the number of speakers.

4. Conclusion

A new speaker recognition scheme is proposed, and the proposed system uses MFCC features for identification and enhanced LBG algorithm for speaker recognition. As can be seen from the results, approach A with overlap gives better results than without overlap. Also maximum accuracy is achieved for 8 code vectors. The proposed system can be extended to text independent speaker recognition system.

References

1. Feng, Ling."Speaker recognition", Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2004.
2. C. Srividya1, S.R. Savithri, Speaker identification using cepstrum in Kannada language, Forensic Science Journal; 2011,10(1):1-10.
3. A. Gersho, R. M. Gray, Vector quantization and Signal Compression, Kluwer Academic Publishers, London 1992.
4. A. Gersho, On the Structure on Vector Quantizers, IEEE Trans. Inf. Theory, IT vol. 28, march 1982.
5. A. Gersho, R. M. Gray - Vector quantization and Signal Compression, Kluwer Academic Publishers, London 1992.
6. Memon, S., M. Lech and N. Maddage, 2009. Speaker verification based on different vector quantization techniques with gaussian mixture models. Proceedings of the 3rd International Conference on Network and System Security, Oct. 19-21, Gold Coast, Queensland, Australia, pp: 403-408.
7. Alsulaiman, M., Y. Alotaibi, M. Ghulam, M.A. Bencherif and A. Mahmoud, 2010. Arabic speaker recognition: Babylon levantine subset case study. J. Comput. Sci., 6: 381-385. DOI: 10.3844/jcssp.2010.381.385.
8. B. Fritzke, The LBG-U Method for vector Quantization – an Improvement over LBG inspired from Neural Network, Neural Processing Letters, vol. 5, no. 1, 1997.
9. G. Patane, M. Russo, The enhanced LBG Algorithm, Neural Networks, vol. 14 (no. 9), nov. 2001.
10. A. Gersho, Asymptotically Optimal Bloc Quantization", IEEE Trans. Inf. Theory, IT vol. 25, April 1979.
11. Du, Qiang, Vance Faber, and Max Gunzburger, Centroidal Voronoi Tessel-lations: Applications and Algorithms, SIAM Review 21 (1999): 637-676.
12. Saha, G. and U.S. Yadhunandan, 2000. Modified Mel- Frequency Cepstral Coefficient. Prince of Songkhla University. http://fivedots.coe.psu.ac.th/~montri/Research/Publications/icep2003_modified.pdf.
13. Vogt, R., J.B. Baker and S. Sridharan, 2005. Modelling session variability in text independent speaker verification. Proceedings of the 9th European Conference on Speech Communication and Technology, Sept. 4-8, Lisbon, Portugal. http://eprints.qut.edu.au/15490/.
14. Yegnanarayana, B., S.R.M. Prasanna, J.M. Zachariah and C.S. Gupta, 2005. Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. IEEE Trans. Speech Audio Proc., 13: 575-82. DOI: 10.1109/TSA.2005.848892.

1/15/2015