Voice Recognition using HMM with MFCC for Secure ATM

Shumaila Iqbal[1], Tahira Mahboob[2] and Malik Sikandar Hayat Khiyal[3]

[1] *Software Engineering, Fatima Jinnah Women University,
Rawalpindi, Pakistan*

[2] *Software Engineering, Fatima Jinnah Women University,
Rawalpindi, Pakistan*

[3] *Software Engineering, Fatima Jinnah Women University,
Rawalpindi, Pakistan*

Abstract

Security is an essential part of human life. In this era security is a huge issue that is reliable and efficient if it is unique by any mean. Voice recognition is one of the security measures that are used to provide protection to human's computerized and electronic belongings by his voice. In this paper voice sample is observed with MFCC for extracting acoustic features and then used to trained HMM parameters through forward backward algorithm which lies under HMM and finally the computed log likelihood from training is stored to database. It will recognize the speaker by comparing the log value from the database against the PIN code. It is implemented in Matlab 7.0 environment and showing 86.67% results as correct acceptance and correct rejections with the error rate of 13.33%.

Keywords: *Voice recognition, Mel Frequency Cepstral coefficients, Hidden Markov Modeling, Forward Backward algorithm, Fast Fourier Transform, Discrete Cosine Transform, K-mean algorithm.*

1. Introduction

A computer system that automatically identifies and verifies the person by capturing the voice from a source like microphone is known as voice recognition. Voice recognition is one of the terms of biometric technology. It uses to provide any authentication to any system on the basis of acoustic features of voice instead of images. The behavioral aspect of human voice is used for identification by converting a spoken phrase from analog to digital format, and extracting unique vocal characteristics, such as pitch, frequency, tone and cadence to establish a speaker model or voice sample. In voice recognition, enrollment and verification processes are involved. Enrollment process describes the registration of speaker by training his voice features [1] [2]. And verification contains to verify the speaker by comparing his current voice features to pre stored features of voice. In real time, the verification process splits into two mechanisms. It first compares the unknown speaker to the pre stored database of known speakers on the basis of 1:N. and then it make decision of speaker to the exact match of 1:1. Where the one voice sample finally matched to only 1 template stored in the database [3]. Voice recognition has two categories text dependent and text independent. Text dependent voice recognition identifies the speaker against the phrase that was given to him at the time of enrollment. Text independent voice recognition identifies the speaker irrespective of what he is saying. This method is very often use in voice recognition as it require very little computations but need more cooperation of speakers. In this case the text in verification phase is different than in training or enrolment phase [2] [4].

In Early research Shi-Huang Chen and Yu-Ren Luo presents in [5] the MFCC as to extract features and trained and recognized using SVM. They defined the MFCC as the unique and reliable feature extraction technique. That was used to find the most usable features in detailed form. In recognition phase SVM (super Vector Machine) technique based on two class classifiers by defining the decision in binary form was introduced. It discriminate claimed speaker and imposter by +1 and -1 by maximizing the margins or minimizing the structural risks. It shows results averaged to 95.1% with ERR of 0.0%. That was considered as the best results under 22nd order of MFCC. In another research [6], Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi also extracted the voice features using MFCC that was trained and recognized using DTW. DTW (dynamic time Wrapping) a non linear sequence alignment is another technique that is used for recognition process. They find it best for time sequence between two speeches. Here the optimal wrapping path is achieved by wrapping the time distance between two signals. One other research [7] has done by Ibrahim Patel and Dr. Y. Srinivas Rao. They represent the voice recognition with improvement of MFCC with frequency decomposition technique. They introduces sub band coding in their research. The integration of MFCC with sub band coding increases its efficincy and accurate classification as compared to MFCC separately. These two features of MFCC and integrated sub band decomposition with MFCC are used in HMM to train and recognize the speaker.

This paper is attentive of providing a security by developing a voice recognition system to secure the ATM (automatic transaction machine) using HMM with MFCC. The usage of MFCC for extracting voice features and HMM for recognition provides a 2D security to the ATM in real time scenario. MFCC is used to describe the acoustic features of speaker's voice. HMM forward backward estimation technique is used to train these features into the HMM parameters and used to find the log likelihood of entire voice. In recognition HMM is used to compare log likelihood to the pre-stored value and intended to recognize the speaker. If the log likelihood is matched then it is granted otherwise failed to use ATM system. The rest of the paper is alienated as. Section 2 demonstrates the features to be extracted through MFCC. Section 3 describes the use of HMM. Section 4 demonstrates the methodology of being using these techniques. Section 5 consists of the experimental results. Finally section 6 contains the principle conclusion.

2. Mel Frequency Cepstral Coefficients

MFCC is used to extract the unique features of human voice. It represents the short term power spectrum of human voice. It is used to calculate the coefficients that represent the frequency Cepstral these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale of frequency. In Mel scale the frequency bands are equally spaced that approximates the human voice more accurate. Equation (1) is used to convert the normal frequency to the Mel scale the formula is used as

$$m=2595 \log_{10} (1+f/ 700) \qquad (1)$$

Mel scale and normal frequency scale is referenced by defining the pitch of 1000 Mel to a 1000 Hz tones, 40 db above the listener's threshold. Mel frequency are equally spaced on the Mel scale and are applied to linear space filters below 1000 Hz to linearized the Mel scale values and logarithmically spaced filter above 1000 Hz to find the log power of Mel scaled signal [8] [9]. Mel frequency wrapping is the better representation of voice. Voice features are represented in MFCC by dividing the voice signal into frames and windowing them then taking the Fourier transform of a windowing signal. Mel scale frequencies are obtained by applying the Mel filter or triangular band pass filter to the transformed signal. Finally transformation to the discrete form by applying DCT presents the Mel Cepstral Coefficients as acoustic features of human voice.

3. Hidden Markov Modeling

HMM is defined as a finite state machine with fix number of states. It is statistical processes to characterize the spectral properties of voice signal. It has two types of probabilities. There should be a set of observation or states and there should be a certain state transitions, which will define that model at the given state in a certain time
In hidden markov model the states are not visible directly they are hidden but the output is visible which is dependent on the states. Output is generated by probability distribution over the states. It gives the information about the sequence of states but the parameters of states are still hidden. HMM can be characterized by following when its observations are discrete:

- N is number of states in given model, these states are hidden in model.
- M is the number of distinct observation symbols correspond to the physical output of the certain model.
- A is a state transition probability distribution defined by NxN matrix as shown in equation (2).

$$A= \{a_{ij}\}$$
$$a_{ij} = p\{q_{t+1} = j | q_t = i\}, \quad 1 \leq i, j \leq N,$$
$$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \leq i \leq N \qquad (2)$$

Where q_t occupies the current state. Transition probabilities should meet the stochastic limitations

- B is observational symbol probability distribution matrix (3) defined by NxM matrix equation comprises

$$b_j(k)= p\{o_t=v_k|q_t=j\}, \quad 1<=j<=N , 1<=k<=M$$
$$\sum_{k=1}^{M} b_j(k) = 1, \quad 1 \leq j \leq N \qquad (3)$$

Where V_k represents the K^{th} observation symbol in the alphabet, and O_t the current parameter vector. It must follow the stochastic limitations

- π is a initial state distribution matrix (4) defined by Nx1.

$$\pi= \{\pi_i\}$$
$$\pi_i = p\{q_1 = i\}, \quad 1 \leq i \leq N \qquad (4)$$

By defining the N, M, A, B, and π, HMM can give the observation sequence for entire model as λ= (A, B, π) which specify the complete parameter set of model [10].

HMM define forward backward estimation algorithm to train its parameters to find log likelihood of voice sample. Segmental k mean algorithm is used to generate the code book of entire features of voice sample.

Forward backward algorithm is used to estimate the unidentified parameters of HMM. It is used to compute the maximum likelihoods and posterior mode estimate for the parameters for HMM in training process. It is also known as Baum Welch algorithm. It computes the $P(X_k \mid o_{1:t})$Posterior marginal or distribution. For all hidden state variables $X_k \in \{X_1, \dots, X_t\}$. By given a set of observations as $o_{1:t} := o_1, \dots, o_t$.

This inference task is commonly known as smoothing [10] [11]. This algorithm uses the concept of dynamic programming to compute the required values for the posterior margins efficiently in two processes first doing the forward estimations and then backward estimation.

Segmental K-mean algorithm is used to clustering the observations into the k partitions. It is the variation of EM (expectation-maximization) algorithm. That is used to determine the k-means of data distributed by Gaussian distribution. Its objective is to minimize (5) the intra-cluster variance or squared error function.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2 \qquad (5)$$

Here, k is the no of clusters and μ define the centroids of mean point of all points of the input vector. K-mean algorithm is used to first partition the input vector into k initial sets by random selection or by using heuristic data. It defines two steps to precede k-mean algorithm. Each observation is assigned to the cluster with the closest mean. And then calculate the new means to be centriod of observation in each cluster by associating each observation with the closest centroids it construct the new partition , the centroids are recalculated for new cluster until it convergence or observations are no longer remains to clustering .

It converges extremely fast in practice and return the best clustering found by executing several iterations. Its final solution depends on the initial set of clusters [12] [13]. For this, the number of clusters k must be defined to find otherwise it gives eccentric results.

4. Proposed Methodology

The methodology proposed in this research paper consists of two techniques MFCC and HMM. MFCC is used to extract the voice features from the voice sample. And HMM is used to recognize the speaker on the basis of extracted features. For this it first train the extracted features in the format of HMM parameters and to find the log value of the entire voice for recognition. Forward backward estimation technique is used to train the extracted features and find its parameters. This section explains the methodology step by step by explaining (Figure 1) the two techniques MFCC and HMM.

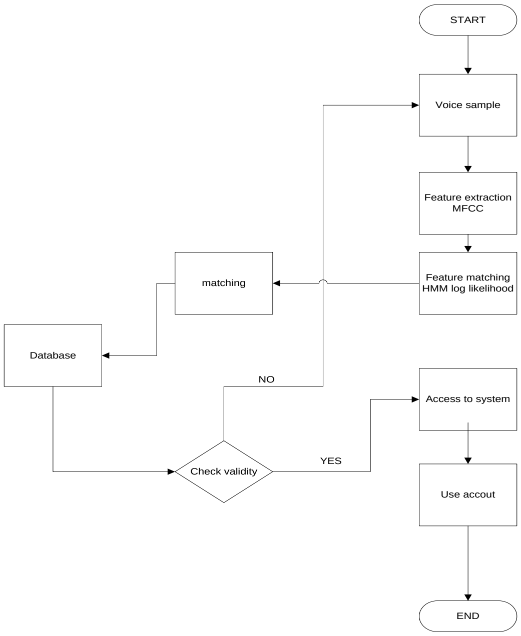


Fig .1 Proposed System

4.1 Voice Processing

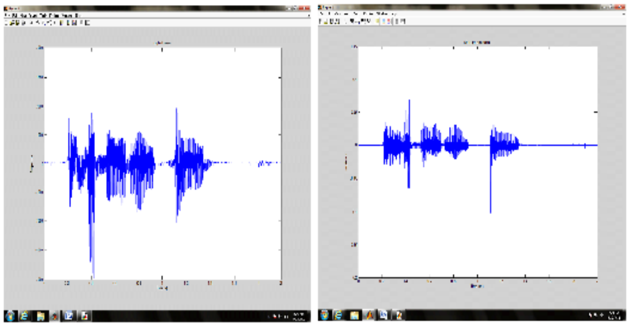
The initial step to develop the proposed research is obtaining the voice sample.Voice sample is taken from the microphone by the speaker. It is digitalized by 8 KHz sampling rate for 2 seconds. Pre-emphasizing the signal make it to normalize. The pre-emphasizing (6) is done to balance the high frequency part of human voice that was covered up when he produce sound. It is also used to increase the high frequency formants in the speech.

$$X2(n) = X (n) - a*X (n-1) \qquad (6)$$

Where the value of a is between 0.9 and 1. Z transform (7) of the filter.

$$H (z) =1-a*z^{1} \qquad (7)$$

This pre-emphasized voice sample (Figure 2) is then stored to a wav file that is used for leading process.



Voice Signal after Pre-emphasized

Fig. 2 voice processing

4.2 Features Extraction

In second step the obtained voice sample is used to frame in MFCC implementation. The pre emphasized voice signal is framed in order to get the stationary part of speech. The speech signal is divided into frames of 30~20 ms with optional overlap of 1/3~1/2 of frame size. With reference to sampling point the frame size usually maintained as the power of two to exploit the FFT. But if it is not then zero padding is done to the nearest length of power of two. Zero padding is used to extend the signal by adding zeros and by increasing its length N to M where M>N. The main purpose of this is to make the signal to the nearest length of power of two and make it feasible for FFT. In this proposed methodology, 256 sample points per frame (Figure 3) and 156 overlap frames are defined.

Framed speech signal is then multiplied with the hamming window in order to remove the discontinuities in the signal. Hamming window returns (8) the symmetric points of integral values framed signal into the column vector w.

$$w(n, \alpha) = (1 - \alpha) - \alpha \cos(2\alpha n/(N-1)) 0 \leqq n \leqq N-1 \qquad (8)$$

Where α shows different curves of hamming window. Its value usually as 0~ 0.5 and window length is L=N. it is obtained (Figure 3) by multiplying each frame to the hamming window.

After windowing FFT is applied to convert (9) the signal into frequency domain from time domain and also used to obtain (Figure 3) the magnitude frequency response of each frame. In doing so it is assumed that the signal in frames in periodic and continuous when wrapping around. In the opposite case of this, there are some discontinuities at the frames start and end points that causes detrimental effects in frequency response. This can be overcome by multiplying each frame by the hamming widow that will help to remove discontinuities at the start and end points of frames.

$$y = fft(b) \qquad (9)$$

Where b is the windowed form of signal.

In Mel filter or triangular band pass filter the magnitude of frequency response is multiplied with the 40 number of triangular band pass filters in order to obtain (Figure 3) the log energy of triangular band pass filter on Mel scale. These filters are equally spaced on the Mel scale (10) and use to calculate the linear frequency.

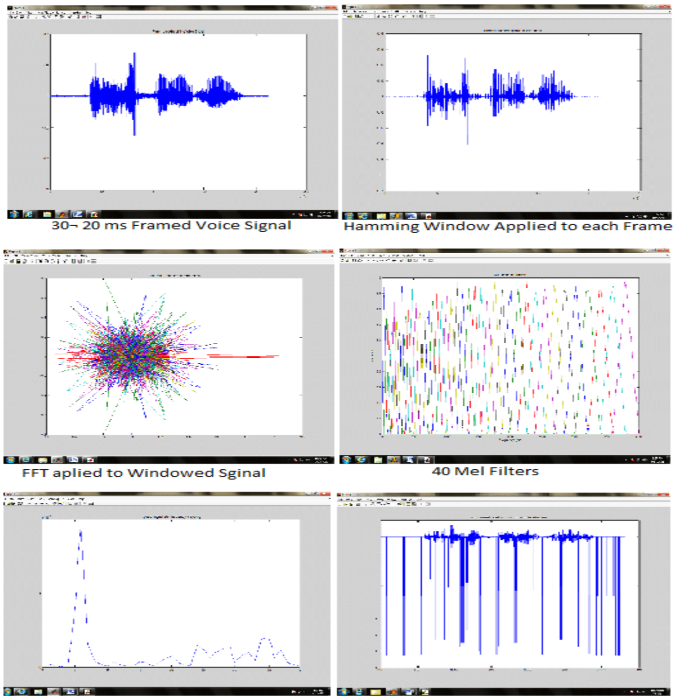
$$m = 2595 \log_{10}(1 + f/700) \qquad (10)$$

The frequency response on Mel scale is reflecting the similar effect of human subjective auditory perception. Triangular band pass filter is used to flatten the magnitude spectrum and to reduce the size of the features occupied. Frequency wrapping (Figure 3) is applied here to keep the useful informational part of the Mel.

At the end by applying DCT cepstral features of voice signal are obtained (11). It is used to convert the log Mel scale cepstrum into time domain from frequency domain (Figure 3).

$$y(k) = w(k) \sum_{n=1}^{N} x(n) \cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right) \quad k = 1, 2, \ldots N \qquad (11)$$

Where N is the length of the computed Mel frequencies. The series starts from n and k=1, because MATLAB vectors starts from 1 instead of 0. The result is known as MFCC. These are the 40 acoustic features of human voice that are used to recognize the person depending upon the filter to be applied.



30~ 20 ms Framed Voice Signal Hamming Window Applied to each Frame

FFT aplied to Windowed Sginal 40 Mel Filters

Frequency Wrapping DCT (Mel Coefficients)

Fig. 3 Features Extraction using MFCC

Log energy is also an important factor of human voice to be recognized. It is computed by obtaining the frame energy of voice signal after framing. It is used to obtain defined number of coefficients of MFCC. Here it calculated the 15 number of coefficients of human voice.

Delta is the first order derivative of original cepstrum. It helps to make speech signal dynamic. There is also 15 coefficients are calculated as delta coefficients. It is the first order derivative of MFCC coefficients.

4.3 HMM Training

For HMM recognition the extracted feature vectors of MFCC are trained into HMM. The training is done in two steps as

- HMM code book
- HMM training by forward backward re-estimation algorithm

First the Code book contains the cluster number specifies to each observation vector, which is obtained by applying the K-mean algorithm. It is used to set the centroids of the observation vector. The observation vectors are represented in the form of matrix Y, and K is the desired number of clusters which are defines as 45; it is used to cluster the featured data Y by random selection. By clustering the model, it returns the centroids, one for each of the cluster k and refers to the cluster number or centroid index of centroid closest to it. K-mean algorithm tries to minimize the distortion that is defined as the sum of squared distances between each observation vector and its dominating centroids. Squared Euclidean distance is ordinary distance between two points which one can measure from ruler. It can be proved by repeated application of Pythagorean formula. In this research study, Euclidean distance (12) is used to find the distance between observation vector and its cluster centroids

$$||Y-Yc|| = sqrt (|| Y ||^2 + || Yc ||^2 - 2* Y. Yc) \qquad (12)$$

After clustering the training to HMM parameters begins by applying Forward-backward algorithm. It uses the principle of Maximum likelihood estimation. It returns the state transition matrix A, observation probability matrix B, and the initial state probability vector π on the basis of defined states as 10 and codebook vectors. In this phase the observation vectors being trained in the form of HMM parameters and resulted as the log likelihood of entire voice. This log likelihood is used to store in speaker's database for recognition in real time.

4.4 HMM Recognition

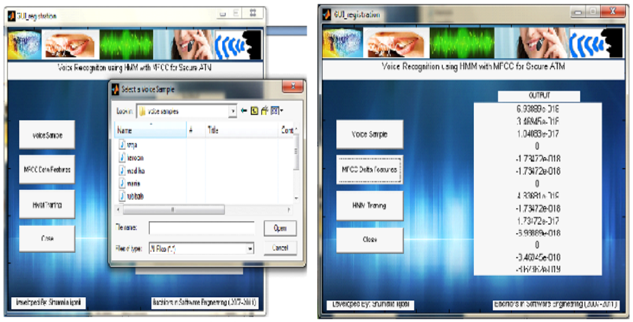
HMM recognition recognized the speaker on the basis of log likelihood. It recalculates the log likelihood of voice vector and compares it to the pre stored value of log likelihood. If it matches the entire log value from the database of specified PIN code then it provides access to the entire speaker to ATM.

5. Experimental Results

In this research, two phases are implemented to obtain experimental results.

- Registration phase
- Recognition phase

In first phase, the static voice sample is used to extract and trained the features and finally stored to the Speaker's database. It stores the entire trained features against the PIN code of specified speaker with his name. Entire features are extracted in 15 MFCC delta coefficients. That represents the pitch of human voice in the form of frequency on Mel scale. Delta coefficients are calculated to these Mel Coefficients and then trained using HMM forward backward algorithm. It results (Figure 4) in the log likelihood of entire voice and used to store in Speaker's database.


GUI_registration
Voice Recognition using HMM with MFCC for Secure ATM
voiceSample
MFCC Delta Features
HMM Training
Close
Select a Voice Sample
Look in: voice samples
Name
Title
smja
keroon
wadiha
maria
rubbab
File name:
Files of type: All Files (*.*)
Open
Cancel
Developed By: Shumaila Iqbal
Bachlors in Software Engineering (2007-2011)
GUI_registration
Voice Recognition using HMM with MFCC for Secure ATM
Voice Sample
MFCC Delta Features
HMM Training
Close
OUTPUT
Developed By: Shumaila Iqbal
Bachlors in Software Engineering (2007-2011)

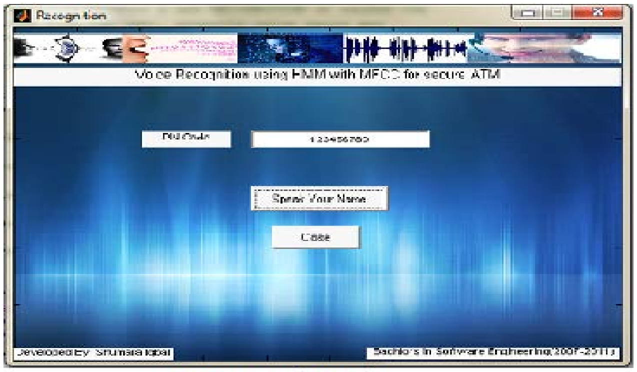
```
6.93889e-018
3.46945e-018
1.04083e-017
0
-1.73472e-018
-1.73472e-018
0
4.33681e-019
-1.73472e-018
1.73472e-017
-6.93889e-018
0
-3.46945e-018
-8.67362e-019
```

Pre-stored/Static voice Sample (GUI)

Features Extraction and Training (GUI)

Fig. 4 Registration phase

Recognition phase works in real time scenario on application ATM. It resulted (Figure 5) to access the user account after verifying the PIN code as well as the log likelihood to the pre stored value against the entire speaker in Speaker's database.


Recognition
Voice Recognition using HMM with MFCC for secure ATM
PIN Code
123456789
Speak Your Name
Close
Developed By: Shumaila Iqbal
Bachlor's in Software Engineering(2007-2011)

Recognition (ATM)

Fig. 5 Recognition phase (ATM)

The results are tested against the specified objectives of proposed system. The developed system is tested by taking 3 speech samples from each speaker with the sampling frequency of 8 KHz. Voice features were extracted from

30 ms frame duration and 20 ms overlapping with the previous frame. The speech sample of 2 sec with noise is used to extract features and then trained using HMM (forward-backward) algorithm. The proposed MFCC features are used to expect the high accuracy in extracting the vocal features of voice. The HMM algorithm is anticipate to get best results in identification system. Accuracy rate shows the percentage of correctly identified test samples by the system. It is obtained by

Accuracy= number of correctly identified test samples/ total number of test samples = 26/ 30*100= 86.67%

Proposed system shows the accuracy of 86.67% as here total 30 test samples of each 10 speakers are being used to identify, where 26 out of total test samples are being correctly identified and correctly rejected with noise factor.

Table 1: Speaker's verification results

Number of speakers	**Correct acceptance (with noise)**	**Correct rejection (with noise)**	**False acceptance (with noise)**	**False rejection (with noise)**
Speaker 1				
Sample 1	√			
Sample 2	√			
Sample 3	√			
Speaker 2				
Sample 1	√			
Sample 2		√		
Sample 3	√			
Speaker 3				
Sample 1		√		
Sample 2	√			
Sample 3			√	
Speaker 4				
Sample 1		√		
Sample 2			√	
Sample 3	√			
Speaker 5				
Sample 1	√			
Sample 2	√			
Sample 3	√			
Speaker 6				
Sample 1				√
Sample 2	√			
Sample 3		√		
Speaker 7				
Sample 1		√		
Sample 2	√			
Sample 3			√	
Speaker 8				
Sample 1	√			
Sample 2	√			
Sample 3		√		
Speaker 9				
Sample 1	√			
Sample 2		√		
Sample 3		√		
Speaker 10				
Sample 1	√			
Sample 2	√			
Sample 3	√			

The error rate is calculated by

Total error of verification system= false accepted+ false rejected =3+1=4/30*100=13.33%

The error rate of total test samples that are being false rejected or accepted is 13.33%

The testing phase shows the efficiency of proposed system up to 86.67% with the error rate of 13.33% as depicted by the test results given in Table 1.

6. Conclusion

It is concluded that the proposed research uses the technique of MFCC to extract unique and reliable human voice feature pitch in the form of Mel frequency and trained and recognized using HMM log likelihood methodology. It comprises two security measures, PIN code as well as voice features to give more security to the ATM application. It represents the best efficiency up to 86.67% with the error rate of 13.33% on the basis of 30 test samples of 10 speakers (3 test samples per speaker).

The future work to leading this system is to provide the secure transmission of voice database to other branches of entire bank. Introducing encryption and decryption to transmit voice database and providing the facility to user to access his ATM account from any branch of bank. By taking the physiological features with behavioral features for recognition of person with voice as well as with his physical movement of mouth to make ATM more secure can tends to be a new research work.

References

[1] Debnath Bhattacharyya, Rahul Ranjan, Farkhod Alisherov A. and Minkyu Choi, "Biometric Authentication: A Review", International Journal of u- and e- Service, Science and Technology Vol. 2, No. 3, September, 2009

[2] Judith A. Markowitz, "Voice Biometrics", September 2000/Vol. 43, No. 9 Communications of the ACM

[3] http://en.wikipedia.org/wiki/Speaker_recognition

[4]http://www.globalsecurity.org/security/systems/biometrics-voice.htm

[5]Shi-Huang Chen and Yu-Ren Luo," Speaker Verification Using MFCC and Support Vector Machine", Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol I, IMECS 2009, March 2009

[6]Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" Journal Of Computing, Volume 2, Issue 3, March 2010

[7]Ibrahim Patel and Dr. Y. Srinivas Rao, "Speech Recognition Using Hmm With Mfcc- An Analysis Using Frequency Specral Decomposion Technique", an International Journal (SIPIJ) Vol.1, No.2, December 2010

[8] Anjali Bala, Abhijeet Kumar and Nidhika Birla, "Voice Command Recognition System Based On MFCC and DTW"

Anjali Bala et al. / International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7335-7342

[9]http://en.wikipedia.org/wiki/Mel_scale

[10] Lawrence R. Rabiner, Fellow, IEEE 'A Tutorial On Hidden Markov Model And Selected Applications In Speech Recognition, Proceedings Of The IEEE, Vol. 77, No. 2, February 1989

[11]http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/HiddenMarkovModels.html

[12]http://algoservice.com/qt/data/used-algoriyhms-reference.pdf

[13] Ms. G. Nathiya, Mrs. S. C. Punitha and Dr. M. Punithavalli," An Analytical Study on Behavior of Clusters Using K Means, EM and K*Means Algorithm", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, March 2010

Shumaila Iqbal She is a graduate student of Software engineering from Fatima Jinnah Women University. She Participated in Quiz Competition held by Sidra Tabassum, Chair IEEE Student Chapter 09.

Tahira Mahboob

She is a registered Engineer with the Pakistan Engineering Council. She received her bachelors degree from University of Engineering & Technology, Lahore in 2007. Currently enrolled in the MS/MPhil Computer Engineering program at Center for Advanced Studies in Engineering CASE(2010), UET Taxila. She has industry experience in telecom sector (Intelligent networks & VAS) Currently serving as a lecturer at Fatima Jinnah Women University, Rawalpindi. She has supervised thesis/projects at Bachelors and Masters Degree Programs in mobile/computer communications, voice recognition, mobile automation and cloud computing. Her area of interests are computer/mobile communications & networks, information security, mobile automation and adhoc/ sensor networks.

Malik Sikandar Hayat Khiyal

Dr. **Malik Sikandar Hayat Khiyal** is Chairperson Department of Computer Sciences and Software Engineering at Fatima Jinnah Women University, Pakistan. He received his M.Sc degree from Quaid-e-Azam University, Islamabad. He got first position in the faculty of Natural Science of the University. He was awarded the merit scholarship for Ph.D. He received his Ph.D. degree from UMIST, Manchester, U.K. He developed software of underground flow and advanced fluid dynamic techniques. His areas of interest are Numerical Analysis, Modeling and Simulation, Discrete structure, Data structure, Analysis of Algorithm, Theory of Automata and Theory of Computation. He has more than hundred research publications in National and International Journals and Conference proceedings.