Dhaka Univ. J. Eng. & Tech. Vol. 1(2) 131-136, 2011 (January)

A Text Dependent Speaker Recognition Using Vector Quantization

Rajib Ahmed, Rifat Ahhmed[1], Md. Moqbull Hossen and Mir Zayed Hasan

Dept. of Applied Physics, Electronics & communication Engineering, University of Dhaka, Dhaka, Bangladesh.

[1]Department of Electronic & Telecommunication Engineering , Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh

rajib.ahmed.apece@gmail.com, rifat.ete07@gmail.com, moqbul_bco@yahoo.com, zayed@univdhaka.edu

Received on 04. 07. 2010. Accepted for publication on 29. 01. 2011

Abstract

Nowadays, security concern is an important issue in keeping own identity secret to others, allowing only specified person to access, storing user database and so on. Speech based identification is used to verify the person's identity for controlling access to services such as database access, banking by telephone, voice dialing, telephone shopping, information services, voice mail, security control for confidential information areas, and remote access to computers. Speech is used for identification of a human as the characteristics of vocal cord are different in each individual. On the basis of discriminatory information in speech waves, a specific speaker can be identified. In this paper, a microcomputer based speech recognition system has been designed using Vector Quantization (VQ) as a basis for identification. All speakers were modeled by a codebook of 32 vectors using LBG (Linde, Buzo and Gray) splitting algorithm. The speakers were prompted to say their nick name, different names for different speaker that is why it is called text dependent speaker recognition. Recognition decision is taken on the basis of the lower VQ distortion value with database speech and new samples of register user. If register user give the voice then the system measures VQ distortion values and based on lower VQ distortion value identification is done. Unknown speaker is discarded, based on comparison with threshold of each database speaker. We able to get 91.67% recognition rate with 12 database speaker. These recognition rate decreases as the numbers of speaker increase as decisions are made on comparisons of various VQ distortion data- causes' greater chance to make mistake.

Keywords: Speaker Recognition, Vector Quantization, Acoustic Vector, Codebook, LBG Algorithm.

1. Introduction

Havę we ever talked with our computer? We mean, have we really, really talked with our computer, where it is actually recognized what we said and did some things as a result? If we have, then we have used a technology known as speech recognition. Speech recognition allows us to provide input to an application with our voice. Simply, it is the process of converting spoken words into texts so that an application understand and do something as a response. Just like clicking with our mouse, typing or pressing a key on our keyboard. In the desktop world, we need help of microphone to able this work. In the voiceXML world, all we need is a telephone. Where we might say something like "checking account balance" to our bank's voiceXML application replies "one million, two hundred twenty-eight thousand, six hundred ninety eight dollar and thirty seven cents. ("We can dream, can't we?) [1].

There are several techniques and methods (already devolapted) for implementation of voice recognition system like Markov Modeling (HMM), Dynamic Time Warping (DTW), Vector Quantization (VQ) and so on [5]. In this paper, a very simple concept on Vector Quantization (VQ) is used for speech signal identification of each individual. Rather than other processes mention above it is simple and easy. Moreover, the system using Vector distortion values as the basis of identification gives higher success rate for small database (10 or 12 User) and decrease success rate for number of speaker increase.

2. Speaker Recognition

If we are able to recognize which of the population of subjects spoke a given utterance then the process is called speaker recognition process. Speaker recognition may be categorized as[2,3,4] (i) speaker identification and (ii) verification. Speaker identification is the process to take the speech signal from known speaker and compares this with database samples of a set of valid users. The best match is then used to identify the speaker. Thus it is the detection process of a particular speaker from a known population. In speaker verification, unknown speaker first claims identity and then claimed model is used for identification. If the matching is above a predefined threshold, the identity is accepted otherwise rejected. Both Speaker identification and verification may be (i) Text Dependent (recognition of specific speaker based on specific text) and (ii) Text independent (recognition of speaker may be for any text) [2,3,4].

In Text dependent Speech recognition systems that require a user to train the system to his/her voice. Thus in speech recognition system requires knowledge of a speaker's individual voice characteristics to successfully process speech. In this case the speech recognition engine can "learn" how we speak words and phrases; it can be trained to our voice. Speech recognition systems that do not require a user to train the system are known as speaker-independent or Text independent systems. Speech recognition in the VoiceXML world is an example of speaker-independent system[1,2,3].

Fig. 1.1 & Fig. 1.2 shows two basic classification of speaker recognition [3, 4] :

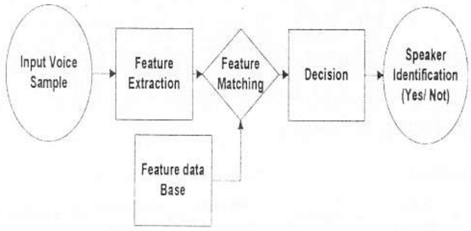

Fig. 1.1: Speaker Identification

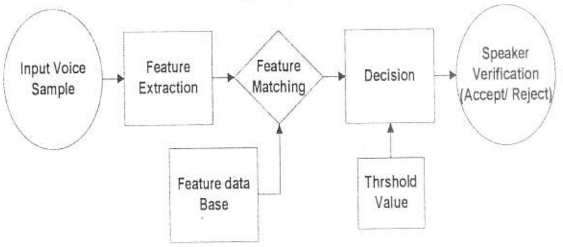

Fig. 1.2: Speaker Verification

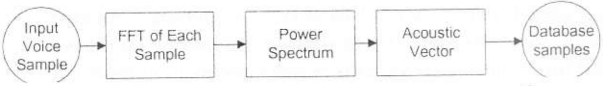

Fig. 1.3: Training phase (Database Formation)

A Speech Recognition process involves 2 modules[3,4] which are (i) Training & (ii) Testing module. In training module, all speakers are requested to give their voice to system to build reference models and store it as database (shown in Fig.1.3). In testing module, new samples are compared with database and make decision according to it. Moreover, all speaker system has to serve two distinct sections [3,4]. Sections are (i) Feature extraction and (ii) Recognition section In feature extraction section, a small amount of data is extracted from voice sample that can later be use for identify the speaker. Actual work is done in recognition section. Here, identification of the speaker is done by comparing the extracted voice data with a database of known speakers and based on this a suitable decision is made. Training module and testing module are part of feature extraction and feature classification section. Now, we are discussing the whole process in terms of speech feature extraction and feature classification.

3. Feature Extraction

The speech signal is quasi-stationary i.e. a slowly timed varying signal and short-time spectral analysis is used to characterize the speech signal. There are several process to extract feature characteristics from input training data that later be used for identification purpose like (1) Real Cepstral Coefficient(RCC) (2)Linear Prediction Coding (LPC), (3) Mel-Frequency Cepstral Coefficient(MFCC) [2,3]. We consider, MFCC to extract voice features and analyze their characteristics. MFCC is a filter, spaced linearly at low frequencies and logarithmically at high to mimic response of the human ear with voice frequency. The MFCC process has five phases or blocks[2, 3] as: (i) Frames blocking phases, (ii) windowing phases, (iii)FFT phase, (iv) Mel-frequency wrapping phase, (v) Cepstrum phase. All the phases are shown in figure bellow:

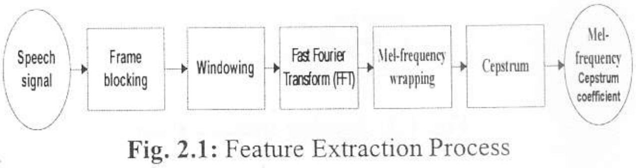

Fig. 2.1: Feature Extraction Process

(i) Frame blocking:

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 overlapping segment, $M = 100$. [2, 3]

(ii) Windrowing:

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. Thus minimize the spectral distortion of the speech signal by using the window to taper the signal to zero at the beginning and end of each frame.

In windowing phase, sample in each frame, x (n) is multiplied by window function, w(n) to get output[2,3]:

Y (n) = x (n) * w (n).................................. (1)

We use Hamming window with window function define as,

W (n) =.54-.46cos ((2*pi*n)/(N-1)) , 0<n<N-1

...........(2)

(iii) Fast Fourier Transform [2,3]

The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples $\{x_n\}$, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,\ldots,N-1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

In general X_k's are complex numbers and we only consider their absolute values (frequency magnitudes). The resulting sequence $\{X_k\}$ is interpreted as follow: positive frequencies $0 \le f < F_s/2$ correspond to values $0 \le n \le N/2-1$, while negative frequencies $-F_s/2 < f < 0$ correspond to $N/2+1 \le n \le N-1$. Here, F_s denote the sampling frequency.

The FFT phase converts time domain signal in each frame in to the frequency domain. Most of the feature of the signal contain in the frequency domain so use FFT. Thus, continuous speech is sampled with N frames, multiplied with Hamming window function and finally transform into frequency domain to get a matrix, each column of which is the power spectrum representation of speech signal. Fig.2.2 shows the power spectrum with different values of M & N.

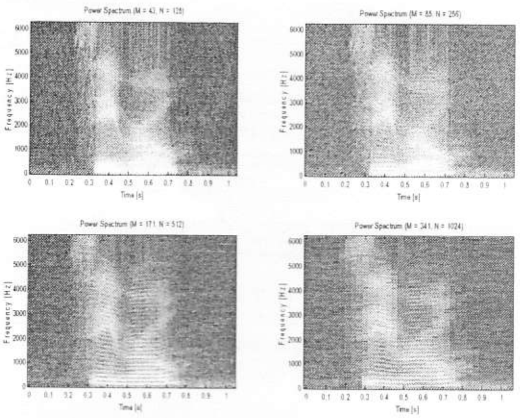

Fig. 2.2: Power spectrum with different values of M & N.

(iv) Mel-frequency wrapping [2,3,4]

Psychophysical studies in human body have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale but follows a scale is called the 'mel' scale. The *mel-frequency* scale which has linear frequency scale below 1kHz and logarithmic scale after 1 KHz known as Mel-Frequency Scale. The Linear frequency f Hz is converted into Mel-frequency by the equation:

M (f) = 2595 log (1+f/700)............................. (4)

In this *Mel-frequency warpping,* the output power spectrum is multiplied with linear [f <1 kHz] and logarithmic [f >1 kHz] Mel-frequency scales to mimic the human ear perception of sound. With this Mel-frequency scale the features of the voice samples will be more visualize. The output of Mel-frequency filter bank has has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The number of mel spectrum coefficients, *K* is typically chosen as 20. For our experiment, we use 30 triangular band pass filter with K= 30.Feature of voice sample in power spectrum before and after the Mel-frequency scale is shown in Fig.2.3. Moreover, from Fig.2.3, we get most of the important characteristics exist within the frequency range bellow 1 KHz within the time scale .3 to .75 second.

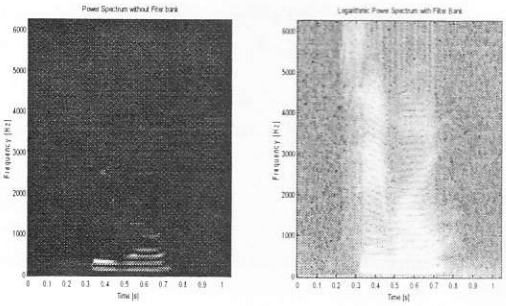

Fig. 2.3: Power Spectrum of Voice Signal

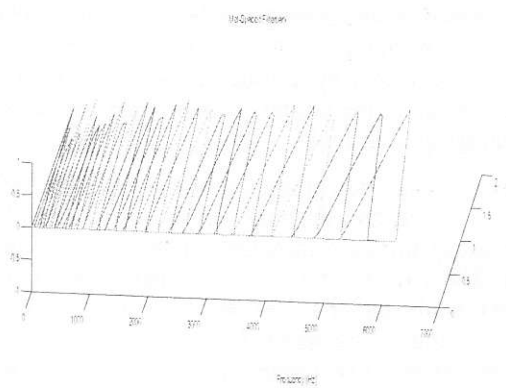

Fig. 2.4: Mel- Frequency filter Bank

(v) Mel-Frequency Cepstral Coefficient [4, 5]

Each power spectrum of speech signal is multiplied by filter gain, added together and perform Discrete Cosine Transform (DCT) to convert the frequency domain signal to previous time domain signal which is known as Mel-Frequency Cepstral Coefficient. This set of coefficient is called the Acoustic Vectors. Therefore each input sample is characterized into a sequence of "Acoustic Vectors". Fig.2.5 shows the two dimensional plot of "Acoustic Vectors" for two samples say Sample1 and Sample2 of speaker1 and speaker2. Similarly, we can plot "Acoustic Vectors" for other speaker also. Therefore, we found different "Acoustic Vectors" for different Speaker that can be used for feature classification in the later section.

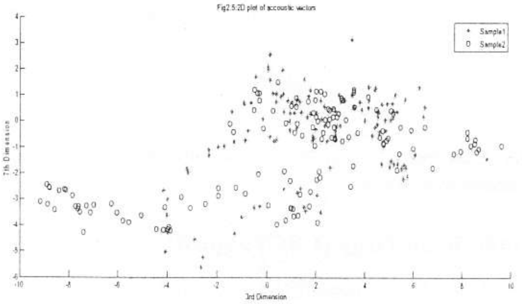

Fig. 2.5: show acoustic vector recognition of two Speaker.

4. Feature classification

Feature classification consists of feature modeling and feature matching subsections. In feature modeling, each speaker is requested to produce a voice signal (say, nick name) to build up a database model (as shown in fig 1.3). In feature matching, decision is made by comparing new samples with stored data base. There are several techniques for this purpose such as Dynamic Time Warping (DTW), Markov Modeling (HMM), Vector Quantization (VQ) and so on. Here, we have used Vector Quantization for finding feature matching.

4.1 Vector quantization (VQ)

Vector quantization (VQ) is a lossy data compression method based on the principle of block coding. VQ is the process of mapping vector to produce regions called clusters, represented by a code vector. Code vector is the average vector of the cluster. Sets of code vector stored in the database as codebooks.

Moreover, recognition process can be shown more visualize with in a conceptual diagrams as shown in Fig.3.1. Here, two speakers in two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker1 while the triangles are from the speaker2. The cluster of acoustic Vectors is known as codebook is shown by doted circles. The center of each code book is called Centroid as shown by black circles and black triangles. Each codebook refers to a specific speaker. The distance from Centroid to an acoustic vector to the closest codeword of a codebook is called a VQ-distortion. In the feature matching phase, an input voice of an unknown person is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the input voice.

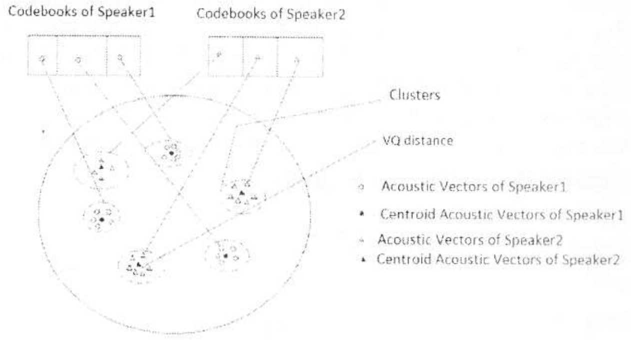
Codebooks of Speaker1
Codebooks of Speaker2
Clusters
VQ distance
Acoustic Vectors of Speaker1
Centroid Acoustic Vectors of Speaker1
Acoustic Vectors of Speaker2
Centroid Acoustic Vectors of Speaker2

Fig. 3.1: Conceptual diagrams for two speakers in two dimensions of the acoustic space[4]

4.2 Linde-Buzo-Gray (LBG) algorithm [2,5]

The codebook is generated by using Linde-Buzo-Gray (LBG) algorithm. The algorithm requires an initial codebook. Then the desired codebooks are obtained by the splitting method. Thus Feature vectors obtain from feature extraction module is averaged to get initial code vector:

$$C_1 = \left(\frac{1}{M}\right) * \sum_{n=1}^{n=M} X(n) \qquad \ldots\ldots\ldots\ldots(5)$$

This is then split into two,

$$C_i = (1+\xi)\, C^*_i \qquad \ldots\ldots\ldots\ldots(6)$$

$$C_{N+i} = (1-\xi)\, C^*_i \ldots\ldots\ldots\ldots(7)$$

Where, C_1 = The initial codebook,

C_i, C_{N+i} = The first i^th^ and its next codebook,

X (n) = The feature vectors,

ξ=small positive number = Splitting parameter, =.oo1

i= 1, 2, 3,.............................N, and

n= 1, 2,3,.......................total no. of codebook.

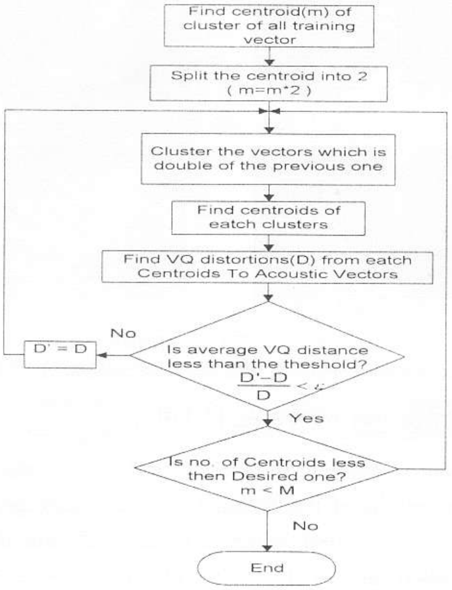
Find centroid(m) of cluster of all training vector
Split the centroid into 2 (m=m*2)
Cluster the vectors which is double of the previous one
Find centroids of eatch clusters
Find VQ distortions(D) from eatch Centroids To Acoustic Vectors
No
D' = D
Is average VQ distance less than the theshold? $\frac{D'-D}{D} < \xi$
Yes
Is no. of Centroids less then Desired one? m < M
Yes
No
End

Fig. 3.2: Linde, Buzo and Gray (LBG) algorithm [2]

The Linde, Buzo and Gray (LBG) algorithm is shown in Fig.3.2. This is a iterative algorithm in which the initial codebook is found from clustering all training vectors of database speaker. Then split centroid of initial codebook into two by using a splitting technique. Then find out closest acoustic vectors around each centroid by the nearest-neighbor search procedure which assigns each training vector to a cluster associated with the closest codeword. Each centroid surrounded with acoustic vectors form a codebook vector. The two code vectors are split into four and the process is repeated until the desired number of code vectors (M) is obtained. Finally, VQ distortions are measured which are the sums of all distance from codewords to centroids. VQ distortions determine whether the procedure has convergence or not.

The generated acoustic Codebooks for two speakers with the LBG algorithm to is shown in Fig.3.3.

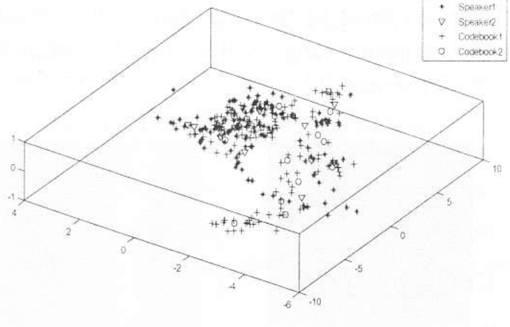

Fig. 3.3: Acoustic Codebooks of Speaker1 and Speaker2

5. Experimental Setup

Using MATLAB a data base of 8 speaker's voice signal is created in .wav format with sampling frequency of 16000 Hz. In the process, each speaker is prompted to say his nick name in clear steady voice. The duration of the database samples and test samples are chosen to be 2 seconds. This database samples are used for recognition process. After introducing itself the system at first requests the speaker to give voice (nick name) and waits for 2 second. If the user is registered, the system will identify the speaker and welcome the user saying the user's name and gives bio-data with his picture and so on. For unknown users who are not in the database, the system indicates that the speaker is an unauthorized person and a request is made again by the system for speaker's voice. The system then waits for 2 second and if no response is found the program terminates. Based on microcomputer, an automatic speaker recognition system has been designed in this paper.

6. Result Analysis

We build up a data base of 8 samples from different speakers with approximately same age (all males). The sample1 is stored in database as s1 and sample2 as s2 and so on. The sound waves that are recorded have the following features:

Bit rate	: 256kbps
Audio sample size	: 16 bit
Channels	: 1(Mono)
Audio sample rate	: 16 kHz
Audio format	: PCM

In the feature extraction section, the voice signals are frame blocked by 30 ms Hamming window and 33.30% overlapping i.e frame size = N = 256 samples and overlap samples ,M = 100 samples. A 30 triangles filter bank is used for preprocessing the signal.

A threshold value for each speaker is calculated using VQ distances. The threshold value is the largest VQ distortion with a speaker with reference to all database speakers.

Table 1: Threshold values for all 8 speakers.

Speaker1	Speaker2	Speaker3	Speaker4	Speaker5	Speaker6	Speaker7	Speaker8
8.2194	8.1904	7.9384	9.0918	7.3609	8.0953	7.1423	7.5819

In VQ Method, decision is made on the basis of VQ distortion value. The VQ distortion values are calculated by comparing the new sample (test sample) with database samples (Train samples) and base on smaller VQ distortion value speaker is identified [5]. For example, when the speaker1 give his voice (nick name), the VQ distortion values are:

Table 2: VQ distortion values for new testing data of speaker1.

D11	D21	D31	D41	D51	D61	D71	D81
4.9100	6.2155	6.4954	5.1050	5.3518	6.1168	6.5403	5.2826

Where, D11= VQ distortion value between new sample1 and database sample1,

D21= VQ distortion value between new sample1 and database sample2 and so on.

From the table, It can be seen that D1 is the lowest of the other values. Therefore, the system takes the decision that the speaker is s1 or speaker1. Results obtained for the rest of the speaker's are given bellow:

Table 3: VQ distortion values for new testing data of speaker 2,3,4,5,6,7 and 8.

Speaker2	
D12	8.2194
D22	**4.3422**
D32	7.9384
D42	9.0918
D52	6.5501
D62	8.0953
D72	6.4791
D82	7.5819

Speaker3	
D13	6.5320
D23	6.7147
D33	**5.6759**
D43	7.9482
D53	6.3164
D63	5.8006
D73	**5.1295**
D83	6.0787

Speaker4	
D14	5.9751
D24	8.1904
D34	6.8804
D44	**5.1336**
D54	6.3193
D64	6.7871
D74	7.1423
D84	6.1991

Speaker5	
D15	7.1998
D25	6.5334
D35	6.8345
D45	6.0461
D55	**4.9119**
D65	6.3415
D75	6.5547
D85	5.9327

Speaker6	
D16	6.8771
D26	7.8014
D36	6.5632
D46	8.0629
D56	7.3609
D66	**4.3500**
D76	6.1707
D86	6.1193

Speaker7	
D17	6.6490
D27	6.2827
D37	6.1691
D47	6.8141
D57	5.8535
D67	6.1827
D77	**5.6466**
D87	5.7092

Speaker8	
D18	6.1822
D28	7.5073
D38	6.1691
D48	5.7452
D58	6.6560
D68	5.2674
D78	6.0457
D88	**4.7932**

From the tables, It can be seen that the distortion value is the lowest when a new voice sample matches the database sample. For example, for speaker2 lowest distortion value is D22=4.3422. Similar results are obtained (D11,D33,D44, D55, D66, D77, D88) for the speaker1,3,4,5,6,7 and 8. An error has occured for speaker3, in which lower distortion value occurred at D73=5.1295 which should be the case for the value D33for the correct recognition of speaker3.

Therefore, when speaker3 give his new sample, there may a chance of wrongly identifying the speaker speaker as s7. Thus in this work we are able to recognize 7 out of 8 speakers which produces an error rate of 12.5%. We performed the experiment on different number database speakers, the result of which is shown below:

Table 4: Variation of recognition rate with increase of database speaker:

Total database Speaker	Recognizable Speaker	Error Rate	Success Rate
5	5	0%	100%
8	7	12.50%	87.5%.
12	11	8.33%	91.67%
14	12	14.28%	85.71%

From this Table4, It is seen that the error rate increase as the number of speaker increases. This is because, as the numbers of speaker increase the rate to making mistake increases as decisions are made on comparisons of various VQ distortion data. fact that the recording conditions are not optimal (background noise, etc).Other factors such as, environment(e.g. room acoustics) ,quality of microphones, emotional state of speaker, computing power, silent parts of speech, overall signal energy etc. can affect the recognition process. Moreover, our success on that without any noise consideration and complexity of filtering, we get higher reorganization rate for small database. Moreover, in this process we may get approximately 100% recognition for small database (4/5 speaker).

6. Further Development

- ❖ Visual identification techniques may also be considered in conjunction with the speech recognition system to improve the success rate.
- ❖ The effect of large database may be further investigated.
- ❖ Use of neural network to identify and match frequency spectrums of the samples.
- ❖ Identification of speaker using this technique for any speaker word (different from the sample stored in the database).
- ❖ Use of controlled environment for recording voice Samples.

7. Conclusion

In this paper, an ideal setting, ignoring any noise effects has been considered. Actually noise affects the system and decreases its performance. The erroneous results are due to the fact that the recording conditions are not optimal (background noise, etc).Other factors such as, environment(e.g. room acoustics) ,quality of microphones, emotional state of speaker, computing power, silent parts of speech, overall signal energy etc. can affect the recognition process. Moreover, our success on that without any noise consideration and complexity of filtering, we get higher reorganization rate for small database. Moreover, in this process we may get approximately 100% recognition for small database (4/5 speaker).

References

1. Kimberlee A. Kemble,Voice Systems Middleware Education, IBM Corporation.: "An Introduction to Speech Recognition."(http://www.docin.com/p-8644426.html).
2. L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, N.J., 1993.
3. L.R Rabiner and R.W. Schafer.:"*Digital Processing of Speech Signals*, Prentice-Hall", Englewood Cliffs, N.J., 1978.
4. F.K. Song, A.E. Rosenberg and B.H. Juang, "A vector quantisation approach to speaker recognition", *AT&T Technical Journal*, Vol. 66-2, pp. 14-26, March 1987.
5. Y. Linde, A. Buzo & R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.
6. S. Furui,"Speaker independent isolated word recognition using dynamic features of speech spectrum", *IEEE Transactions on Acoustic, Speech, Signal Processing*, Vol. ASSP-34, No. 1, pp. 52-59, February 1986.
7. Wan-Chen Chen, Ching-TangHsieh and Chih-Hsu ˙ Hsu "Robust Speaker Identification System Based onTwo Stage Vector Quantization",TamkangJournalofScienceandEngineering,Vol. 11,No.4,pp. 357-366 (2008)
8. Jr., J. D., Hansen, J., and Proakis, J. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, 2000.
9. Zhong-Xuan, Yuan & Bo-Ling, Xu & Chong-Zhi, Yu. (1999). "Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification" in IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 1, January 1999. IEEE, New York, U.S.A.
10. comp.speech Frequently Asked Questions WWW site, http://svr-www.eng.cam.ac.uk/comp.speech/