FULL PAPER
International Journal of Recent Trends in Engineering, Vol 2, No. 5, November 2009

Modified PCA Speaker Identification Based System Using Wavelet Transform and Neural Networks

Omar Daoud[1], Abdel-Rahman Al-Qawasmi[1], and Khaled daqrouq[1]
[1]Philadelphia University/Communications and Electronics Dept., Amman, Jordan
odaoud@philadelphia.edu.jo, qawasmi@philadelphia.edu.jo, haleddaq@yahoo.com

***Abstract*—This work investigates to improve the robustness of the speaker identification systems based on a modified version of Principal Component Analysis (PCA) and Continuous Wavelet Transform (CWT). Therefore, this work proposes a robust feature extraction method based on MPCA instead of Mel Frequency Cepstral Coefficient (MFCC) that is used in the literature, which is based on converting the common Eigen matrix from two dimensional into a one dimensional one.**

A simulation program has been built to proof the given mathematical model for the proposed work. At a certain SNR level of the CWT (6dB) the achieved improvement in the classification process was approximately 7.3% (85-92.3%) over the previously published work that was based on the MFCC with CWT.

***Index Terms*— Speaker identification, Continuous wavelet transform, MPCA, Neural network**

I. INTRODUCTION

Speaker identification plays an important role in electronic authentication. In an operational environment speech is degraded by many kinds of interferences. The interference can be classified broadly as stationary or non-stationary. Stationary interference is noise which can be dealt with by using de-noising and noise reduction techniques; whereas non-stationary interference could be speech from a different speaker. Such interference is a common occurrence and the corrupted speech is known as co-channel speech [1-4].

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique can make it possible to use the speaker's voice to verify their identity and control access to various services. These services include voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. From the viewpoint of technology, speaker recognition is a general term, which refers to any task to discriminate people based upon their voice characteristics [5-7].

Meeting transcription is a desirable, but extremely challenging, feature of automated meeting systems. A preliminary step to enabling meeting transcription is identifying who is speaking for each frame of audio data. This problem is commonly referred to as speaker identification. By first pre-processing the audio signal with a speaker identification algorithm, individual speech recognition models can then be used on the segmented speech pertaining to each of the meeting participants [8-10].

The modern computerized global, authentication and privacy are taking a place to identify the access to different systems using passwords, identification IDs or pin numbers. Therefore the researchers have turned their interests to the proposition of good classifiers to access these services. During these developments, sophisticated systems are created, discussed and debated every day or other more natural solutions as an alternative to the conventional authentication patterns. Many natural features could be used to identity the users such as the fingerprints, faces, eyes, and the speech signals. Due to simplicity and being the object of interest particularly in mobile and telephony systems, the speech feature attracts the researchers' attentions [11-13].

Number of methods using robust feature extraction method based on PCA instead of MFCC [14] is proposed.

In this work the Modified Principal Component Analysis (MPCA) is used in statistics to extract the main relations in data of high dimensionality.

The introduced system depends on two features extraction stages; WT and MPCA due to its simplicity and better accuracy comparing to linear prediction coding or FFT based method. The system works with capability of features tracking even with 6 dB SNR, which accomplished due to MPCA and WT features extraction methods.

The paper is divided into 5 sections. A mathematical derivation of the CWT is described in Section 2. Section 3 introduces Feature Extraction using Modified Principal Component Analysis. The Proposed work main parts and results are depicted in Sections 4 and 5, respectively. After that, the conclusion is drawn in Section 6.

II. CONTINUOUS WAVELET TRANSFORM

The Fourier transform is very widely used tool for many mathematical or scientific applications, but it is well suited only to the study of stationary signals where all frequencies have an infinite coherence time. The Fourier analysis brings only global information which is not enough to detect compact patterns. Gabor introduced a local Fourier analysis, taking into account a sliding window [18], leading to a time frequency-analysis. This methodology is only applicable to problems where the coherence time is independent of the frequency. This is the case for instance for singing signals which have their coherence time determined by the geometry of the oral cavity. Morlet has introduced the Wavelet Transform [18] in order to have a coherence time proportional to the period.

The Morlet-Grossmann definition of the continuous wavelet transform [20] for a one dimensional 1D signal $f(x) \in L^2(R)$ is:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x) \psi^* \left(\frac{x-b}{a} \right) dx \qquad (1)$$

where z* denotes the complex conjugate of z, $\psi^*(x)$ is the analyzing wavelet (Figure.2), a (>0) is the scale parameter and b is the position parameter.

The transform is characterized by the following three properties:

1. It is a linear transformation,
2. It is covariant under translations:

$$f(x) \to f(x-u) \quad W(a,b) \to W(a,b-u) \qquad (2)$$

3. It is covariant under dilations:

$$f(x) \to f(sx) \; W(a,b) \to s^{-\frac{1}{2}} W(sa,sb) \qquad (3)$$

The last property makes the wavelet transform very suitable for analyzing hierarchical structures. It is like a mathematical microscope with properties that do not depend on the magnification.

In Fourier space, we have:

$$\hat{W}(a,v) = \sqrt{a} f(v) \psi^*(a,v) \qquad (4)$$

When the scale a varies, the filter $\psi^*(av)$ is only reduced or dilated while keeping the same pattern.

Now consider a function $W(a,b)$ which is the wavelet transform of a given function *f(x)*. It has been shown that *f(x)* can be restored using the formula:

$$f(x) = \frac{1}{C_\chi} \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{a}} W(a,b) \chi \left(\frac{x-b}{a} \right) \frac{da.db}{a^2}$$

where

$$C_x = \int_0^{+\infty} \frac{\psi^*(v)\hat{\chi}(v)}{v} dv = \int_{-\infty}^{0} \frac{\psi^* \hat{\chi}(v)}{v} dv \qquad (5)$$

Generally $\chi(x) = \psi(x)$, but other choices can enhance certain features for some applications. The reconstruction is only available if C_χ is defined (admissibility condition). In the case of $\chi(x) = \psi(x)$, this condition implies $\hat{\psi}(0) = 0$, i.e. the mean of the wavelet function is 0.

The wavelet defined by Morlet [19] is a complex wavelet which can be decomposed in two parts, one for the real part, and the other for the imaginary part.

$$\begin{aligned} g_r(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cos(2\pi v_0 x) \\ g_i(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \sin(2\pi v_0 x) \end{aligned} \qquad (6)$$

where v_0 is a constant. The admissibility condition is verified only if $v_0 > 0.8$.

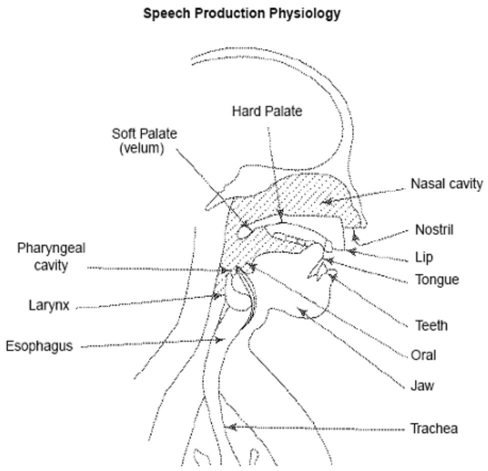

Figure 1. Speech Production Physiology.

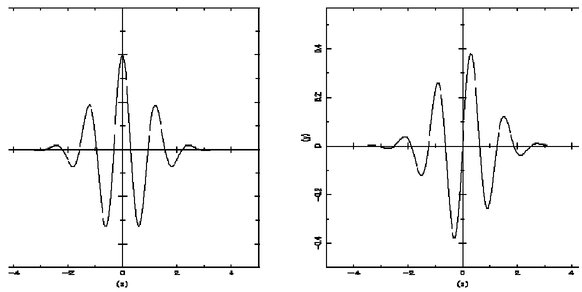

Figure 2 Morlet's wavelet: real part at left and imaginary part at right.

III. Feature Extraction using Modified Principal Component Analysis

In this work the MPCA is used in statistics to extract the main relations in data of high dimensionality. A common way to find the principal components of a data set is by calculating the eigenvectors of the data correlation matrix. These vectors give the directions in which the data cloud is stretched most. The projections of the data on the eigenvectors are the principal components. The corresponding eigenvalues give an indication of the amount of information the respective principal components represent. Principal components corresponding to large eigenvalues represent much information in the data set and thus tell us much about the relations between the data points [15, 16].

So the basic idea in PCA is to find the components $S_1, S_2, S_3, \ldots\ldots S_n$ so that they explain the maximum amount of variance possible by n linearly components. PCA can be defined in an intuitive way using a recursive formulation. Define the direction of the first principal component, say w_1, by

$$w_1 = \arg \max_{\|w\|=1} E\{(w^T x)^2\} \tag{7}$$

where w_1 is of the same dimension *m* as the data vector *x* of the speech signal. Thus the first principal component is the projection on the direction in which the variance of the projection is maximized. Having determined the first *k-1* principal component of the residual as

$$w_k = \arg \max_{\|w\|=1} E\{(w^T (x - \sum_{i=1}^{k-1} w_i w_i^T)^2\} \tag{8}$$

The principal components are then given by

$$S_i = w_i^T x \tag{9}$$

In practice, the computation of the w_i can be simply accomplished using the (sample) covariance matrix $E\{x^T x) = C$. The w_i are the eigenvectors of C that correspond to the *n* largest eigenvalues of C. Thus, this technique orthogonalizes the components of the input vectors so that they are uncorrelated with each other, orders the resulting orthogonal components (principal components) so that those with largest variation come first, and eliminates those components that contribute the least to the variation in the data set.

The used procedure to calculate the principal components for the speech signals is as follows:

1. Based on the application, adjust the desired number of eigenvectors that will be used related to the largest number of eigenvalues.
2. Process each signal alone
   - Get a zero mean one by subtracting it is mean,
   - Find the covariance matrix for the zero mean signal, and then take it is eigenvectors to form the eigensignal.
3. Project each zero mean signal on its own eigensignal, so we get an uncorrelated data matrix related to the first signal, then save it as the first vector in the input matrix. Repeat these steps to complete the data matrix.

Trying to improve the classification performance, a MPCA is proposed in this work based in getting more data reduction.

Thus, for more dependency reduction and after finding a set of two-dimensional eigenobjects using PCA, they will be converted to a set of one-dimensional eignobjects (a skeleton) by a second projection of them once again using the PCA for each signal alone. This proposed work is summarized as follows:

1. Enter the total number of signals,
2. To reduce the dimensionality for the zero mean data, calculate an average signal of the input ones, and then subtract it from the signals database, thus we can work with zero mean input signals.
3. Arrange the input signals in one matrix (*has the same number of rows as the images, and columns equal to the number of columns of any image multiplied by the total number of signals*). This matrix will be processed to find the initial skeleton matrix (finding the covariance matrix, and then taking its eigenvectors. This matrix has the same size of any signal).
4. Process each signal alone by the projection on the initial skeleton transpose matrix, to get a matrix with the same size as the input signal.
5. Find the covariance matrix for each resultant signal in step 4 after finding the zero mean column signal. Then after calculating the eigenvectors, the skeleton matrix will be generated.
6. Project the zero mean column for each signal on its own transpose skeleton matrix.
7. Rearrange the projected data matrix, to be the d column on the output data matrix. Returns to step 4, to complete the output data matrix.
8. To understand the procedure for this algorithm, the example will demonstrate its implementation procedure:

1. Assume two signal matrices :

f1= $\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$, f2= $\begin{bmatrix} 5 & 4 \\ 2 & 1 \end{bmatrix}$

2. Find the average for these two signals:

F_av= $\begin{bmatrix} 3 & 3 \\ 2 & 2 \end{bmatrix}$

3. Find the zero mean signals:

e1= f1- f_av= $\begin{bmatrix} -2 & -1 \\ 0 & 1 \end{bmatrix}$,

e2= f2-f_av= $\begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix}$.

4. Rearrange the two zero mean signals in the same matrix as:

A=[e1 e2] = $\begin{bmatrix} -2 & -1 & 2 & 1 \\ 0 & 1 & 0 & -1 \end{bmatrix}$

5. Find the covariance matrix:

C= A AT = $\begin{bmatrix} 10 & -2 \\ -2 & 2 \end{bmatrix}$

6. Find the eigenvalues, and their corresponding eigenvectors:

[u, d] = eig(C)

$$u=\begin{bmatrix}-0.9732 & -0.2298\\ 0.2298 & -0.9732\end{bmatrix}$$

$$d=\begin{bmatrix}10.4721 & 0\\ 0 & 1.5279\end{bmatrix},$$

$$u1=\begin{bmatrix}-0.9732\\ 0.2298\end{bmatrix},$$

$$u2=\begin{bmatrix}-0.2298\\ -0.9732\end{bmatrix}.$$

7. The processing in this step will be for each zero mean signal alone:

take e1:

w11= u1T * e1 = [1.9465 1.2030].

w22 = u2T * e1 = [0.4595 - 0.7435].

* notice that e1 = u1 w11 + u2 w22.

so the resultant projected matrix is fw = [fw1 fw2] , where

$$fw1 = w11T = \begin{bmatrix}1.9465\\ 1.2030\end{bmatrix}, \text{ and}$$

$$fw2 = w22T = \begin{bmatrix}0.4595\\ -0.7435\end{bmatrix}.$$

8. Find the column average for the resultant matrix

$$fw\_av = \begin{bmatrix}1.2030\\ 0.2298\end{bmatrix}$$

9. Find the zero mean column matrix:

Aw = [ew1 ew2], where

$$ew1 = fw1 - fw\_av = \begin{bmatrix}0.7435\\ 0.9732\end{bmatrix}, \text{ and } \quad ew2 = fw2 - fw\_av = \begin{bmatrix}-0.7435\\ -0.9732\end{bmatrix}$$

10. Find the covariance matrix:

$$Cw = Aw * AwT = \begin{bmatrix}1.1056 & 1.4472\\ 1.4472 & 1.8944\end{bmatrix}.$$

11. Find the eigenvalues and their eigenvectors

[uw, dw] = eig(Cw)

$$uw=\begin{bmatrix}0.7947 & 0.6071\\ -0.6071 & 0.7947\end{bmatrix}, \quad dw=\begin{bmatrix}0 & 0\\ 0 & 3\end{bmatrix},$$

$$uw1=\begin{bmatrix}0.7947\\ -0.6071\end{bmatrix},$$

$$uw2=\begin{bmatrix}0.6071\\ 0.7947\end{bmatrix}.$$

12. Project the zero mean column matrix on uw

$$v = \begin{bmatrix}v11 & v21\\ v12 & v22\end{bmatrix}, \text{ where}$$

v11= uw1T ew1 = 0, v21= uw2T ew1 =1.2247

v12 = uw1T ew2 = 0, v22= uw2T ew2 = -1.2247

save these four elements a column in the first column of the output data matrix.

* notice that

ew1= uw1 v11 + uw2 v21

ew2= uw1 v21 + uw2 v22

13. Repeat steps 7 to 12 for the second signal matrix.

Note that the size of the matrix can be reduced in step 6 by removing the eigenvector corresponding to the smallest eigenvalues; the same can be done in step 11.

IV. PROPOSED WORK MAIN PARTS

Figure 3 shows the main parts of the proposed work. The first block feature extracting is accomplished by WT and MPCA, while the second block presents identification process via verification by FFBP-NN.

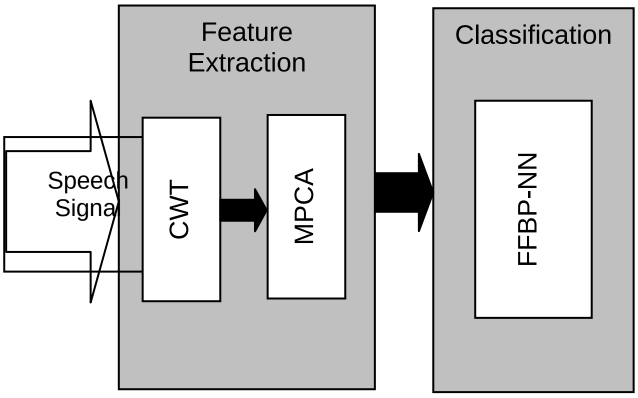

Figure 3. Block diagram of speaker identification system

The first stage of this method decomposes the speech signal into CWT sub-signals of a given scale that based on the speaker own feature frequency related to the vocal tract as shown in Figure 1. The CWT scale-determination is very challenging problem because of its non-stationary nature contained in speech signals, therefore, the used scale-determination is chosen experimentally by studying a huge database of about 1000 speech signals. This assists greatly in finding out a scale matching all speech signals' database.

The second stage of feature extraction method is accomplished by applying the MPCA to the WT coefficients. MPCA has been exploited in different tasks, mainly, in image, speech and speaker recognition. MPCA is superior to most feature tracking popular methods such as Fast Fourier Transform (FFT) and linear prediction coding. This is due to its capability of removing the redundancy in data since it eliminates those components that contribute the least to the variation in the data set.

Specifically, the speech signal is decomposed into CWT sub-signals $(d_1, d_2,\ldots, d_J)$, where each of them is generated using (1) of particular level ($j=1,2,\ldots,J$), as we said above. This is accomplished by convolving the signal with mother wavelet function. After that, it goes to

MPCA block which extracts the eigenvectors of the input database as shown previously in Figure 3.

To take matching decision, this matrix is given to a FEBNN to be trained. Now if these features match any of our models stored in the system, system will accept, otherwise, system will cancel the trail.

V. RESULTS AND DISCUSSION

The achieved results here just from about 28 ones (the input + noisy ones). After the processing stage I have taken the average as shown in the results section.

As a testing speech signals data base is used. The signals were recorded via PC-sound card, with spectral frequency 4000 Hz and sampling frequency 16000 Hz, over about 3 sec. time duration. Each speaker recorded "Eftah" Arabic word that means "Open" in English. Each utterance of eight "Eftah" words was recorded 4 times by the speaker.

From above speech signal discretion, we can notice that presented recognition system is text-dependent system, because prompts are common across all speakers that can share secrets (passwords or PINs).

In order to create multi-factor authentication scenarios, the speaker in each trail is compared to all models stored in database.

For speech signals identification via verification, Neural Networks is studied in term of a Feed Forward Back Propagation Neural Network FFBNN method with a "Tansig" transfer function. Tab.1 summarizes the recognition rate improvement for the average of 30 runs of using the MPCA over the previously published work in [17]. These results were achieved for a FFBNN of noisy speech signals of about 6dB SNR. The use of MPCA increases the system robustness and improves the recognition rate of about 7% when the used *J* equals to 0. We can notice that the use of MPCA improves the recognition process to be a monotonically increasing process.

TABLE I.
THE EFFECT OF WT ON NOISY SIGNALS RECOGNITION

6 dB SNR		
	Using MFCC	Using MPCA
J	Recognition Rate [%]	
0	85	92.3
1	89.9	94.1
2	95.9	97.6
3	89.9	98.4
5	99.89	99.92
15	95	99.97

The achieved improvement is clearly shown in Figure 4. This figure expresses the total average of the different *J*'s used.

In Table 2 the effect of number of hidden neurons in the hidden layer on the recognition rate is studied. We can notice the ability of identification based on this algorithm. In this table the FFBNN is used with a "Tansig" transfer functions. Number of epochs is 200 and WT level is 8 for all cases Different SNR ratios are used.

The Network is designed as three layers: one input of

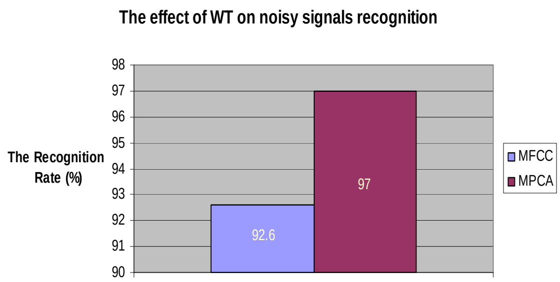

Figure 4. The total average recognition rate for the speaker identification system with different J-parameters

13 neurons, one hidden of 4 neurons and one output of 4 neurons. The maximum recognition rate was 95% for the use of MFCC, while this ratio is improved to 99.7% when the MPCA was used. Same design is used in Tab.3 with 13 hidden neurons. Thus, as an observation, increasing the hidden neurons will improve the recognition rate.

TABLE II.
ILLUSTRATES THE EFFECT OF NUMBER OF 4 HIDDEN NEURONS AT THE RECOGNITION RATE

J=8, 200 epochs and 4 hidden neurons		
SNR (dB)	Recognition Rate [%]	
	MFCC	MPCA
4	79.9	87.2
2.2	89.9	94.7
14	90	96
28	95	98.5

Figure 5 shows the total average improvement of the recognition ratio at different SNR values.

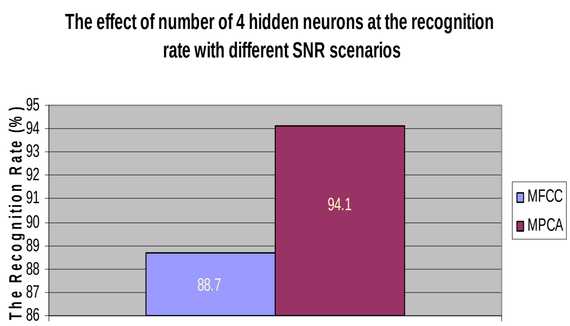

Figure 5. The total average recognition rate for a 4 hidden neurons speaker identification system with different SNR scenarios

From this figure, the MPCA improves the systems' recognition ratio of about 6% over the previously published work.

TABLE III.
ILLUSTRATES THE EFFECT OF NUMBER OF 13 HIDDEN NEURONS AT THE RECOGNITION RATE

J=8, 200 epochs and 13 hidden neurons		
SNR (dB)	Recognition Rate [%]	
	MFCC	MPCA
4	93	96.8
2.2	95	98.2
14	98.7	99.94

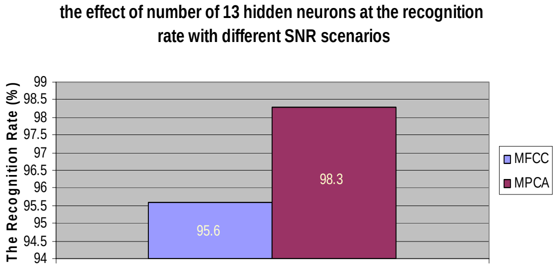

Figure 6. The total average recognition rate for a 13 hidden neurons speaker identification system with different SNR scenarios.

Moreover, there is a slight improvement on the recognition rate when the hidden neurons are increased from 4 to 13. This could be seen in Figure 6, where the difference in recognition ratio is about 3% for the MPCA over the MFCC.

VI. CONCLUSIONS

In this paper, the effect of Wavelet Transform on speaker feature extracting is studied. The introduced system in this paper depends on two features extraction stages; WT and MPCA due to its simplicity and better accuracy comparing to linear prediction coding or FFT based method. The system works with capability of features tracking even with 6 dB SNR, which accomplished due to MPCA and WT features extraction methods.

NN classification method has been imposed in this work for text-dependant system; so that the system can be applied to clarify passwords, PINs, or any identification patterns in any security system, since up to 99.93% identification rate was achieved in the proposed system.

REFERENCES

[1] D. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Communications, vol. 17, pp. 91–108, 1995.
[2] Alex Park and Timothy J. Hazen, ASR Dependent Techniques for Speaker Identification, the 7th International Conference on Spoken Language Processing, Sep. 16-20, 2002, Denver, Colorado, pp. 1337-1340.
[3] R. E. Yantorno, Co-channel speech study, final report for summer research faculty program, tech. rep., Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, 1999.
[4] Saurabh S. Khanwalkar†, Brett Y. Smolenski†, Robert E. Yantorno† and S. J. Wenndt‡, Enhancement of Speaker Identification Using SID-Usable Speech, 13th European Signal Processing Conference, Antalya 2005.
[5] Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (eds.): Survey of the State of the Art in Human Language Technology. Cambridge University Press, 1997.
[6] Doddington G.R.: Speaker Recognition-Identifying People by Their Voices. Proceedings of IEEE, Vol. 73, No. 11, 1986, pp. 1651-1644.
[7] Asterios Toutios, K. G. Margaritis, Development of a Text-Dependent Speaker Identification System with the OGI Toolkit, 2nd Hellenic Conf. on AI, SETN-2002, 11-12 April 2002, Thessaloniki, Greece, Proceedings, Companion Volume, pp. 525-530
[8] J. Ortega-Garcia and J. Gonzalez-Rodriguez, “Providing single and multi-channel acoustical robustness to speaker identification systems,” in Proc ICASSP, 1997, vol. 2, pp. 1107–1110.
[9] J.P. Campbell, “Speaker recognition: a tutorial,” in Proc. of the IEEE, 1997, vol. 85, pp. 1437 – 1462.
[10] J.S. Garofolo, C.D. Laprun, and J.G. Fiscus, “The rich transcription 2004 spring meeting recognition evaluation,” in The Rich Transcription 2004 Spring Meeting Recognition Workshop, 2004.
[11] Rita H Wouhaybi, Mohamad Adnan Al-Alaoui, Comparison of Neural Networks for Speaker Recognition. IEEE Member, IEEE Senior Member IncoNet sal American University of Beirut.
[12] Fernando L. Podio1 and Jeffrey S. Dunn2, Biometris from the movies, National Institute of Standards and Technology.
[13] Elisabeth Zetterholm, Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success. Lund University. (2003).
[14] Hironori Matsumasa,Tetsuya Takiguchi, Yasuo Ariki, I-Chao LI, Toshitaka akabayashi, Integration of Metamodel and Acoustic Model for Dysarthric Speech Recognition, Journal of Multimedia , VOL. 4, NO. 4, August 2009.
[15] Hagan,M., Demuth, H., and Beale, M. 1996. Neural Network Design. MA:PWS Publishing Company, Boston.
[16] Ham, F., and Kostanic, I. 2001. Principles of Neurocomputing for Science and Engineering, McGraw-Hill Company, New York.
[17] Wael S., Khaled. D., Omar D. and Abdel-Rahman Q., Speaker Identification System-based Mel frequency and Wavelet Transform using Neural Networks Classifier, Submitted for publication.
[18] Meyer, Wavelets, Ed. J.M. Combes et al., Springer Verlag, Berlin, p. 21, 1989.
[19] P. Goupillaud, A. Grossmann, J. Morlet, Cycle-octave and related transforms in seismic signal analysis, Geoexploration, 23, 85-102, 1984-1985.