Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 53(67), Fascicola 2, 2008

# A Human Person Recognition System using Face and Voice Biometrics

Tudor Barbu[1], Mihaela Costin

**Abstract – We propose a biometric system using two human recognition techniques, both of them using the same kind of supervised classifier. The first one represents a Eigenface based facial recognition approach, while the second is a text-independent speaker recognition method. A minimum mean distance classification technique is provided for person identification. Threshold-based verification approaches are used by both recognition methods.**

**Keywords: face recognition, eigenvectors, eigenface, feature vector, training set, supervised classification, text-independent speaker recognition, mel-cepstral analysis.**

## I. INTRODUCTION

*Biometric recognition* refers to the use of distinctive *physiological* (like voice, fingerprints, face, retina, iris) and *behavioral* (like gait, signature) characteristics, called *biometric identifiers* (or simply biometrics), for automatically identifying and verifying the human persons. In this paper we focus on biometric authentication using two important identifiers: face and voice.

Thus, we propose a two-level biometric system. First, the persons are recognized by their faces [1-7]. Then, their voice recognition is performed for a better authentication [8-11].

In the next section we describe a novel eigenimage based face recognition approach. A speech-independent voice recognition technique, using a mel-cepstral analysis, is provided in the third section.

The two biometric methods use similar supervised person classification schemes in the identification stage. A minimum mean distance classifier is thus proposed by us. Also, threshold-based verification procedures are used by both recognition techniques.

Experiments performed using the described techniques are also presented. The article ends with a conclusions section.

## II. A FACE RECOGNITION TECHNIQUE

Facial recognition represents a very important biometric authentication domain. A face recognition system is a computer-driven application for automatically identifying a person from a digital image. It performs that operation by comparing selected facial features in the input image and a face database.

The most popular recognition techniques include Eigenface [1,2], Fisherface [3], the Hidden Markov Model (HMM) [4], Gabor Filters [5], and the neuronal based Dynamic Link Matching [6]. We provide an eigenimage-based face recognition method in this paper [1,2]. Proposed in 1991 by M. Turk and A. Pentland, the Eigenface approach was the first genuinely successful system for automatic recognition of human faces [1]. It represented a breakway from contemporary research trend on face recognition which focused on identifying some individual features such as eyes, nose, mouth and head outline and developing a face model based on position and size of these characteristics.

Our work is based on this influential work of Turk and Pentland [1]. Let us briefly describe their facial recognition method. One starts with a large set of face images, the training set. Then, each image is represented as a vector $\Gamma_i$, $i = 1,..., M$ and the average vector $\Psi$ is computed. Next, the covariance matrix is obtained as $C = A \cdot A^T$, where $A = [\Phi_1,..., \Phi_M]$, $\Phi_i = \Gamma_i - \Psi$. The eigenvectors and eigenvalues of $C$, a very large matrix, are computed from those of $A^T \cdot A$. Thus, $A \cdot A^T$ and $A^T \cdot A$ have the same eigenvalues and their eigenvectors are related as follows: $u_i = Av_i$. One keeps only $M'$ eigenvectors, corresponding to the largest eigenvalues. Each of these eigenvectors represents an eigenimage or eigenface. Each face image is projected onto each of the eigenfaces, its feature vector, containing $M'$ weights, thus being obtained. Any new input face image is identified by computing the Euclidean distance between its feature vector and each feature training vector. Next, verification procedures may be needed to determine if the input image represents a face at all or if it represents one of the persons from the training set. Threshold values are used for face verification.

[1] Institutul de Informatică Teoretică al Academiei Române, filiala Iaşi, e-mail tudbar@iit.tuiasi.ro

As any pattern recognition system, our face recognition system consists of two main steps: feature extraction and classification. Its last part, face classification, is composed of two processes: face identification, and verification. An additional procedure, namely face detection, may be necessary.

A. *Face feature extraction approach*

Let us consider $M$ matrices of size $N_1 \times N_2$, representing the discrete images. Their corresponding $N_1 \cdot N_2 \times 1$ image vectors are $I_1,..., I_M$. We have $\tilde{Q} = A^T \cdot A$, where

$$A = \begin{Vmatrix} \Phi_1,...,\Phi_M & 0 & 0 \\ 0 & W_1^1,...W_M^1 & 0 \\ 0 & 0 & W_1^2,...W_M^2 \end{Vmatrix} \quad (1)$$

and

$$\begin{cases} \Phi_k = \left\| \Phi_k(i,j) \right\|_{i,j=1}^{N_2,N_1} \\ W_k^1 = \left\| \Phi_k(i+1,j) - \Phi_k(i,j) \right\|_{i,j=1}^{N_2,N_1} \\ W_k^2 = \left\| \Phi_k(i,j+1) - \Phi_k(i,j) \right\|_{i,j=1}^{N_2,N_1} \end{cases} \quad (2)$$

Thus, we determine the eigenvectors $\psi_i$ of the matrix $\tilde{Q}$. Then, the eigenvectors of the covariance operator $Q$ are computed as:

$$\tilde{\varphi}_i = A \cdot \psi_i, \ i = 1,..., M \quad (3)$$

We keep only $M' < M$ eigenimages corresponding to largest eigenvalues and consider the space $X = linspan \ \{\tilde{\varphi}_i\}_{i=1}^{3M'}$. Then, the projection of $[\Phi_i, W_i^1, W_i^2]$ on $X$ is given by:

$$P([\Phi_i, W_i^1, W_i^2]) = \sum_{j=1}^{3M'} w_i^j \cdot \tilde{\varphi}_j, \ i = 1,...,3M \quad (4)$$

where $w_i^j = \tilde{\varphi}_j^T \cdot [\Phi_i, W_i^1, W_i^2]^T$. Therefore, for each face image $I_i$ a corresponding training feature vector is extracted as the sequence of all these weights:

$$V(I_i) = [w_i^1,..., w_i^{3M'}]^T, \ i = 1,..., 3M \quad (5)$$

Thus, the feature training set of the recognition system is obtained as $\{V(I_1),..., V(I_M)\}$, each feature vector being given by (5). Euclidean distance can be used to classify the image feature vectors.

B. *Supervised face classification approach*

We propose a supervised classification approach for face recognition [2]. A minimum mean distance classifier is used for feature vector classification [2, 12,13]. Thus, we provide an extension of the classical variant of minimum distance classifier.

So, let us consider an unknown input image $\Gamma$ to be recognized using the face training set $\{I_1,..., I_M\}$. The feature training set of our classifier is $\{V(I_1),...,V(I_M)\}$ and its metric is the Euclidean distance for vectors. There are three processes to be performed for this image: face detection, face identification and face verification. The first step can be omitted if the input image is a human face for sure.

First the input image is normalized: $\Phi = \Gamma - \Psi$, where $\Psi = \frac{1}{M} \sum_{i=1}^{M} I_i$. The vectors $W^1$ and $W^2$ are computed from $\Phi$ using relation (2). Then it is projected on the eigenspace, using the formula $P(\Phi) = \sum_{i=1}^{M'} w^i \tilde{\varphi}_i$, where $w^i = \tilde{\varphi}_i^T \cdot [\Phi, W^1, W^2]^T$.

Therefore, its feature vector is obtained as $V(\Gamma) = [w^1,..., w^{M'}]^T$.

A threshold-based face detection procedure should be performed to determine if the given image represents a face or not. Therefore, if $\left\| P(\Phi) - \Phi \right\| \le T$, where $T$ is a chosen threshold value, then $\Gamma$ represents a face, otherwise it is a non-face image. If it is a human face, an identification process has to be performed.

Let us suppose there are $K$ persons whose faces are represented in the training set $\{I_1,..., I_M\}$. We could name them registered persons (users). Therefore, we choose to re-note the training images as

$$\{I_1^1,..., I_1^{n(1)},..., I_i^1,..., I_i^{n(i)},..., I_K^1,..., I_K^{n(K)}\}, K < M \quad (6)$$

where $\{I_i^1,..., I_i^{n(i)}\}$ represents the training subset provided by the $i^{th}$ advised person, $n(i)$ being the number of its image faces. We propose a minimum mean-distance classification technique for face identification. This means that for each registered user, the mean distance between its feature training subset and the feature vector of the input face is computed. The input image is associated to the user corresponding to the minimum distance value. That user must be the $k^{th}$ registered person, where:

$$k = \arg\min_i \frac{\sum_{j=1}^{n(i)} d(V(\Gamma), V(I_i^j))}{n(i)} \quad (7)$$

where $d$ is the Euclidean metric. Thus, each input face image $\Gamma$ is identified as an advised person by the above formula.

The face identification procedure has to be completed by a facial verification step. We propose a threshold-based verification technique. Any such method implies the task of choosing a proper threshold value [2]. We provide a novel threshold detection idea, considering the overall maximum distance between any two feature vectors belonging to the same training feature subset as a threshold value. So, each time an input image is to be associated to an registered user of the system, the computed minimum mean distance value has to be compared with that threshold, first. Thus, the input face $\Gamma$ represents the $k^{\text{th}}$ registered person if the following condition is satisfied:

$$\min_i \frac{\sum_{j=1}^{n(i)} d(V(\Gamma), V(I_i^j))}{n(i)} \le T \qquad (8)$$

where $k$ is given by (7) and the threshold value is computed as:

$$T = \max_i \max_{j \ne t} d(V(I_i^j), V(I_i^t)) \qquad (9)$$

If the condition expressed by formula (8) is not satisfied, then the input facial image is rejected by our recognition system. It is labeled as corresponding to an unregistered person.

C. *Experiments*

We performed a lot of experiments on various face datasets and obtained satisfactory face recognition results with our system. It has a high recognition rate, approximately 90%. We used "Yale Face Database B" [7], that contains thousands of $192 \times 168$ face images corresponding to many subjects, for our tests.

Fig. 1. The face image training set

The following example utilizes a training set containing 30 face images representing three people. and illustrated in Fig.1. We compute a total of 90 eigenimages for this set but only the most significant 27 eigenfaces ($M' = 9$) are further used. They are depicted in Fig. 2. Using these eigenimages, the feature training set, containing the 30 face feature vectors, is obtained.

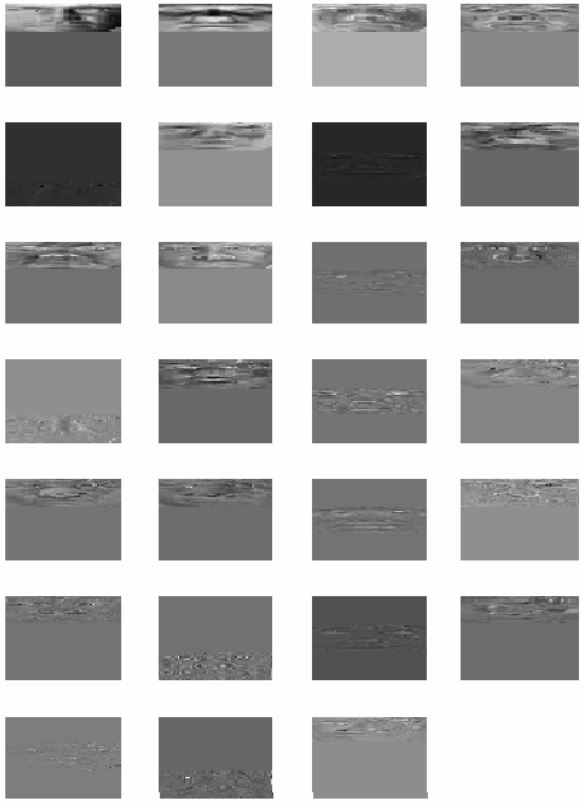

Fig. 2. The most significant eigenfaces

Next, we consider an input image set, composed of the 10 photos displayed in Figure 5.3. Then, each of them is detected as an human face image, so an identification process is performed.

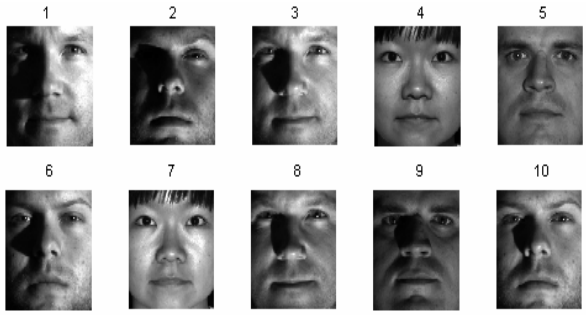


Fig.3. The input image set

The corresponding feature vectors are then determined and the mean distance values between these input image feature vectors and the three feature subsets are computed. These values are registered in the next table, where each column $i$, marked by $D_i$, contains the distance values from the ten feature vectors to the $i^{\text{th}}$ feature subset.

Table 1. The mean distance values

| $D_1$ | $D_2$ | $D_3$ | |
|---|---|---|---|
| 2.573 | 3.090 | 2.351 | 1 |
| 1.919 | 3.859 | 3.257 | 2 |
| 2.557 | 2.788 | 2.273 | 3 |
| 2.736 | 1.954 | 2.534 | 4 |
| 5.795 | 6.716 | 6.103 | 5 |
| 2.270 | 2.595 | 3.135 | 6 |
| 2.108 | 1.789 | 3.293 | 7 |
| 3.147 | 2.623 | 2.331 | 8 |
| 5.101 | 6.959 | 5.143 | 9 |
| 1.923 | 3.467 | 2.926 | 10 |

It results from Table 1 that the input images 1, 3 and 8 are identified as faces of the third registered person from Fig.1, the images 2, 6 and 10 are identified as faces of the first person and images 4 and 7 are identified as the second person. Also, the face images 5 and 9 are identified as belonging to the first registered person but their distance values, 5.795 and 5.101, are found greater than the threshold. In this case we get $T = 2.568$ , so the verification procedure labels the person with faces 5 and 10 as unregistered, the other faces being identified correctly.

## III. SPEECH-INDEPENDENT VOICE RECOGNITION APPROACH

Voice recognition methods can be divided into *text-dependent* and *text-independent* techniques [8]. The former approaches discriminate the users based on the same spoken utterance, while the latter do not rely on a specific speech.

The most successful speech-independent recognition methods are based on Vector Quantization (VQ) or Gaussian Mixture Model (GMM). The VQ-based methods are parametric approaches which use VQ codebooks consisting of a small number of representative feature vectors [9,10], while the GMM-based methods represent non-parametric techniques using $K$ Gaussian distributions [11].

The speaker identification system proposed in this paper uses the melodic cepstral analysis in the feature extraction stage and a minimum mean distance classifier in the classification part. We utilize a threshold –based approach for speaker verification.

### A. *Speech feature extraction approach*

The Mel Frequency Cepstral Coefficients (MFCCs) are the dominant features used for speech and speaker recognition ([8], [10], [12]).

Other voice recognition methods, such as those based on Vector Quantization, utilizes MFCC-based unidimensional feature vectors [10]. Our proposed speech feature extraction technique obtains bidimensional feature vectors.

Thus, a short-time analysis is performed on the sound signal to be featured [12]. The speech signal is divided in overlapping frames having the length 256 and overlaps of 128 coefficients. Then, each resulted segment is windowed, by multiplying it with a Hamming window of length 256. The spectrum of each windowed sequence is then computed, by applying FFT to it. The cepstrum of each windowed frame $s[n]$ is then computed as:

$$C[n] = \text{IFFT}(\log|\text{FFT}(s[n])|) \quad (10)$$

where IFFT represents the Inverse FFT. Next we use the mel-scale, which translates regular frequencies to a scale that is more appropriate for speech. So, a sequence of mel-frequency cepstral coefficients (MFCCs) are obtained for each frame. Each such MFCC set represents a melodic cepstral acoustic vector. Next, a derivation process is performed on these MFCC acoustic vectors.

*Delta mel cepstral coefficients* (DMFCC) are computed as the first order derivatives of mel cepstral coefficients. Then, the *delta delta mel frequency cepstral coefficients* (DDMFCC) are obtained as the second order derivatives of MFCCs. These derivative processes are used because of the intra-speaker variability. Thus, a set of DDMFCC acoustic vectors result for the initial voice signal. Each of them is composed of 256 samples, but the speech information is codified mainly in the first 12 coefficients. So, each acoustic vector is truncated at its first 12 samples and then it is positioned on a matrix column.

The resulted DDMFCC acoustic matrix constitutes a powerful speech feature vector. Let us note $V(S)$ the feature vector corresponding to input speech $S$. Each feature vector has 12 rows and a number of columns depending on the length of the speech signal. So, because of their different dimensions, these feature vectors cannot be classified using the well-known Euclidean distance. For this reason, a special nonlinear metric is introduced in the next section [12].

### B. *A supervised speaker classification and verification scheme*

As mentioned in the introduction, we use a similar supervised classification approach as in the face recognition case. The same minimum mean distance classifier is used for voice classification, but with a different metric. While the face classifier uses the Euclidian metric, the speaker classifier uses a special nonlinear metric.

Thus, we propose a new metric which is able to compute the distance between different sized matrices having a single common dimension, like the acoustic matrices representing our voice feature vectors. It is derived from the Hausdorff distance for sets [12], described as:

$$H(A,B) = \{\max_{a\in A}\inf_{b\in B} dist(a,b), \max_{b\in B}\inf_{a\in A} dist(a,b)\} \quad (11)$$

where *dist* can be the Euclidean metric. After transforming the formula above, we obtain the following metric:

$$d(A,B)=\max\left\{\begin{array}{l}\sup\limits_{1\le k\le p}\inf\limits_{1\le i\le m}\sup\limits_{1\le i\le n}\left|b_{ik}-a_{ij}\right|,\\ \sup\limits_{1\le i\le m}\inf\limits_{1\le k\le p}\sup\limits_{1\le i\le n}\left|b_{ik}-a_{ij}\right|\end{array}\right\} \quad (12)$$

where $A=(a_{ij})_{n\times m}$ and $B=(b_{ij})_{n\times p}$ are matrices having the same number of rows. The nonlinear function $d$ does not represent a Hausdorff metric anymore, but it verifies the three distance properties: positivity, simmetry and triangle inequality [12].

The training set of our classifier contains a collection of spoken utterances, generated by the registered (advised) speakers. Each vocal utterance from the training set constitutes a vocal prototype and represents the same speech. We consider a large spoken text containing all the English language phonemes. Each registered speaker should provide this speech several times.

Therefore, the resulted training set receives the form $P=\{P_1,...,P_N\}$, where each $P_i=\{s_1^i,...,s_{n(i)}^i\}$ represents the set of signal prototypes corresponding to the $i^{\text{th}}$ speaker, $N$ being the number of advised speakers. For each $s_j^i$ the previously described vocal feature extraction is performed, the feature training set $\{\{V(s_1^i),...,V(s_{n(1)}^1)\},...,\{V(s_1^N),...,V(s_{n(N)}^N)\}\}$ being obtained. There are $N$ classes, each of them corresponding to a different advised speaker. Our classification procedure inserts each input vocal signal in the class of the closest registered speaker, which is the speaker corresponding to the smallest mean distance between the feature vector of the input signal and the prototype vectors of the speaker. So, we identify the $p_i^{\text{th}}$ speaker as being the closest to $S_i$, where:

$$p_i=\arg\min_j\frac{\sum_{k=1}^{n(j)}d(V(S_i),V(s_k^j))}{n(j)},\forall i\in[1,n] \quad (13)$$

The $N$ classes of vocal utterances, represents the result of the speaker identification process. The next stage of the recognition process, speaker verification, has to decide if an identified speaker represents a registered user of the system. Let the identified classes be $C_1,...,C_N$. We propose a threshold-based approach, setting a threshold value $T$ and then compare the resulted minimum mean distance values with it. Thus, the following condition has to be tested:

$$\forall i\in[1,N],\forall S\in C_i\,|\,\frac{\sum_{k=1}^{n(j)}d(V(S_i),V(s_k^j))}{n(j)}\le T, \quad (14)$$

where the threshold $T$ is chosen from the numerical experiments. If condition (14) becomes true for a voice sequence $S$ and a class $C_i$, then the vocal utterance $S$ is accepted by the recognition system as an advised vocal input, otherwise being rejected.

## C. *Experiments*

The experiments we have performed prove the effectiveness of our voice recognition system. We obtain a high speaker recognition rate, that is approximately 85%.

Let us describe now a simple speaker recognition example using our approach. We consider three registered speakers and a long sequence of words, containing all the English phonemes, to be spoken by each of them.

The training set contains four vocal utterances having that sequence of words as text. We got one recording for the first advised user, two recordings for the second user and one recording for the last one. The prototype speech signals and the corresponding training feature vectors, are represented as RGB color images in Fig. 4.

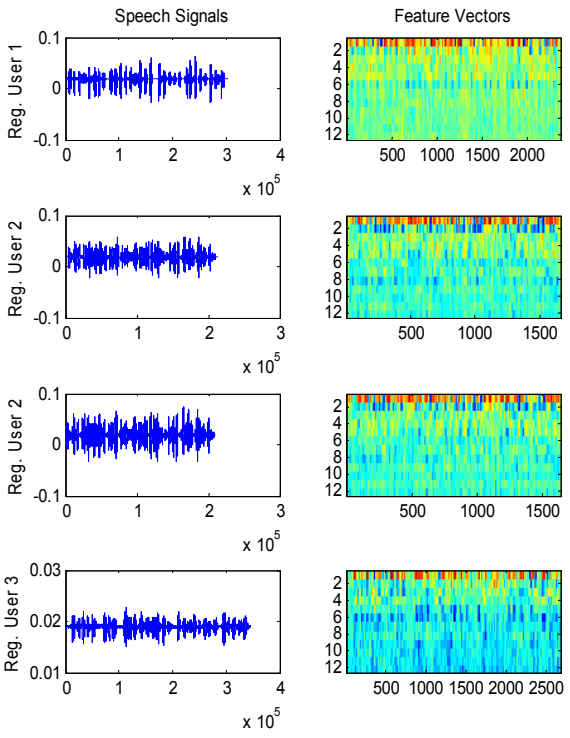


Fig. 4. The prototype speech signals and their feature vectors

Then, we consider a sequence of nine input speech utterances to be recognized, each of them having a different spoken text. Their speech signals, $\{S_1,S_2,S_3,S_4,S_5,S_6,S_7,S_8,S_9\}$, are represented in the fifth figure.

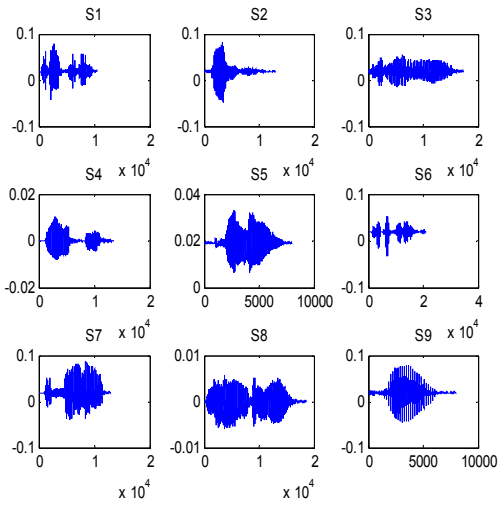


Fig. 5. The input speech signals

Next, the feature vectors $V(S_i)$ are computed, and they are displayed in Fig. 6. The mean distance values between the input feature vectors and the training feature subsets are registered in the Table 2.

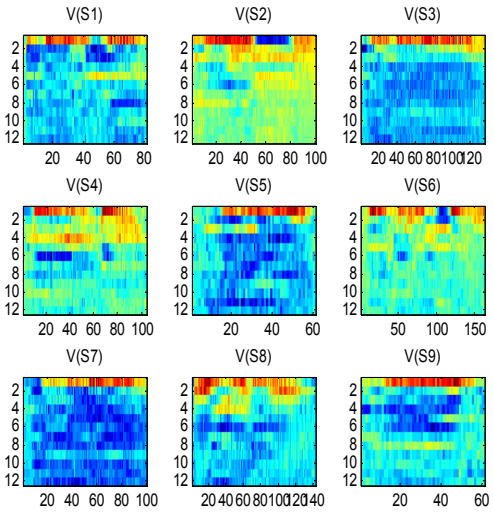


Fig. 6. The speech feature vectors

Table 2. The mean distance values

| | Speaker 1 | Speaker 2 | Speaker 3 |
|---|---|---|---|
| Input 1 | 5.6678 | 3.4676 | 6.2476 |
| Input 2 | 4.1606 | 7.2581 | 6.4327 |
| Input 3 | 5.9543 | 3.8976 | 4.3671 |
| Input 4 | 6.0946 | 4.7342 | 2.9857 |
| Input 5 | 10.6853 | 9.7366 | 8.6545 |
| Input 6 | 5.1522 | 5.7855 | 6.3879 |
| Input 7 | 7.5031 | 4.8871 | 10.8775 |
| Input 8 | 8.7624 | 7.9964 | 5.8976 |
| Input 9 | 3.2465 | 8.0245 | 4.9082 |

First, we get the following identification result, based on the values registered in the table. The input signals 2, 6 and 9 belong to the first speaker, the input speeches 1, 3 and 7 are associated to the second registered speaker, and the input signals 4, 5 and 8 belong to the third speaker. Using the threshold value $T$ =7.5, we then obtain the final speaker recognition result: Speaker 1 =>$\{S_2, S_6, S_9\}$, Speaker 2 =>$\{S_1, S_3, S_7\}$, Speaker 3 =>$\{S_4, S_8\}$ and finally, Unregistered Speaker =>$\{S_5\}$.

## IV. CONCLUSIONS

A two-level human person recognition system has been proposed in this paper. We have used two biometrics in this paper, face and voice, and obtained satisfactory results.

The main contributions of this paper are as follows: the Eigenface based face feature extraction, the DDMFCC based voice featuring, the supervised classification scheme, the proposed nonlinear metric and the threshold based speaker verification approach. Our future research will focus on adding recognition operations based on other human identifiers to this biometric system.

## REFERENCES

[1] M. A. Turk, A. P. Pentland, "Face recognition using eigenfaces", In *Proc. Of Computer Vision and Pattern Recognition*, pp. 586-591, IEEE, June 1991b.

[2] T. Barbu, "Eigenimage-based face recognition approach using gradient covariance", *Numerical Functional Analysis and Optimization*, Vol. 28, Issue 5 & 6 May 2007, pp. 591-601.

[3] B. Yin, X. Bai, Q. Shi, Y. Sun, "Enhanced Fisherface for Face Recognition", *Journal of Information and Computational Science*, No. 3, pp. 591-595, 2005.

[4] Ara V. Nefian ,"A hidden Markov model based approach for face detection and recognition", *PhD Thesis,* 1999.

[5] K. C. Chung;, S. C. Kee, S. R. Kim, "Face recognition using principal component analysis of Gabor filterresponses", *Proceedings of International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 53-57, 1999.

[6] L. Wiskott, C. Malsburg, "Recognizing Faces by Dynamic Link Matching", ***NeuroImage,*** Volume 4, Issue 3, December 1996, pp. S14-S18.

[7] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose", *IEEE Trans. Pattern Anal. Mach. Intelligence*, Vol. 23, No. 6, 2001, pp. 643-660.

[8] R. A. Cole, J. Mariani, H. Uszkoreit , A. Zaenen, V. Zue, "Survey of the State of the Art in Human Language Technology". Cambridge University Press, 1997.

[9] H. Gish, M. Schmidt, "Text-Independent Speaker Identification", *IEEE Signal Processing Magazine*, IEEE,oct. 1994, pp. 1437-62.

[10] N. Bagge, C. Donica, "ELEC 301: Final Project Text Independent Speaker Recognition", *ELEC 301 Signals and Systems Group Projects*, 2001.

[11] D. A. Reynolds, R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, 1995, pp. 72-83.

[12] T. Barbu, "Discrete speech recognition using a Hausdorff-based metric", In *Proceedings of the 1st Int. Conference of E-Business and Telecommunication Networks*, *ICETE 2004*, Setubal, Portugal, Vol. 3, Aug. 2004, pp.363-368.

[13] R. Duda, P. Hart, D. G. Stork, *Pattern Classification*, John Wiley & Sons, 2000.