Text-Dependent Audio Biometric Person Authentication

Ekhlas Falih Naseer *

e-mail: @yahoo.com

Received on: 20/5/2012
Accepted on: 10/9/2012

Abstract: Audio biometric person authentication is the task of verifying the identity of a person based on the information in the speech signal that occurs during the production of speech. In this research, audio person authentication is focus on acoustic text-dependent speaker verification system. The proposed recognition process begins by converting that audio into frequency domain by applying discrete cosine transform (DCT), then compute the seven moments as a features for that audio and build a database depends on these features, then compute the Kohonen neural network for person identification, and then compute the dynamic time wrapping (DTW) for verification and patterns matching, and later give the decision logic for accepting or rejecting a claimant.

Keywords: Audio Biometric Person Authentication; Multi modal Speech Processing; Discrete Cosine Transform, Rules of Moments, Kohonen Network Rules, Dynamic Time Wrapping.

*** Computer Science Department / University of Technology**

1. Introduction

The technique for audio person authentication is developed as an extension over a baseline text dependent acoustic speaker verification system by using the features extracted from the person sound to improve the performance of the system. The acoustic speaker verification exploits the uniqueness in the voice of the speaker. This uniqueness is due to the physiological characteristics of the speech production mechanism. The speech signal captured during the utterance is subjected to preprocessing and feature extraction. The acoustic features used in the baseline system are segmental features that are extracted from 10 to 30 ms of speech signal. The system operates in two modes: enrollment and verification. During enrollment, the features extracted from the speech signal are stored as reference template for each speaker. During verification, the features extracted from the test speech signal are matched with the reference template [1].

2. Preprocessing of Acoustic Speech Using Discrete Cosine Transform (DCT) [1]

Speech signal captured by the microphone is preprocessed before extracting features. This involves detection of begin and end points of the utterance in the speech waveform, pre-emphasis and windowing of the frame. The process of pre-emphasis provides high frequency emphasis and windowing reduces the effect of discontinuity at the ends of each frame of speech. The speech samples in each frame are preprocessed using a difference operator to emphasize the high frequency components. If the speech signal s(n) is represented as a sequence of samples where n = 0, 1,2 ,3,4,…,N–1 , then the pre-emphasized speech is obtained by:

f(n) = s(n) + s(n + 1) ……….(1)

where: f(n) represents the pre-emphasized speech and n is the sample index [1]. Discrete Cosine Transform (DCT) is the basis for many image and video compression algorithms, especially the baseline JPEG and MPEG standards for compression of still and video images respectively [2].
The one-dimensional *forward discrete Cosine transform* (1D FDCT) of *N* samples is formulated by

For u = 0, l, . . . ,*N* - 1 and
v = 0, 1, . . . , *M* - 1, where

$$C(k) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } k = 0 \\ 1 & \text{otherwise.} \end{cases}$$

$$F(u) = \sqrt{\frac{2}{N}} C(u) \sum_{x=0}^{N-1} f(x) \cos\left[\frac{\pi(2x+1)u}{2N}\right] \qquad \textbf{(2)}$$

The function *f (x)* represents the value of the x th sample of the input. The one-dimensional *inverse discrete Cosine transform* (1D IDCT) is for signal. *F (u)* represents a DCT coefficient for *u* = 0, l,.,, N–1 formulated in a similar fashion as follows[2],

$$f(x) = \sqrt{\frac{2}{N}} \sum_{u=0}^{N-1} C(u) F(u) \cos\left[\frac{\pi(2x+1)u}{2N}\right] \qquad \textbf{(3)}$$

For x = 0, l,. . . *N* - 1

3. Authentication

Authentication ensures that users are who are who they claim to be. It also ensures that all processes and transactions are initiated only by authorized end users. User authentication couples the lginid and the password, providing an identifier for the user, a mechanism for assigning access privileges, and an auditing "marker" for the system against which to track all activity, such as file accesses, process initiation, and other actions (e.g., attempted logons). Thus, through the process of authentication, one has the means to control and track the "who" and the "what" [3].

4. Begin and End Points Detection

For a text-dependent person authentication system using the DTW algorithm, the leading and trailing silence regions of a recorded speech utterance have to be removed by detecting the begin and end points of speech. Correct detection of begin and end points of a speech utterance improves the accuracy of aligning the reference with the test utterance. Techniques for begin and end detection of an utterance are generally based on the amplitude of the signal. But these techniques are not robust to ambient noise and reverberation. Since the data used in this work is captured by a microphone in an ordinary live room [4].

5. Acoustic Features Extraction by Computing the Seven Moments

The acoustic features extracted from the speech signal represent the characteristics of the vocal tract system. In the baseline system, *Moment invariant is* to use region-based geometric moments that are invariant to translation and rotation. It identified seven normalized central moments as shape features, which are also scale invariant.

Let F (x, y) denote an image in the two-dimensional spatial domain.

Geometric moment of order p + q is denoted as:

$$m_{p,q} = \sum_{x}\sum_{y} x^{p} y^{q} F(x,y) \qquad \ldots \textbf{(4)}$$

for p , q = 0 , 1 , 2 , . . N. The central moments are expressed as

Where

$$x_c = m_{1,0} / m_{0,0}$$

$$y_c = m_{0,1} / m_{0,0}$$

Where $m_{1,0}$ mentioned in Eq. (4) and *(x*$_c$*,* y$_c$) is called the center of the region of object [5]. Hence the *central moments*, of order up to 3, can be computed as :

$$\mu_{0,0} = m_{0.0}$$

$$\mu_{1,0} = 0$$

$$\mu_{0,1} = 0$$

$$\mu_{2,0} = m_{2,0} - \chi_c m_{1,0}$$

$$\mu_{0,2} = m_{0,2} - \gamma_c m_{0,1}$$

$$\mu_{1,1} = m_{1,1} - \gamma_c m_{1,0}$$

$$\mu_{3,0} = m_{3,0} - 3\chi_c m_{2,0} + 2m_{1,0}\chi_c{}^{2}$$

$$\mu_{1,2} = m_{1,2} - \gamma_c m_{1,1} - \chi_c m_{0,2} + 2\gamma_c{}^{2} m_{1,0}$$

$$\mu_{2,1} = m_{2,1} - 2\chi_c m_{1,1} - \gamma_c m_{2,0} + 2\chi_c{}^{2} m_{0,1}$$

$$\mu_{0,3} = m_{0,3} - 3\gamma_c m_{0,2} + 2\gamma_c{}^{2} m_{0,1}$$

The normalized central moment, denoted $\eta_{p,q}$, is defined as

$$\eta_{p,q} = \mu_{p,q} / \mu^{\gamma}{}_{0,0} \qquad \ldots (5)$$

Where

$$\gamma = p + q/2 \qquad \ldots (6)$$

For p + q = 2, 3, a set of seven *transformation invariant moments* can be derived from the second- and third-order moments as follows [5]

$$\left.\begin{aligned}
&\phi 1=\eta_{2,0}+\eta_{0,2}\\
&\phi 2=(\eta_{2,0}+\eta_{0,2})^2+4\eta_{1,1}\\
&\phi 3=(\eta_{3,0}-3\eta_{1,2})^2+(3\eta_{2,1}-\eta_{0,3})^2\\
&\phi 4=(\eta_{3,0}+3\eta_{1,2})^2+(3\eta_{2,1}+\eta_{0,3})^2\\
&\phi 5=(\eta_{3,0}-3\eta_{1,2})(\eta_{3,0}+3\eta_{1,2})[(\eta_{3,0}+3\eta_{1,2})^2\\
&-3(\eta_{2,1}+\eta_{0,3})^2]+(3\eta_{2,1}-\eta_{0,3})(\eta_{2,1}+\eta_{0,3})\\
&[3(\eta_{3,0}+\eta_{1,2})^2-(\eta_{2,1}+\eta_{0,3})^2]\\
&\phi 6=(\eta_{2,0}+\eta_{0,2})[(\eta_{3,0}+\eta_{1,2})^2-(\eta_{2,1}-\eta_{0,3})^2]\\
&+4\eta_{1,1}(\eta_{3,0}+\eta_{1,2})(\eta_{2,1}-\eta_{0,3})\\
&\phi 7=(3\eta_{2,1}-\eta_{0,3})(\eta_{3,0}+\eta_{1,2})[(\eta_{3,0}+\eta_{1,2})^2\\
&-3(\eta_{2,1}+\eta_{0,3})^2]+(3\eta_{1,2}-\eta_{3,0})(\eta_{2,1}+\eta_{0,3})\\
&[3(\eta_{3,0}+\eta_{1,2})^2-(\eta_{2,1}-\eta_{0,3})^2]
\end{aligned}\right\}\quad \textbf{(7)}$$

6. Kohonen Network for Learning Prototypes

.

Classification of data and the role of prototypes in learning are constant concerns of psychologists, linguists, computer scientists, and cognitive scientists [6]. Kohonen learning has a strong inductive bias in that the number of desired prototypes is explicitly identified at the beginning of the algorithm and then continuously refined; this allows the net algorithm designer to identify a specific number of prototypes to represent the clusters of data [7].

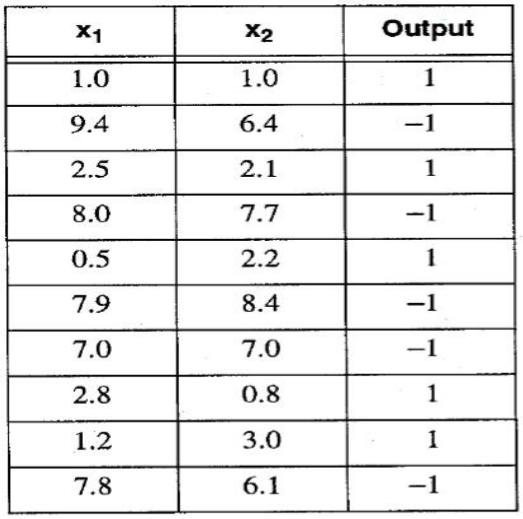
This set of normalized central moments is invariant to translation, rotation, and scale changes in an image.

The proposed system can be entered the values of the seven moments as input to the Kohonen neural network.

Figure (2) is a Kohonen learning network for classification of the data of Table (1). The data are represented in Cartesian two dimensional spaces, so prototypes to represent the data clusters will also be ordered pairs. We select two prototypes, one to represent each data cluster.

We have randomly initialized node A to (7, 2) and node B to (2, 9). Random initialization only works in simple problems such as ours; an alternative is to set the weight vectors equal to representatives of each of the clusters.

Table (1): A classification of data set



x_1	x_2	Output
1.0	1.0	1
9.4	6.4	−1
2.5	2.1	1
8.0	7.7	−1
0.5	2.2	1
7.9	8.4	−1
7.0	7.0	−1
2.8	0.8	1
1.2	3.0	1
7.8	6.1	−1

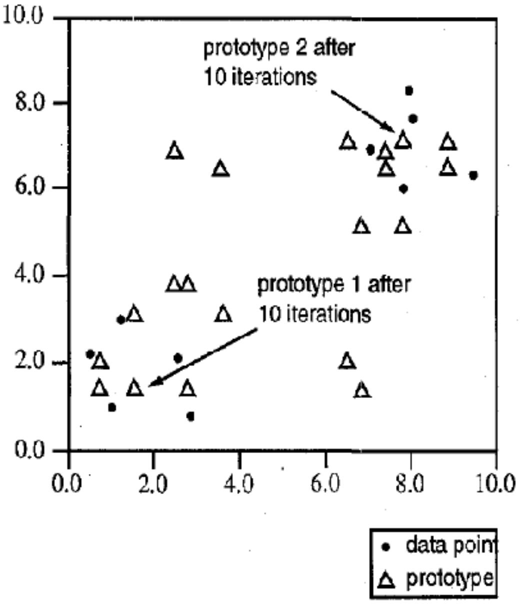


Figure (1): The use of a Kohonen layer, unsupervised, to generate a sequence of prototypes to represent the classes of Table (1).

Kohonen learning is unsupervised, in that a simple measure of the distance between each prototype and the data point allows selection of the winner. Although Kohonen learning selects data points for analysis in random order, we take the points of Table(1) in top to bottom order. For point (1, 1), we measure the distance from each prototype:

$\| (1, 1) - (7, 2) \| = (1 - 7)^2 + (1 - 2)^2 = 37$, and

$\| (1, 1) - (2, 9) \| = (1 - 2)^{2+} (1 - 9)^2 = 65.$

Node A (7, 2) is the winner since it is closest to (1, 1). $\| (1, 1) - (7, 2) \|$ represents the distance between these two points; we do not need to apply the square root function in the Euclidean distance measure because the relation of magnitudes is invariant. We now reward the winning node, using the learning constant c set to 0.5. For the second iteration:

$$W^2 = W^1 + c(X^1 - W^1)$$
$$= (7,2) + .5((1,1) - (7,2)) = (7, 2) + .5((1 - 7), (1 – 2)) \qquad \textbf{(8)}$$
$$= (7, 2) + (-3, -.5) = (4, 1.5)$$

The proposed system depends on the equation (8) for computing it values.

At the second iteration of the learning algorithm we have, for data point (9.4, 6.4):

$$\| (9.4, 6.4) - (4, 1.5) \| = (9.4 - 4)^2 + (6.4 - 1.5)^2 = 53.17 \text{ and}$$
$$\| (9.4, 6.4) - (2, 9) \| = (9.4 - 2)^2 + (6.4 - 9)^2 = 60.15$$

Node A is the winner. The weight for the third iteration is computed from applying eq. (8) again as shown in figure (2):

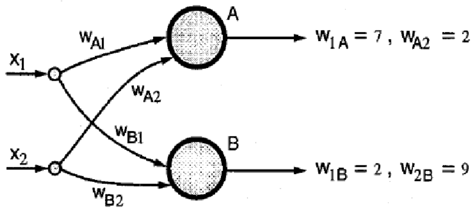


Figure (2) The architecture of Kohonen based learning network for the data of Table (1) and classification of figure (1).

$w^3 = W^2 + c(X^2 - W^2)$
$= (4, 1.5) + .5((9.4, 6.4) - (4, 1.5))$
$= (4, 1.5) + (2.7, 2.5) = (6.7, 4)$
At the third iteration we have, for data point (2.5, 2.1):
$\| (2.5, 2.1) - (6.7, 4) \| = (2.5 - 6.7)^2 + (2.1 - 4)^2 = 21.25$, and
$\| (2.5, 2.1) - (2, 9) \| = (2.5 - 2)^2 + (2.1 - 9)^2 = 47.86$.
Node A wins again and we go on to calculate its new weight vector. Figure (1) shows the evolution of the prototype after 10 iterations. The algorithm used to generate the data of Figure(1) selected data randomly from Table (1), so the prototypes shown will differ from those just created. The progressive improvement of the prototypes can be seen moving toward the centers of the data clusters. Again, this is an unsupervised, winner take- all reinforcement algorithms. It builds a set of evolving and explicit prototypes to represent the data clusters [8].

7. Dynamic Time Warping

Dynamic Time Warping (DTW) is a method to compensate for the variability in speaking rate in template-based system. It was originally developed for isolated word recognition application, and later adapted for text-dependent speaker verification. The model or the reference template is a sequence of feature vectors $\{y_1, y_2, \ldots, y_n\}$. The DTW algorithm is used to find a match between the reference template and the input sequence $\{x_1, x_2, \ldots, x_{m\}}$. In general, N ≠ M. The matching score, (D), is given by

$$D = \sum_{i=1}^{M} d(\mathbf{x}_i, \mathbf{y}_{j(i)}) \qquad \ldots (9)$$

Where the template indices j(i) is obtained by the DTW algorithm, and d(.) represents the distance between the feature vectors. The index j(i) corresponds to the index of the feature vector of the template sequence which matches with the ith vector of the input sequence. Given the reference and test input data, the DTW algorithm performs a constrained, nonlinear mapping of one time axis onto the other to align the two, by minimizing D. The accumulated stance is used as the matching score. The performance of DTW based system critically depends on the accuracy with which the end-points of the speech utterance are located [9].

8. Pattern Matching using Dynamic Time Warping Algorithm

During enrollment, the acoustic features are extracted and stored as reference templates.
During verification, the features extracted from the test utterance are matched with the reference templates using the dynamic time warping (DTW) algorithm [10]. DTW is a dynamic programming based pattern matching algorithm useful for nonlinear time normalization. This nonlinear time normalization takes care of the timing difference between the two speech patterns by warping the time axis of one so that the best match is attained with the other. The process of time normalization requires certain constraints on the warping path.

The sequence of feature vectors corresponding to the speech utterance during testing (A) and during enrollment (B) are given as

$A = \{a_1, a_{2,\ldots,}\ a_{x,\ldots,}\ a_X\}$

$B = \{b_1, b_{2,\ldots,}\ b_{y,\ldots,}\ b_Y\}$

where X and Y are the number of test and reference feature vectors respectively. Due to variation in speaking rates, $X \neq Y$ even for the same utterance recorded during different sessions. The DTW algorithm gives a warping path between A and B such that the feature vectors of the same sound units can be compared [11].

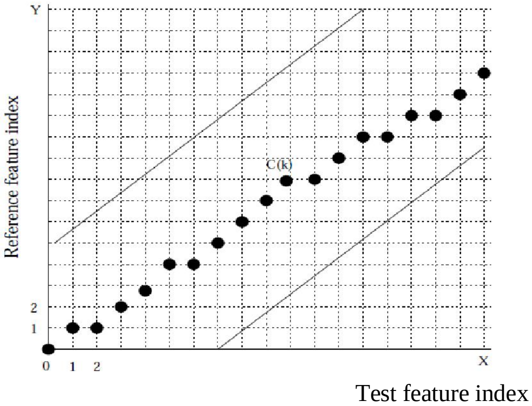


Figure (3): Warping path obtained using the DTW algorithm

In Figure (3), the patterns A and B are represented along the x and y axes, respectively. The match between them can be represented by a sequence of K points C(1),C(2),…,C(k)…;C(K), where each C(k) = ((x(k),y(k)), and x(k) is the frame index of the test utterance and y(k) is the frame index of the reference utterance. Then the sequence, F = C(1),C(2),…,C(k) , … ,C(K) represents a mapping from the time axis of the pattern A to that of the pattern B. This mapping is called the warping function.

As a measure of the difference between the test and reference feature vectors a_x and b_y, a distance d(C) is computed between them which is given by

$$d(C(k)) = \|\mathrm{a}_x - \mathrm{b}_y\| \qquad \textbf{(10)}$$

The accumulated distance

$$\textstyle\sum_{k=1}^{K} d(C(k))$$

For all K points on the warping path is used as the similarity score for making the decision. [12].

9. The Proposed Technique

The system is organized as follows: Preprocessing of the acoustic speech signal by applying Discrete Cosine Transform (DCT) this is followed by a description of the acoustic features used in the baseline system by applying seven moments rules and the pattern matching technique using Kohonen neural network for person identification and DTW algorithm for verification and give the decision logic used for accepting or rejecting a claimant.

The algorithm of the proposed system is shown as follow:

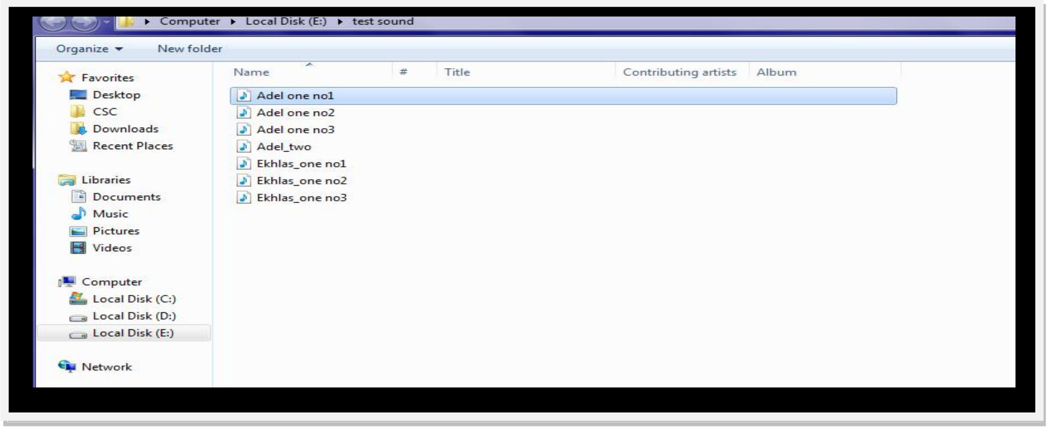
Algorithm (Person Authentication)

```
Input :   text dependent  audio person
Output :  verify authorization

Processing:
Step1: 1)Enter the template sound Temp
       2)Apply the discrete Cosine Transform as mentioned in Eq.2 and put the result in DCTemp
       3)Extract the features from DCTemp as mentioned in Eq.7 and put the result in template data- base TempDB
       4) Apply Kohonen network for person identification as mentioned in Eq.8 and put the result in KTemp.
Step2: 1 )Enter the tested sound Tst
       2) Apply the discrete Cosine Transform  as mentioned in Eq.2 and put the result in DCTst
       3)Extract the features from DCTst as mentioned in Eq.7 and put the result in tested database TstDB
       4)Apply Kohonen network for person  identification as mentioned in Eq.8  and put the result in KTst
Step 3: Apply dynamic time warping (DTW) as mentioned in Eq.10 to verify the authority between KTemp and KTst.
Step4: Give the logic decision for accepting or rejecting the tested sound Tst.
Step5:End
```

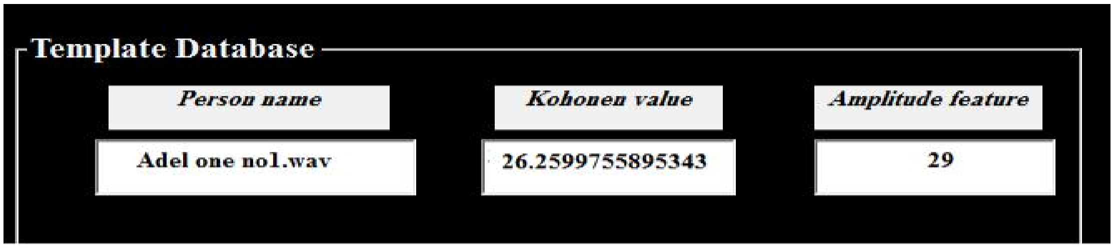
10. **Experimental Results**

Suppose the template database contain single person with audio text dependent ("***one***")



Name	#	Title	Contributing artists	Album
Adel one no1				
Adel one no2				
Adel one no3				
Adel_two				
Ekhlas_one no1				
Ekhlas_one no2				
Ekhlas_one no3				

Figure (1) Template sound selection



Template Database

Person name	Kohonen value	Amplitude feature
Adel one no1.wav	26.2599755895343	29

Figure (2) The features of the template audio.

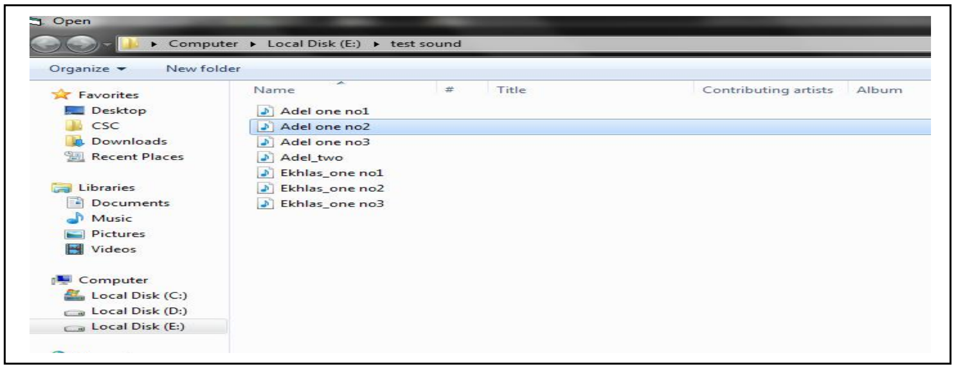
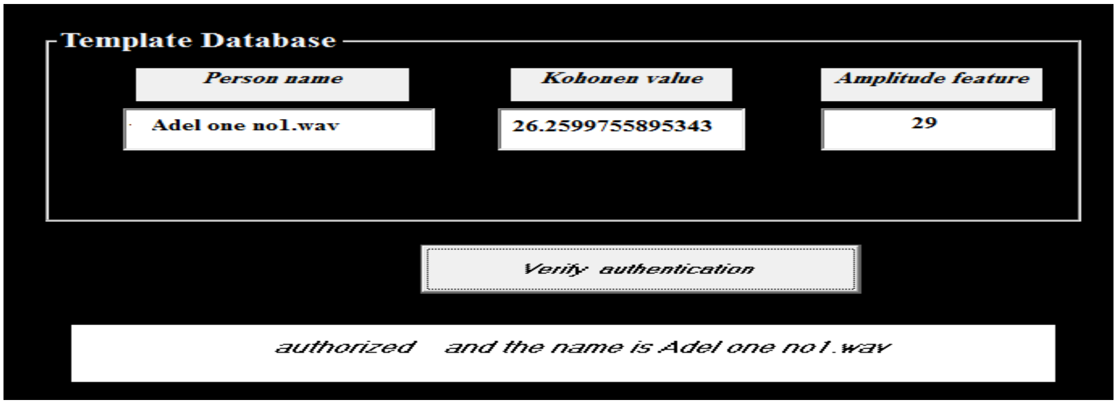


Figure (3) The selection of the same claimant for the same audio text (“*one*”)

Table (1) The features of the claimant that can be selected from figure (3)

Claimant database				
Person name	Kohonen value	Amplitude feature	Decision logic	Identify person name
Adel one no.2	26.2583	29	Authorized	Adel one no. 1

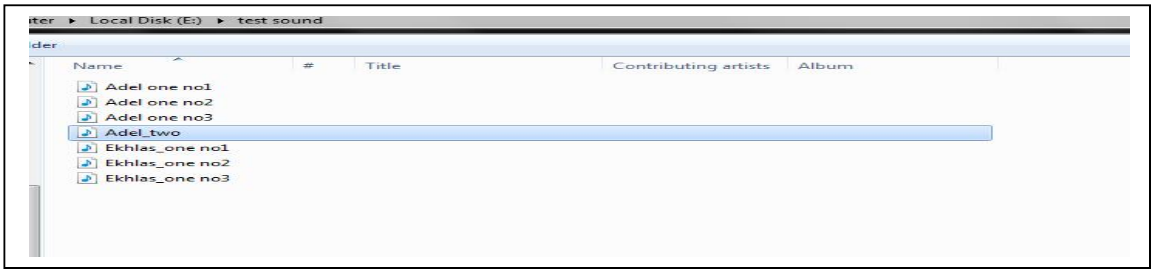


Template Database

Person name	Kohonen value	Amplitude feature
Adel one no1.wav	26.2599755895343	29

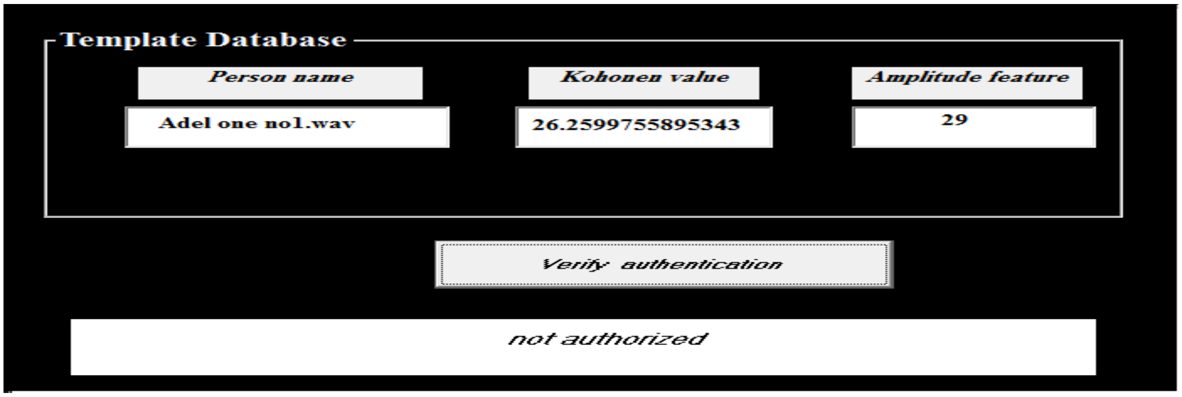
authorized and the name is Adel one no1.wav

Figure (4) Verification of claimant for authorization.



Adel one no1
Adel one no2
Adel one no3
Adel_two
Ekhlas_one no1
Ekhlas_one no2
Ekhlas_one no3

Figure (5) Selection of the same claimant for the different audio text (“*two*”)


Template Database

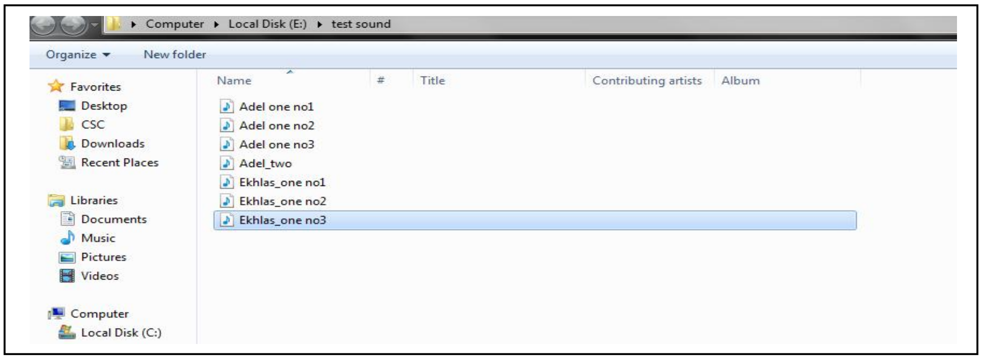
Person name	Kohonen value	Amplitude feature
Adel one no1.wav	26.2599755895343	29

not authorized

Figure (6) The verification of claimant that can be selected form figure (5)

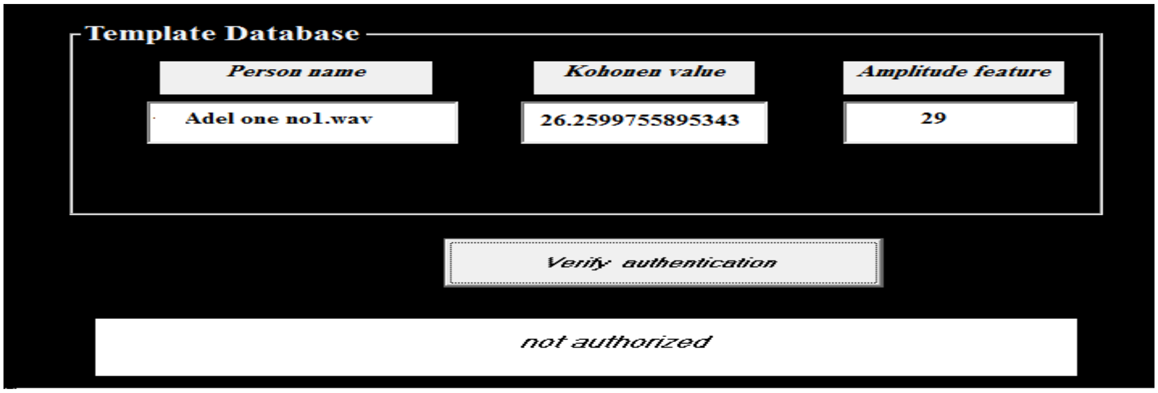
Table (2) The features database of claimant that can be selected from figure (5).

Claimant Database				
Person name	**Kohonen value**	**Amplitude feature**	**Decision logic**	**Identify person name**
Adel two no.1	24.71194	37	Not Authorized	



Adel one no1
Adel one no2
Adel one no3
Adel_two
Ekhlas_one no1
Ekhlas_one no2
Ekhlas_one no3

Figure (7) Different claimant selection for the same audio text (“*one*”)

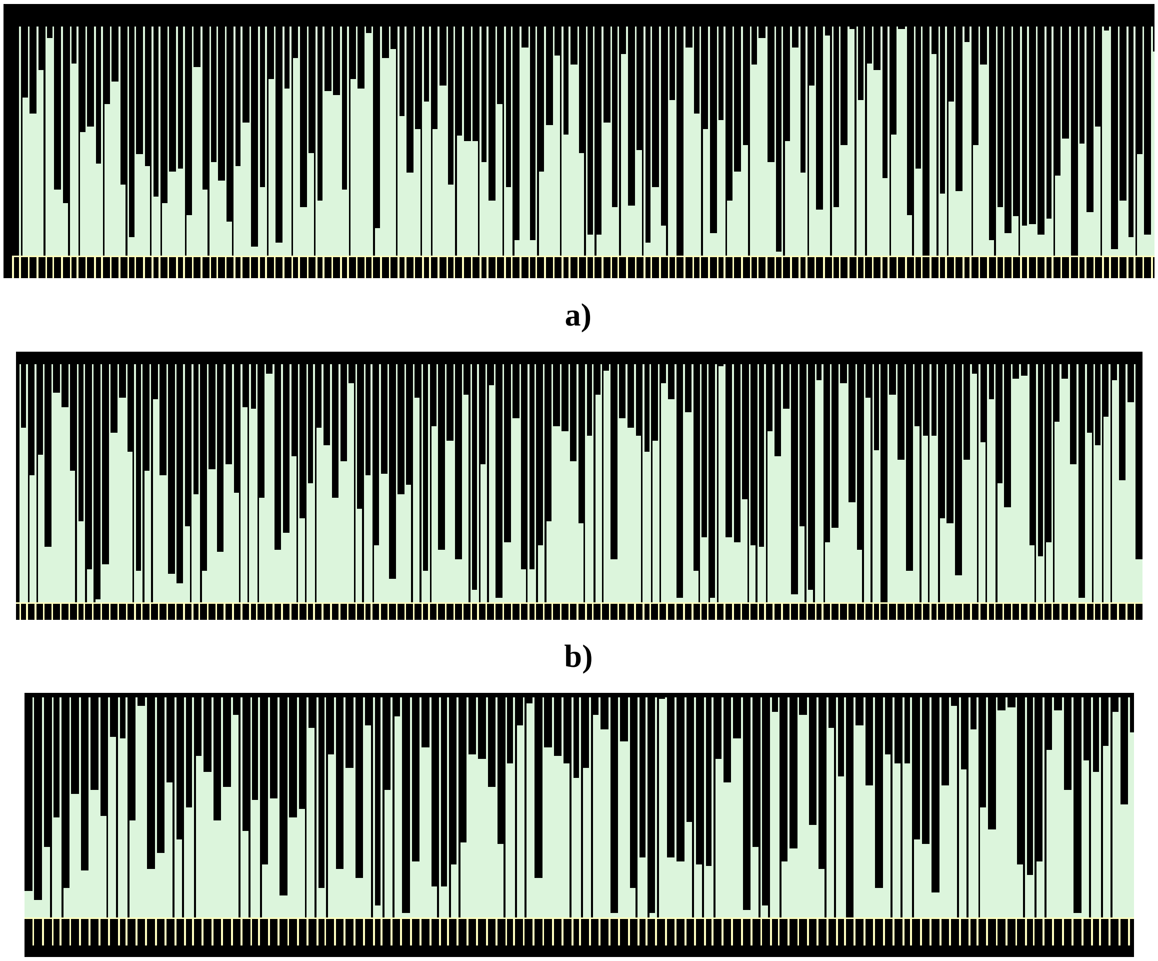


Template Database

Person name	Kohonen value	Amplitude feature
Adel one no1.wav	26.2599755895343	29

not authorized

Figure (8 The verification of claimant that can be selected from figure (7).



a)

b)

c)

Figure (9) Different histograms a) for the template person with audio text dependent ("one"), b) presents Authorization status for text dependent("one"); c) present not Authorization status

Table (3) The features database of claimant that can be selected from figure (7).

Claimant Database				
Person name	**Kohonen value**	**Amplitude feature**	**Decision logic**	**Identify person name**
Ekhlas one no3	28.19122	12	Not Authorized	

11. Conclusion

Audio biometric person authentication is the task of verifying the identity of a person based on the information in the speech signal that occurs during the production of speech.

Uniqueness of a speaker in the voice characteristics is due to the shape and size of the vocal tract, the movements of the articulators, the excitation source, the vibration of the vocal folds, the accent imposed by the speaker, and the speaking rate.

The proposed system can be used Discrete Cosine Transform (DCT) to compress the audio file ,after compression it can be used the seven moments to extract the features, after extraction these features can be entered to the neural network as input and later it can be used dynamic time warping (DTW) for making decision.

During verification, the extracted features are stored as a reference template that represents the biometric identity of the speaker. During authentication, test features are extracted from the speech signal of the claimant, and are matched with the corresponding reference template features to verify the authenticity of the claimant using dynamic time warping (DTW) algorithm. A pattern matching algorithm is used to obtain a measure of dissimilarity between the test and reference patterns. This is followed by application of a decision strategy to determine whether the claim can be accepted or rejected. Decision logic is used to determine whether the speaker is genuine or impostor based on the results of the pattern matching stage. If the speaker is genuine then the system determines the name of genuine person.

The performance is evaluated for two sets of data with 25 and 50 speakers each with different of constraints on the speaker.

12. Recommendations

The proposed system can suggest the following re commendations:

1- It can apply the proposed system to another type of file format such as MP3 or MP4.
2- It can be used another neural network such as back propagation to identify any person.
3- It can be mixed more than one technique for pattern matching to increase the authority.

References

[1] Vinod P., ***"Text dependent audio Visual Biometric Person Authentication"***, A Thesis for the award of the degree of Master of Science Department of Computer Science and Engineering, Indian Institute of Technology Madras, June 2003.

[2] Acharya T. and Ray A.K., ***" image processing principles and applications"***, TA1637, A3, 2005.

[3] Harold F. Tipton, Cissp,***," Information Security Management"*** Handbook, Fifth Edition,2004.

[4] Jinu Mariam ,***"Text-Dependent Speaker Verification Using Segmental, Supraseg and Source Features"***, M.S. thesis, Indian Institute of Technology Madras, Dept. of Computer Science and Engineering, Mar,2002.

[5] Ghosal S.***,"A moment based identified approach to image feature detection",*** IEEE transaction on image processing, 2000.

[6] George F. luger***, "Artificial Intelligence Structures and Strategies for Complex Problem Solving",*** fifth edition, 2005.

[7] James Bezdek***, "Fuzzy Models and Algorithms for Pattern Recognition and Image Processing",*** *University of West Florida,* Springer, 2005.

[8] Toshinori Munakata, ***"Fundamentals of the New Artificial Intelligence",*** Computer and Information Science Department, Cleveland State University, second Edition, 2008

[9] Sadaoki Furui, ***"Cepstral analysis technique for automatic speaker verification,"*** IEEE Trans. Acoustic., Speech, Signal Processing, vol. 29, no. 2, pp. 254{272, Apr. 1981.

[10] Sakoe H.,***" Two level Dp Matching a dynamic programming based pattern matching algorithm for connected word recognition",*** IEEE Trans. Acoustic., Speech, signal Processing, vol. ASSP-27, pp. 588{595, 1998.

[11] Douglas A. Reynolds, ***"An Overview of Automatic Speaker Recognition Technology"*** *ICASSP Orlando, Florida,* pp.4072-4075, 2002.

[12] Mariam J. and Cheedella S. Gupta, ***"Combining Evidence from Source, Supra-segmental and Spectral Features for a Fixed-Text speaker Verification System"***, *IEEE Transaction on Speech and Audio Processing,* pp.575-582, 2005.