Face and Speech Recognition Fusion in Personal Identification

Ibiyemi T.S
Dept. of Electrical Engineering
University of Ilorin,
Ilorin, Nigeria

Aliu S.A
Dept. of Electrical Engineering
University of Ilorin,
Ilorin, Nigeria

Akintola A.G.
Dept. of Computer Science
University of Ilorin,
Ilorin, Nigeria

ABSTRACT
Security personnel manning access points most often based their access authorisation on recognition of faces. And it is also very common to base access decision for a person knocking at the door on recognition of his/her voice. The conventional manual method of drawing attention by knocking the door, pressing door bell, or/and shouting one's presence in order to gain access are inefficient and risky. A better method is automatic personal identification based on face and speech recognition which is the subject of this paper. Eigenface method is used for the face recognition. While the speaker and spoken command recognition are both based on the same mel-frequency cepstral coefficients as feature vectors extracted from English and Yorùbá utterances. Experiments yielded 90% face recognition while recognition rates for speakers and the spoken commands were 87% and 74% for utterances in English and Yorùbá respectively. However, the final recognition decision for authorisation or access activation is based on the recognition outcomes of the face recognition and the speech recognition.

General Terms
Biometric Pattern Recognition

Keywords
Face Recognition, Speaker and Speech Recognition, English and Yorùbá utterances, Access control.

1. INTRODUCTION
Face and speech are the two most popular means of personal identification particularly when the person to be identified is physically present. Hence, these two modes become handy in machine effected personal identification using biometrics for automatic access control [1,2]. Security personnel manning access points most often based their access authorisation on recognition of faces. And it is also very common to base access decision for a person knocking at the door on recognition of his/her voice. Human face does not only provide information on personal identification but also information on gender, race, and age group. Speech conveys information on "what is said" and in addition information on the speaker identity, gender, accent, and age group. The conventional manual method of drawing attention by knocking the door, pressing door bell, or/and shouting one's presence in order to gain access are inefficient and risky. Access authorisation by this method can easily lead to play into the hands of armed intruders. But using biometric automates the process of access authorisation and eliminates risk of inadvertently opening the door for armed intruders.

In this paper, the development of personal identification based on face and speech recognition is presented. One problem with face recognition by human is that people find it relatively easier to recognise faces of their own race than other races. But face recognition by machine eliminates the problem of racial subjectivity. People find it very easy to recognise other familiar people by the voice even if they are out of sight. This project exploits the machine ability to recognise human face and the possibility of porting man's ability to recognise people by their voices to machine.

The face recognition part of this work is based on the principle of principal component analysis, PCA, also known as eigenface [1 ,3,4]. Face recognition by eigenface method is becoming very popular because of its effectiveness and modest computational requirement [4,5,6]. However, face recognition rate using eigenface is easily degraded by variation in illumination conditions, inconsistency in size of acquired face image, and rotation in face. Hence, measures are put in place in this work to handle these sources of degradations. The speech recognition aspect includes recognition of what is said and the speaker that said it. The method used is primarily based on using mel frequency cepstral coefficients as feature vectors fed into vector quantisation algorithm and simple Euclidean distance measure for recognition. This work is particularly useful in door/gate access control since not only that intruders will be denied access but they presence will cause an alert.

2. FACE RECOGNITION
The PCA, which is similar to Karhunen-Loeve's linear transformation, maps a high dimensional image space to a lower dimensional face image sub-space [6]. The basis vectors of this face image sub-space correspond to directions of maximum variances in the original image space. These basis vectors are the eigenvectors of the face image covariance matrix and also known as eigenfaces which are

face-like when displayed. The problem can be stated as: For a given set of training face images, extract and store the most dominant eigenfaces as templates then compute the eigenface of a query face image and compare it, using distance metric, with the templates for recognition decision. Therefore, face recognition algorithm based on the eigenface method can be described as:

For a given K training colour face images, each of size M x N:

Convert each RGB colour image to grey scale:

$$\Gamma_{i,j} = \sum_{l=1}^{3} \begin{bmatrix} 0.299 \\ 0.587 \\ 0.114 \end{bmatrix}^T .G_{i,j,l} \quad ,i = 1,2,\cdots,M \; ; \quad j = 1,2,\cdots,N \tag{1}$$

$where$:

$\Gamma_{i,j} \Rightarrow grey\ scale\ image$

$G_{i,j,l} \Rightarrow Colour\ RGB\ image$

Create training image matrix (M.N x K) with each column representing a vectorised image of length , M.N :

$$F = \left[\Gamma_1, \Gamma_2, \cdots\cdots\cdots, \Gamma_K\right] \tag{2}$$

$where$:

$$\Gamma_i = \left[x_0, x_1, \cdots\cdots\cdots, x_{M.N-1}\right]^T$$

Perform illumination normalisation:

$$F' = (F_i - \mu_0)\frac{\sigma_i}{\sigma_0} + \mu_i \quad ,i = 1,2,\cdots,K \tag{3}$$

$where$:

$$\mu_i = \frac{1}{M.N}\sum_{j=0}^{M.N-1} x_j \quad , \sigma_i = \sqrt{\left(\frac{1}{M.N}\sum_{j=0}^{M.N-1}(x_j - \mu_i)^2\right)}$$

$\mu_0 \Rightarrow desired\ mean, typically = 100$

$\sigma_0 \Rightarrow desired\ standard\ deviation, typically = 100$

Obtain zero-mean training image matrix:

$$\Phi_i = F_i' - \Psi \qquad ,i = 1,2,\cdots,K \tag{4}$$

$where$:

$$\Psi = \frac{1}{K}\sum_{i=1}^{K} F_i'$$

$$A = \left[\Phi_1, \Phi_2, \cdots, \Phi_K\right]$$

Obtain covariance matrix:

$$C = A.A^T \tag{5}$$

$where\ the\ dimension\ of\ C\ is$:

$$C = (M.N.K)x\ (M.N.K) \Rightarrow (M.N)x(M.N) \quad matrix$$

$$\therefore \quad L = A^T.A$$

$where$:

$$L \Rightarrow K\ x\ K\ matrix$$

Calculate eigenvalues and eigenvectors:

The covariance matrix, C, formed by outer product of AA^T has a very high dimension of (M.N) by (M.N) hence making direct computation of its eigenvalues and corresponding eigenvectors intractable. However, the inner product of $L = A^T A$ has a more manageable dimension of (K.K)

From which the eigenvalues and eigenvectors are indirectly computed. Hence, it suffices to obtain K eigenvalues and corresponding K eigenvectors of length K each; and then extrapolate the K-length eigenvectors to full length (M.N) eigenvectors.

Let the eigenvectors of reduced matrix L be $\nu_i \ ,i = 1,2,\cdots,K$, and from the eigenvectors of the reduced matrix, the eigenvectors of large matrix C can be extrapolated as:

$$U_i = \sum_{j=1}^{K} \nu_{i,j}\Phi_j \quad ,i = 1,2,\cdots,K \tag{6}$$

These eigenvectors, U_i , are like the face images hence they are called eigenfaces.

Each of the eigenface has varying significance depending on the magnitude of its eigenvalue. Hence, it suffices to select a subset K' of the K eigenfaces corresponding to K' highest valued eigenvalues as characterising the entire training face images.

These reduced subset K' eigenfaces are stored , in addition to the mean face.

And then calculate weight vectors by projecting training face images onto the stored eigenfaces.

The contribution of a stored eigenface to a zero-mean training face image can be calculated as a scalar weight. Therefore, a weight vector of length K' whose elements represent the degree of contribution of the corresponding eigenface to that zero-mean image is obtained by projection:

$$\omega_j = U_j^T.\Phi_j \quad , j = 1,2,\cdots,K' \tag{7}$$

$$\therefore \quad \Omega_i = \left[\omega_1, \omega_2, \cdots\cdots, \omega_{K'}\right] \quad ,i = 1,2,\cdots\cdots,K'$$

The calculated $\left(K'\,x\,K'\right)$ weight matrix is stored as reference templates.

The recognition phase, i.e. recognition of a query face, process involves conversion of the colour query image to grey image, normalisation of the converted image, and subtraction of the stored mean of the training images from the query image. Then calculate weight vector for the new image:

$$\Omega_{query} = U_i.\left(F'_{query} - \Psi\right) \quad ,i = 1,2,\cdots,K' \tag{8}$$

Finally, perform classification by matching using Euclidean distance metric between the query and the templates:

$$\varepsilon_{face} = \arg\min_i \left\| \Omega_{query} - \Omega_i \right\|_2 \quad , i = 1,2,\cdots\cdots, K' \qquad (9)$$
$$if\ \varepsilon_{face} < T_{face}\ then\ F_{query}\ recognised\ as\ i-th\ training\ \ image\ else\ unknown$$

3. SPEECH RECOGNITION

The speech recognition tasks in this work are twofold, namely, recognition of the speaker, and recognition of the uttered word. Fortunately, a speech data contains information on both hence in one process both the speaker and the uttered word can be simultaneously recognised. The algorithm used for the speech recognition consists of pre-processing, feature extraction, and recognition modules [2,7,8].

The pre-processing module is made up of word segmentation by endpoints detection; pre-emphasis filtering, frame blocking, and application of smooth window to speech word data.

The frame energy and zero crossing rate, ZCR, of a word is used in conjunction with thresholds to segment the activity area of the word from the silent background. The frame energy and ZCR are determined by eqns(10 and 11).

$$Let\ s_i\ , i = 1,2,\ldots, N\ be\ a\ frame\ speech\ samples;\ N \Rightarrow no.of\ samples \qquad (10)$$
$$E = \frac{1}{N}\sum_{i=1}^{N} s_i^2$$

And the frame ZCR calculated as:

$$z = \frac{1}{N}\sum_{j=2}^{N} \frac{\left|\mathrm{sgn}(s_j) - \mathrm{sgn}(s_{j-1})\right|}{2} \qquad (11)$$
$$where:$$
$$\mathrm{sgn}(s()) = \begin{cases} +1 & , s() \geq 0 \\ -1 & , s() < 0 \end{cases}$$

The pre-emphasis filter is implemented as high pass FIR filter with transfer function of:

$$H(z) = 1 + az^{-1}$$
$$where:\ a = 0.95 \qquad \textbf{(12a)}$$

The time response of this filter is:

$$y(n) = s(n) - s(n-1) \qquad (12b)$$

For the frame blocking, the word samples are partitioned into short blocks in order to make the signal stationary as in eqn.13:

$$x_{i,j} = y((K-M).i + j)\ ,\ j = 0,1,\cdots, K-1; i = 0,1,\cdots, L-1; \qquad (13)$$
$$where:$$
$$N \Rightarrow no.\ of\ samples\ per\ word$$
$$L = \left(\frac{N-M}{K-M}\right) \Rightarrow no.of\ \ frames\ per\ word$$
$$M \Rightarrow no.of\ overlapped\ samples\ per\ \ frames$$
$$K \Rightarrow no.of\ \ samples\ \ per\ \ frame$$

Frame blocking using rectangular window as defined in eqn (13) leads to ringing frequency response also known as Gibb's phenomenon as a result of sharp edges. Hence, Hamming window which is a raised cosine window is used to smooth the edges. Hamming window is defined by eqn (14):

$$y(n) = x(n).w(n)$$
$$where:$$
$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad ,0 \leq n \leq N-1 \qquad (14)$$
$$w(n) \Rightarrow Hanm\min g\ window$$
$$x(n) \Rightarrow speech\ \ frame\ samples$$

The characterisation of speech signal data for the purpose of simultaneous recognition of what is said and who said it, that is, speech recognition and speaker verification respectively is most efficiently handled by mel frequency cepstral coefficients, MFCC. This is because MFCC characterisation is very similar to that of the human aural perception. Hence, MFCCs are used as feature vectors in this work for simultaneous speaker authentication and speech recognition [2,9,10,11].

The process of obtaining the MFCCs involves transformation of the windowed frame of speech data from time domain to frequency domain and then back to time domain after processing. Firstly, the spectrum magnitude of the windowed speech signal data is obtained on a linear scale frequency by FFT. This output magnitude is converted power magnitude which is convolved with the frequency response magnitude output of filter bank on mel-frequency scale. In order to convert the obtained mel-scaled power spectrum to time domain, an the inverse discrete cosine transform, DCT, is taken. The output of the inverse DCT are the mel frequency cepstral coefficients . The algorithm for obtaining the MFCCs is described in eqn(15) to eqn(19).

Apply DFT to each of the windowed speech signal of eqn(8):

$$Y(n) = \sum_{k=0}^{N-1} w(k)x(k)e^{-j2\pi kn/N}, \quad n = 0,1,\cdots, N-1 \qquad (15)$$

Get the power spectrum of eqn(15):

$$Y'(n) = \left(\sqrt{\left((Y'_{real}(n)^2) + (Y'_{imag}(n)^2\right)}\right)^2 \quad , n = 0,1,\cdots, N-1 \qquad (16)$$
$$= \left(Y'_{real}(n)^2 + Y'_{imag}(n)^2\right)$$

Convert power spectrum of eqn(16) in linear frequency to power spectrum in mel-scale frequency:

$$P_{mel}(m) = \sum_{k=0}^{N-1} Y'(k).H(k,m) \quad ,m = 0,1,\cdots,L$$

where:

$L \Rightarrow$ *no. of mel filters*

$$H(k,m) = \begin{cases} 0 & ,f(k) < f_c'(m-1) \\ \dfrac{f(k) - f_c'(m-1)}{f_c'(m) - f_c'(m-1)} & ,f_c'(m-1) \le f(k) < f_c'(m) \\ \dfrac{f(k) - f_c'(m+1)}{f_c'(m) - f_c'(m+1)} & ,f_c'(m) \le f(k) < f_c'(m+1) \\ 0 & ,f(k) \ge f_c'(m+1) \end{cases} \qquad (17a)$$

$$f(k) = f_{\min} + (k-1).\Delta f \quad ,k = 1,2,\cdots,N$$

where:

$$\Delta f = \frac{f_{\max} - f_{\min}}{N-1} \qquad (17b)$$

$f_{\min} \Rightarrow$ *Minimum speech frequency*

$f_{\max} \Rightarrow$ *Maximum speech frequency*

$$f_c'(m) = f_{\min}' + (m-1).\Delta f' \quad ,m = 1,2,\cdots,L$$

where:

$$f'(m) = 2595 \log_{10}\left(\frac{f(m)}{700} + 1\right) \quad ,m = 1,2,\cdots,L$$

$$\Delta f' = \frac{f_{\max}' - f_{\min}'}{L-1}$$

$f_{\min}' \Rightarrow$ *Minimum mel frequency*

$f_{\max}' \Rightarrow$ *Maximum mel frequency*

$f \Rightarrow$ *speech frequency*

$f' \Rightarrow$ *mel speech frequency*

(17c)

Obtain the logarithm of the mel scale power spectrum:

$$P'_{mel}(m) = \log_{10}(P_{mel}(m)), \quad m = 1,2,\cdots,L \qquad (18)$$

Convert logarithm mel frequency power spectrum to time domain cepstral coefficients (mfcc) using inverse DCT:

$$C_i = \sum_{j=1}^{L} P'_{mel}(j) \cos(\pi.i(j-0.5)/M) \quad ,i = 1,2,\cdots,M \qquad (19)$$

where: $M \Rightarrow$ *no. of coefficients* , $L \Rightarrow$ *no. of mel – filters*

It is typical in speaker and speech recognition involving multiple utterances of words to generate very large number of feature vectors per word during the training phase. The total number of feature vectors can easily become unmanageable in term of storage as templates or in matching computation. These two problems can render speech recognition in embedded application unrealisable. This makes the use of vector quantisation imperative as data compression method. [11,12,13] . The VQ problem can be defined as:

Given: $T = \{X_1, X_2, \cdots\cdots, X_N\}$ *feature vectors*; *& no.of desired codevectors M*

Find $C = \{c_1, c_2, \cdots\cdots, c_M\}$ *codevectors & the codevectors's partition regions*

$P = \{s_1, s_2, \cdots, s_M\}$ *such that average distortion* D_{ave} *is* min *imised*

This problem is solved, in our case, using LBG-VQ algorithm of Fig. 1.

_start_LBG_VQalgorithm

{

step0: Codebook Initialisation

-Input N, M, X_i = { $x_{i,1}$, $x_{i,2}$,…..,$x_{i,k}$} , i=1,2...,N ; k=dimension of feature vector

(N→total no. feature vectors in training set; X → training set feature vectors*

*M→no. of codevectors/vocabulary words *)*

-Calculate 1-codevector codebook:

$$c_1 = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (20a)$$

-Set $\varepsilon = 0.01$

-Set m = 1

-Set $n(h) = 0 \quad ,h = 1,2,\cdots,M$

-Calculate distortion, D:

$$D = \frac{1}{N.k}\sum_{i=1}^{N}\sum_{j=1}^{k}\left\| x_{i,j} - c_{1,j} \right\|^2$$

step1: Double Codebook Size by Splitting

$$\begin{aligned} &\text{for } j = 1 \text{ to } m \text{ do} \\ &(\\ &\quad c_i' = (1+\varepsilon).c_i \\ &\quad c'_{m+i} = (1-\varepsilon).c_i \\ &) \\ &\quad m = 2m \\ &\quad c_i = c_i' \quad ,i = 1,2,\cdots,m \end{aligned} \qquad (20b)$$

step2: Distribute feature vectors by clustering

-Set $D' = D$

$$\begin{aligned}&for\ i=1\ to\ N\ do\\&(\ for\ \ j=1\ to\ M\ \ do\\&\quad(\ j^{*}=\arg\min_{j}\left(\left\|x_i-c_j\right\|^{2}\right))\\&\quad s_{j^{*}}=x_i\ ;\\&\ n(j^{*})=n(j^{*})+1\\&)\end{aligned} \qquad (20c)$$

step3: Update Centroids/Codevectors

$$c_j=\frac{1}{n_j}\sum_{h=1}^{n_j}s_j(h)\quad ,j=1,2,\cdots,M \qquad (20d)$$

step4: Calculate new Distortion

$$D'=\frac{1}{N.M.k}\sum_{i=1}^{N}\sum_{h=1}^{M}\sum_{j=1}^{k}\left\|x_{i,j}-c_{h,j}\right\|^{2} \qquad (20e)$$

$$if\ \left(\frac{D-D'}{D}>\varepsilon\right)Then\ goto\ step2\ otherwise\ goto\ step5$$

step5: Repeat until desired number of codewords

$$if\ \ (m<M)\ Then\ goto\ step1\ otherwise\ step6$$

step6: Output codebook

$$Output\quad c_j\ ,\ j=1,2,\cdots\cdots,M$$

step7: stop

}*_end_LBG_VQalgorithm.*

Fig.1 LBG VQ Process

The computed code-vectors are stored in codebook as feature vector templates but indexed as scalar values.

In the operational mode, a query word passes through the same process as the training phase except that instead of the storing code-vectors, they are matched with those in template database to determine the closest match. A simple Euclidean distance metric is used, and recognition decision is taken by comparing the closest match with a threshold as defined in eqn(21).

$$\begin{aligned}&\varepsilon_{speech}=\arg\min_{i}\left\|C_{query}-C_i\right\|_2\quad ,i=1,2,\cdots\cdots,M\\&if\ \varepsilon_{speech}<T_{speech}\ then\ W_{query}\ recognised\ as\ i-th\ spoken\ by\ i-th\ spea\ker\\&else\ unknown\end{aligned} \qquad (21)$$

4. RECOGNITION FUSION

The final recognition decision for authorisation or access activation is based on the recognition outcomes of the face recognition and the speech recognition. If the person requesting service is recognised by the face recognition process and the same person is recognised as the speaker of the command word by the speech recognition module then the recognised command is effected. But if the same person is not recognised by the two biometric modes authorisation is denied.

5. EXPERIMENT AND RESULT

Table I shows the 11 possible command phrases in English and Yorùbá. These phrases are formed from 11 English words (Open, Close, On, Off, Door, Gate, Light, Generator, Security, Siren, Alert) and the equivalent 11 isolated Yorùbá words (Sì, Tì , Tán, Pá, Lẹkun, Lẹkun-nlà, Inà, Ẹro-Inà, Ẹro-Férè, Pè, Àsó-Òdé).

Table I: The 11 possible Command Phrases in English and Yorùbá (in parentheses)

Open Door (Sì Lẹkun)	Open Gate (Sì Lẹkun-là)	Close Door (Tì Lẹkun)	Close Gate (TìLẹkun-Nlà)
On Light (Tán Inà)	On Generator (Tàn Ẹro-Inà)	Off Light (Pá Inà)	Off Generator (Pá Ẹro-Inà)
On Siren (Tán Ẹro-Férè,)	Off Siren (Pá Ẹro-Férè)	Alert Security (Pè Àsó-Odé)	

The "Alert Security" command phrase auto-dials pre-defined telephone numbers. Experiments were conducted using 50 speakers to pronounce each of the 11 phrases in the vocabulary 10 times.

Half of the 5500 phrases were used for training while the other half were used for testing the operational performance of the developed system. Each phrase was sampled at 8 kHz and quantised into 16 bits per sample.

The pre-emphasis filter coefficient used was 0.97 at a framed window of 256 samples with 128 samples overlap. The mel frequency cepstral coefficients of 16 dimension derived from 20 mel filter bank constituted a feature vector.

There are 50 codebooks, with one codebook for each speaker. Each codebook has 11 codebook-lets, each codebooklet represents each phrase of the vocabulary. A codebooklet contains 10 codevectors representing the 10 utterances per phrase per speaker. The codebooks, codebooklets, and codevectors are appropriately populated during training. An adaptive threshold is determined for each of the codevector during training. A simple Euclidean distance measure is used in matching a test phrase utterance with the templates in the codebooks. The computed distance is compared with the stored thresholds in determining the speaker and recognition of the spoken command for granting access. Both the face recognition and speech recognition algorithms were coded in C language and ran on a Pentium duo core 2.6 GHz with 2GB

RAM on board. The experiments yielded 90% face recognition. The recognition rates for speakers and the spoken commands were 87% and 74% for utterances in English and Yorùbá respectively.

6. CONCLUSION

The development of personal identification based on face and speech recognition was presented. This bi-modal biometrics was utilised for access control. Eigenface method was used for the face recognition while the speaker and spoken command recognition were both based on the same mel-frequency cepstral coefficients as feature vectors. Experiments yielded 90% face recognition. The recognition rates for speakers and the spoken commands were 87% and 74% for utterances in English and Yorùbá respectively.

7. ACKNOWLEDGEMENT

We acknowledge with great appreciation the generous research and development grant received from Federal Government of Nigeria through the STEP-B project to execute this work.

8. REFERENCES

[1] Ibiyemi T.S., Ogunsakin J., Daramola S.A. 2012. "Bi-Modal Biometric Authentication by Face Recognition and Signature Verification", International Journal of Computer Applications, vol.42, no. 20, pp 17-21.

[2] Ibiyemi T.S., Akintola A.G. 2012. "Speaker Authentication and Speech Recognition Enabled Telephone Auto-Dial in Yorùbá", International Journal of Science and Advanced Technology, vol.12, no.4, pp 88-187.

[3] Ibiyemi T.S., Aliu S.A. 2003. "Automatic Face Recognition by Computer", Abacus: Mathematics Series, vol 30, no. 2B, September, pp180-188

[4] Ibiyemi T.S., Aliu S.A. 2002. "On Computation of Optimum Basis Vector for Face Detection and Recognition", Abacus: Mathematics Series, 29(2), pp 144-149

[5] Brian Harding, Cat Jubinski, "A Standalone Face Recognition Access Control System", ECE4760 Final Project Report,

[6] Turk M., Pentland A. 1991. "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, vol. 3, no.1, pp71-86

URL: http://people.ece.edu/land/courses/ece4760

[7] Lipeika Antanas, Lipeikiene Joana, Telksnys Laimutis. 2002. " Development of Isolated word Speech Recognition", Informatica, vol.13, no.1, 37-46

[8]E-Hocine Bourouba, et al. 2006."Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM", Informatica 30 373-384

[9] Satyahad Singh, and Rajan E.G. 2011. "MFCC VQ based Speaker Recognition and its Accuracy Affecting Factors", International Journal of Computer Applications, vol. 21, no.6, pp.1-6

[10] Rashidul Hasan, Mustafa Jamil, Golam Rabbani, Saifur Rahman. 2004. "Speaker Identification using Mel Frequency Cepstral Coefficients", Proc. 3rd International Conference on Electrical and Computer engineering, ICECE 2004, 28-30 December, Dhaka Bangladesh, pp. 565-568

[11] Srinivasan A.. 2012. "Speaker Identification and Verification using Vector Quantisation and Mel Frequency Cepstral coefficients", Research Journal of Applied Sciences, Engineering and technology", vol.4, no.1, pp. 33-40

[12] Allam Musa. 2011,."MareText Independent Speaker Identification based on K-Means Algorithm", International Journal on Electrical engineering and Informatics, vol.3, no.1, pp100-108

[13] Kekre H.B., and Vaishali Kulkarni. 2010. Performance Comparison of speaker Recognition using Vector Quantization by LBG and KFCG", International Journal of applications, vol.3, no.10, pp.32-37