European Journal of Scientific Research
ISSN 1450-216X Vol.41 No.4 (2010), pp.515-525
© EuroJournals Publishing, Inc. 2010
http://www.eurojournals.com/ejsr.htm

Speaker Identification System-based Mel Frequency and Wavelet Transform using Neural Network Classifier

Wael Al-Sawalmeh
Department of Communication and Electronics Engineering, Philadelphia University, Jordan
E-mail: narin912007@yahoo.com
Tel: + 962- 79 678 6786

Khaled Daqrouq
Department of Communication and Electronics Engineering, Philadelphia University, Jordan
E-mail:haleddaq@yahoo.com

Omar Daoud
Department of Communication and Electronics Engineering, Philadelphia University, Jordan
E-mail: odaoud@philadelphia.edu.jo

Abdel-Rahman Al-Qawasmi
Department of Communication and Electronics Engineering, Philadelphia University, Jordan
E-mail: qawasmi@philadelphia.edu.jo

Abstract

In The robustness to noise in speaker identification systems is improved by applying Continuous Wavelet Transform (CWT). In this work, essential speaker features are used to investigate the identification accuracy in non-stationary signals. These features are extracted using Mel Frequency Cepstral Coefficients (MFCC) and CWT for speech signals. In order to classify extracted features, a Feed Forward Back Propagation Neural Network (FFBNN) is imposed, since it gives better classification accuracy over conventional methods. A simulation program used to test the performance of the proposed method at certain level of SNR (-6dB), showed a classification ratio equal to 99.7%.

Keywords: Speaker identification; Continuous wavelet transform; Cepstral coefficients; and neural network.

1. Introduction

When a person speaks the lungs work like a power supply of the speech generating system. The glottis supplies the input with the certain pitch frequency (F0). The vocal tract, which contains the pharynx, mouth and nose cavities, works like a musical instrument to generate a sound. In fact, different vocal tract characters or shapes would generate a different sound (wave). To form distinct vocal tract shapes, the mouth cavity plays an important role. The nasal cavity is often included as a part of the vocal tract system. The nasal cavity and the mouth cavity are connected in parallel. The vocal tract model is shown in Figure. 1.

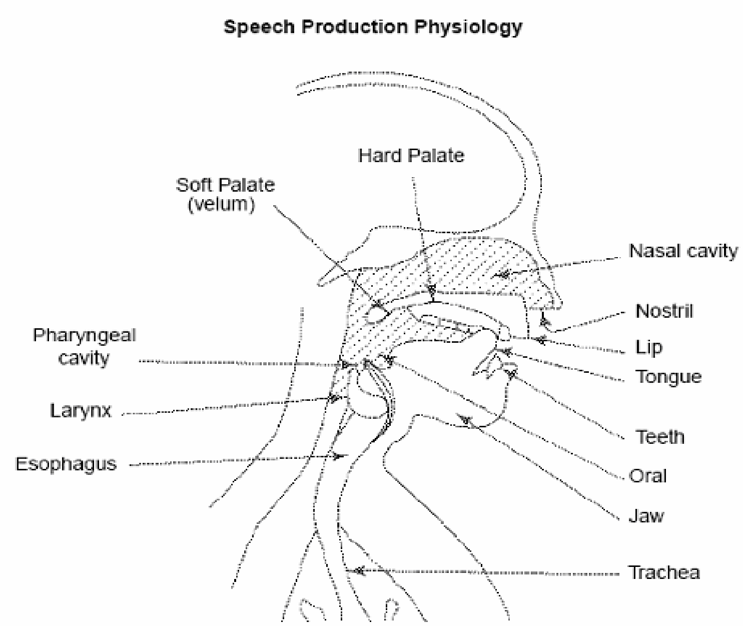
The glottal pulse produced by the glottis is used to generate vowels or sounds. And the noise-like signal is used to produce consonants or unvoiced sounds. Pitch frequency F0 = (1/T0) varies in different people. A child's pitch frequency is as high as 400 Hz. An adult male's pitch frequency is as low as 100 Hz while an adult female's pitch frequency varies from 200 Hz to 300 Hz [1, 23].

In a modern computerized globe, authentication and privacy methods are the primary means of identification for accessing different systems by way of passwords, identification IDs or pin numbers. Therefore the researchers have turned their interests to the proposition of good classifiers by which to access these services. During these developments, sophisticated systems and other natural solutions are created, discussed and debated every day as an alternative to the conventional authentication patterns.

Due to simplicity and being the object of interest particularly in mobile and telephony systems, the speech feature has attracted the researchers' attention [1-3].

Over last four decades many solutions for speaker recognition have appeared in various literatures [4-12]. The Al-Alaoui algorithm for pattern classification [4-7] was motivated by Patterson and Womack's [8] and Wee's [9] proofs that the Mean Square Error (MSE) solution of the pattern classification solution gives a minimum mean-square-error approximation to Bayes' discrimination weighted by the probability density function of the sample. All audio techniques start by converting the raw speech signal into a sequence of acoustic feature vectors carrying distinct information about the signal. This feature extraction is also called "front-end".

Figure 1: Vocal tract



The most commonly used acoustic vectors are MFCC [13, 14], linear prediction Cepstral coefficients [15-17], and perceptual linear prediction Cepstral coefficients. All these features are based on the spectral information derived from a short-time windowed segment of speech signals one of the most common short-term spectral measurements currently used is Linear Predictive Coding (LPC) derived from Cepstral coefficients and their regression coefficients. A spectral envelope reconstructed from a truncated set of Cepstral coefficients is much smoother than one reconstructed from LPC coefficients. Therefore, it provides a more stable representation from one repetition to another of a particular speaker's utterances. As for the regression coefficients, typically the first and second-order coefficients are extracted at every frame period to represent the spectral dynamics. These coefficients are derivatives of the time functions of the Cepstral coefficients and are respectively called the delta and delta-delta-Cepstral coefficients.

Text-dependent methods are usually based on template-matching techniques. In this approach, the input utterance is represented by a sequence of feature vectors, generally short-term spectral feature vectors. The time axis of the input utterance and each reference template or reference model of the registered speakers are aligned using a Dynamic Time Warping (DTW) algorithm and the degree of similarity between them, accumulated from the beginning to the end of the utterance, is calculated.

The Hidden Markov Model (HMM) can efficiently model statistical variation in spectral features. Therefore, HMM-based methods were introduced as extensions of the DTW-based methods, and have achieved significantly better recognition accuracies [18].

One of the most successful text-independent recognition methods is based on Vector Quantization (VQ). In this method, VQ codebooks consisting of a small number of representative feature vectors are used as an efficient means of characterizing speaker-specific features. A speaker-specific codebook is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker and the VQ distortion accumulated over the entire input utterance is used to make the recognition decision.

A method using statistical dynamic features has recently been proposed. In this method, a Multivariate Auto-Regression (MAR) model is applied to the time series of Cepstral vectors and used to characterize speakers [19].

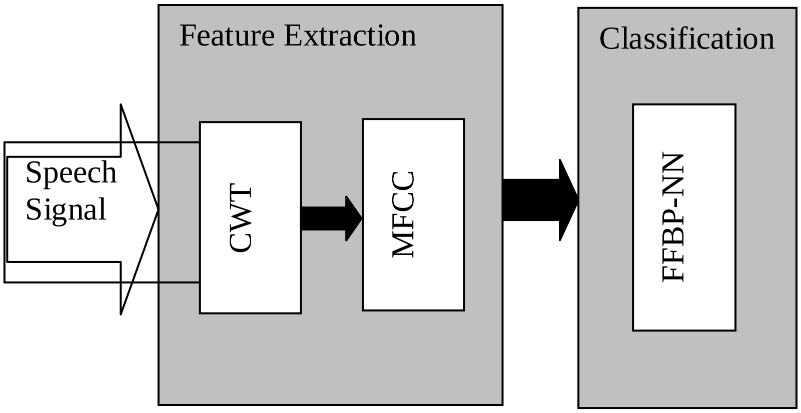
This work presents an investigation of utilizing the effect of CWT on system noise immunity. This system is divided into two main blocks; feature extraction and classification. The first part deals with speech signal feature extraction based on both CWT and MFCC. CWT depends on the convolution with wavelet functions that tracks the very quick variations in frequency changes in non-stationary signals. Then, MFCC is applied to the resulted signals from CWT to extract the desired features. The second part forms the classification process through a Neural Network.

The paper is divided into 4 sections. The proposed work and simulation results are depicted in Sections 2 and 3, respectively. After that, the conclusion is drawn in Section 4.

2. Proposed Work Main Parts

Figure 2 shows the main parts of the proposed work. The first block presents feature extraction that is accomplished by CWT and MFCC, while the second block presents the identification process via verification by FFBNN.

Figure 2: 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 is based on the speaker’s own feature frequency related to the vocal tract, and discarding other unwanted frequencies, such as high pass band frequencies which are more than 5 KHz. Morlet wavelets are chosen empirically, after investigation several wavelets in term of higher recognition rate,

The Morlet-Grossmann definition of the continuous wavelet transform [20-23] for a one dimensional 1D signal $f(x) \in L^2(R)$ is:

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

where $\psi^*(x)$ is the analyzing wavelet (Figure.3), the nonnegative value a is the scale parameter and b is the parameter that determines the position.

The transform is characterized by the following three properties:

1. It is a linear transformation,
2. It is covariant under translations:
$$f(x) \rightarrow f(x-u) \;\; W(a,b) \rightarrow W(a,b-u) \tag{2}$$
3. It is covariant under dilations:
$$f(x) \rightarrow f(sx) \; W(a,b) \rightarrow s^{-\frac{1}{2}} W(sa,sb) \tag{3}$$

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

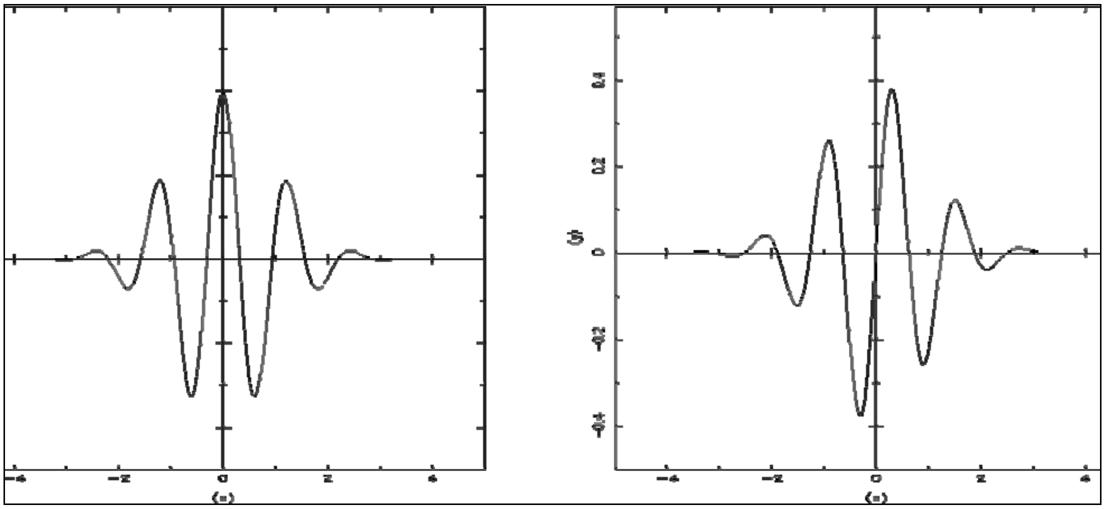
Now consider a function $W(a,b)$ which is the wavelet transform of a given function f(x). It has been shown that f(x) can be restored using the formula:

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

where:

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

Figure 3: Morlet's wavelet: real part at left and imaginary part at right



The CWT scale-determination is a very challenging problem because of its non-stationary nature, which is contained in speech signals. Therefore, the applied scale is chosen experimentally by studying a large database of approximately1000 speech signals. This assists greatly in creating a scale which matches all the speech signals within the database. To investigate whether or not the proper CWT scale is chosen, the PSD could be used as shown in Figures 7 and 8. The effect of using WT on the energy concentration can be seen in Figure 9 using a spectrogram.

Figure 6: Standard deviation of cepstral coefficients of two speakers

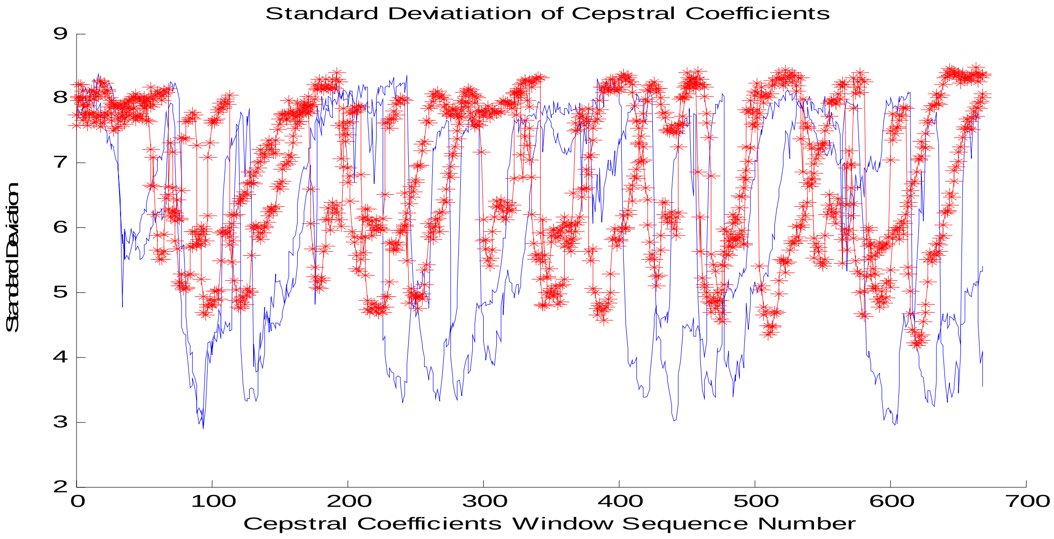


Figure 7: PSD coefficients of two speech signals of speaker1

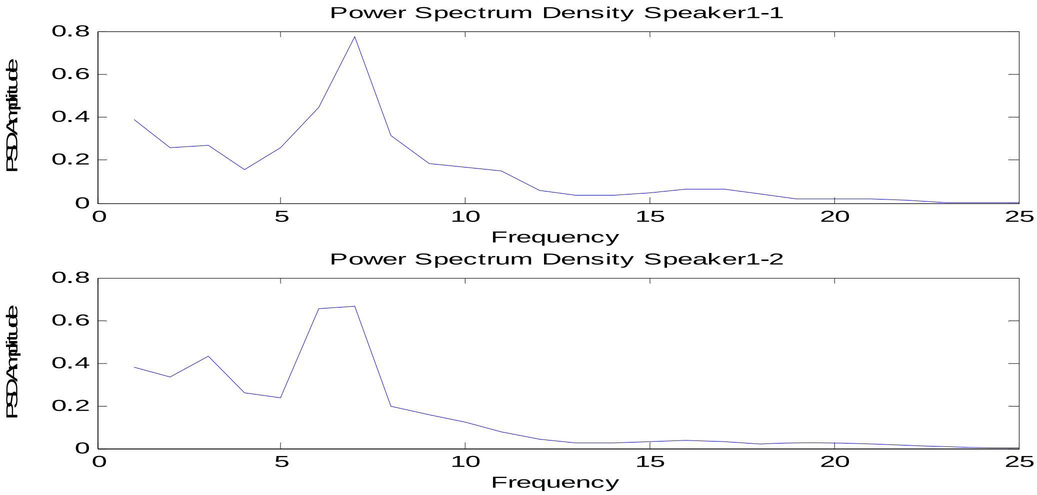


Figure 8: PSD coefficients of two speech signals of speaker2

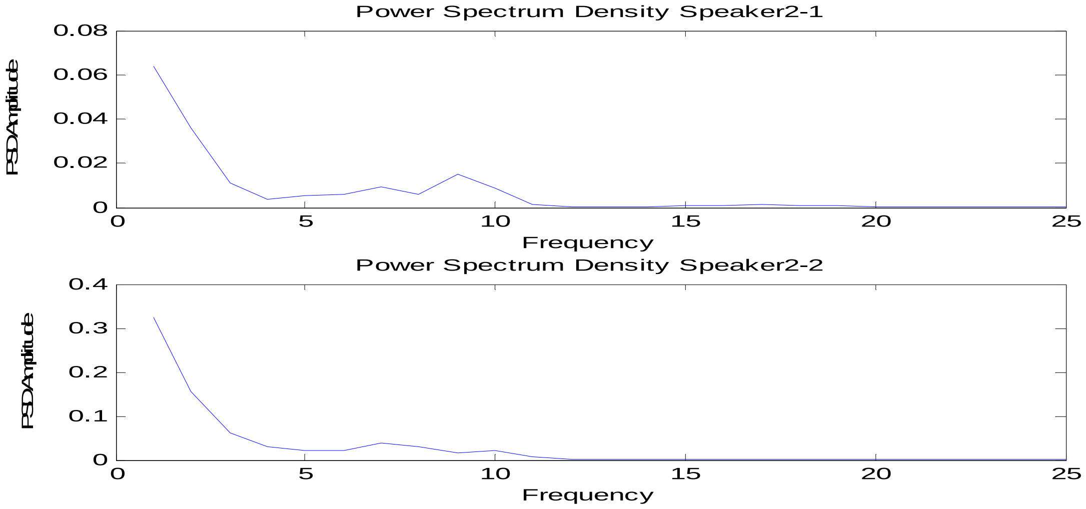
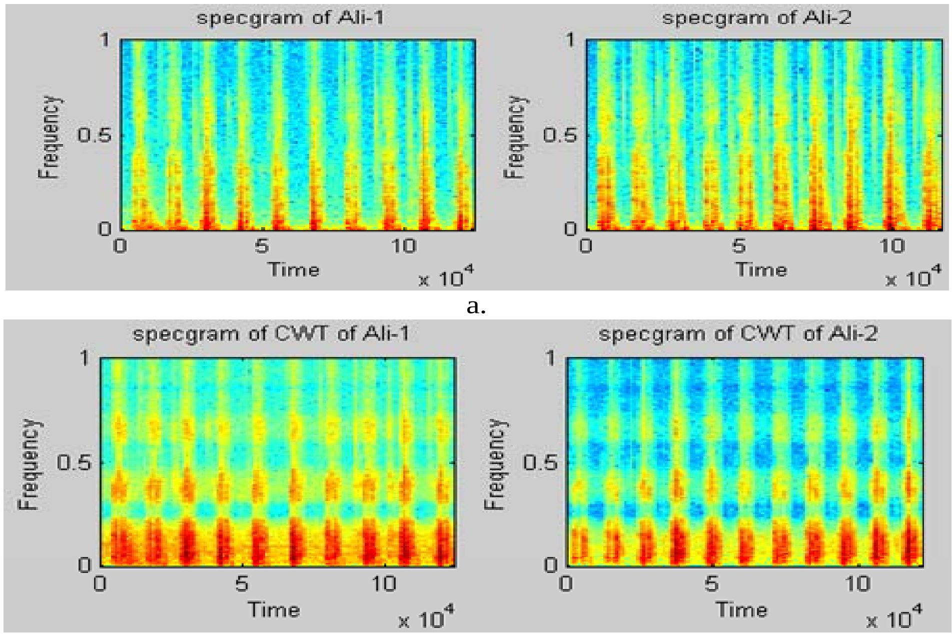


Figure 9: Effect of WT on the spectrogram density. a. Ali two utterances without CWT and b. Ali two utterances with CWT.



a.

b.

The second stage of the feature extraction method is accomplished by applying the MFCC to the WT coefficients. MFCC has been exploited in different tasks, but primarily in speech and speaker recognition. MFCC is superior to most popular feature tracking methods such as Fast Fourier Transform (FFT), linear prediction coding or PSD. This is due to its capability of imitating the human ear process.

Specifically, the speech signal is decomposed into CWT sub-signals (d1, d2… dJ), where each is generated using (1) a particular scale or level (j=1,2…J), as mentioned above. This is accomplished by convolving the signal with the mother wavelet function. After that, the speech signal goes through the MFCC algorithm different stages:

The speech signal is first windowed with the Hamming window and the discrete Short Time Fourier Transform (STFT) is computed.

The magnitude is then weighted by a series of filter frequency responses where center frequencies and bandwidths match those of the auditory critical band filters. These filters follow the Mel scale whereby band edges and center frequencies of the filters are linear for low frequency and logarithmically increase with increasing frequency as shown in Fig. 3. These filters are called the Mel-Scale Filter (MSF) bank. Figure 4 shows the MSF bank with N triangularly shaped frequency responses, which approximate the actual auditory critical band filters that cover the 4 kHz range.

Using equation 7 to compute the log energy in the STFT that is weighted by each MSF frequency response, we get:

$$melf = 2595\log10\left(1+\frac{f}{7000}\right) \tag{6}$$

Finally, the Cepstral coefficients are calculated using the Discrete Cosine Transform (DCT) of the filter bank output. [24, 25].

Figure 4: A Mel-scale filter bank

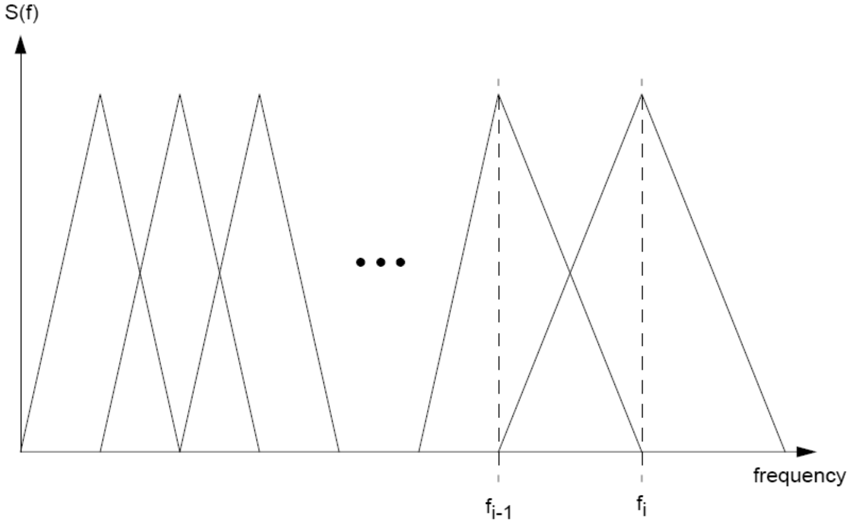
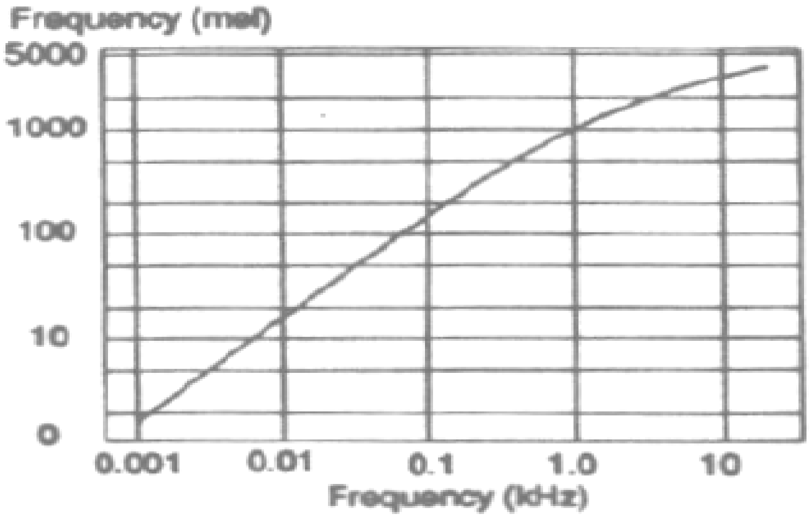


Figure 5: A mel-frequency



Roughly speaking, we start from firming the signal spectrum spatially using Hamming window then the resultant windowed signal goes through FFT and triangular filter bank. The logarithmic scale is applied to the output of the given filter bank to be processed through the DCT stage; this will produce the Cepstral coefficients.

Thirteen coefficients have been calculated for each window as follows:

$$MFCC_{x(n)} = \begin{bmatrix} C1 & C1 & \dots & C1_N \\ C2 & C2 & \dots & C2_N \\ & & . & . \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ C13 & C13 & \dots & C13_N \end{bmatrix} \qquad (7)$$

Where N is the number of widows. The window length is chosen as 12ms. To take matching decision, this matrix is given to a FFBNN to be trained with the following target

$$T = \begin{bmatrix} 1 & 0 & 1 & 0 & ... \\ 0 & 1 & 0 & 0 & ... \\ 0 & 0 & 1 & 0 & ... \\ 0 & 1 & 0 & 1 & ... \end{bmatrix} \qquad (8)$$

The second block is classification by FFBNN; to implement FFBNN we can use matlab neural network toolbox by function newff, tansig transfer function and trainlm back propagation training function:

net=newff(minmax($MFCC_{X(n)}$),[13 4],{'tansig''tansig'},'trainlm');

This commend builds a network of three layers: 13 neurons input layer, 13 neurons hidden layer and 4 neurons output layer. After training with the target by [net,tr]= train($MFCC_{X(n)}$,T);

we can simulate the network outputs (the weights and the biases) with each model stored in the system by T_result=round(sim(net, $MFCC_{X_{Model}(n)}$))>0.5;

Now error between impostor and the model is calculated by a=T_result-T;

Now classification or recognition rate between impostor and the model is calculated by RecognitionRate=(4*N-nnz(a))/ 4*N);

This action is repeated for each model. The maximum recognition rate for any of these models means that the imposter belongs to this model.

3. Results and Discussion

1000 speech signals as the testing speech signals database were used. The signals were recorded via a PC-sound card, with a spectral frequency of 4000 Hz and a sampling frequency of 16000 Hz, over about a 3 second time duration. Each speaker recorded the Arabic word "Eftah" which means "Open". Each utterance of eight "Eftah" words was recorded 4 times by the speaker.

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

Figure 6 illustrates the standard deviation values of 3344 windows, each window contains 13 MFCC coefficients calculated for four speech signals of two persons, where (-blue) coefficients present speaker 1 and (* red) coefficients present speaker 2. We observe that speaker 1's values are, definitely, in different ranges. This is due to the capability of MFCC to separate speaker features.

For speech signals identification via verification, Neural Networks were studied in terms of Feed Forward Back Propagation Neural Network FFBNN method. Table 1 shows the results of FFBNN of noisy speech signals of -6dB SNR. We observe that the use of CWT improves the recognition rate from 85% in the case where J=0 (no use of CWT) to 99.89 in the case where J=5, which is due to the use of CWT and MFCC. This result proves the system robustness by CWT.

Table 1: The effect of using WT on noisy signals recognition

-6 db SNR				
J	**Network**	**Training network**	**Transfer Function**	**Rate [%]**
0	FF	BP	Tansig	85
1	FF	BP	Tansig	89.9
2	FF	BP	Tansig	95.9
3	FF	BP	Tansig	89.9
5	FF	BP	Tansig	99.89
15	FF	BP	Tansig	95

In Table 2 the effect of the number of hidden neurons in the hidden layer on the recognition rate is studied. We are able to observe the ability of identification based on this algorithm. In this table the

Feed Forward (FF) NNT is used and trained by the Back-propagation (BP) network training function. The number of epochs was 200 and WT level was 8 for all cases of Different SNR ratios used. The Network was constructed of three layers: one input of 13 neurons, one hidden of 4 neurons and one output of 4 neurons. The maximum recognition rate was 95%. The same design is used in Table 3 with 13 hidden neurons.

The use of the hidden neurons number as the input neurons number can improve the recognition rate.

Table 4 presents RR results of four identification methods of different classification concepts (CC) calculated for 480 numbers of testing signals (NTS). The first one is the proposed method, the Continuous Wavelet Transform Neural Network CWTNN. The second one is the formants of the Neural Network method FNN [26], and final one is the Discrete Wavelet formants K-means WFDKM. The results show that proposed method is superior.

Table 2: Illustrates the effect of number of 4 hidden neurons at the recognition rate

J=8, 200 epochs and 4 hidden neurons				
SNR (dB)	**Network**	**Training network**	**Transfer Function**	**Rate [%]**
-4	FF	BP	Tansing	79.9
2.2				89.9
14				90
28				95

Table 3: Illustrates the effect of number of 13 hidden neurons at the recognition rate

J=8, 200 epochs and 13 hidden neurons				
SNR (dB)	**Network**	**Training network**	**Transfer Function**	**Rate [%]**
-4	FF	BP	Tansing	93
2.2				95
14				98.7
28				89.9

Table 4: Recognition rate (RR) results of four methods with Classification concepts (CC)

Method	**CC**	**j**	**NTS**	**NH**	**CR**
CWTNN	FFBNN	1	480	0	97.45
FNN	FFBNN	0	480	0	89.40%
WFDKM	k-means	1,2,…,5	480	18.7%	93.7%

4. Conclusions

In this paper, the effect of Wavelet Transform on speaker feature extraction is studied. The introduced system in this paper depends on two feature extraction stages; WT and MFCC due to its simplicity and better accuracy compared to linear prediction coding or FFT based method. The system works with the capability of features tracking even with 6 dB SNR, which is accomplished due to MFCC and WT feature extraction methods. NN classification method has been imposed in this work for the 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 a 99.87% identification rate was achieved in the proposed system.

References

[1] 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.

[2] Fernando L. Podio1 and Jeffrey S. Dunn2, Biometris from the movies, National Institute of Standards and Technology.

[3] Elisabeth Zetterholm, Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success. Lund University. (2003).

[4] M. A. Al-Alaoui, Some Applications of Generalized Inverse to Pattern Recognition, Electrical Engineering Department, Georgia Institute of Technology, December, 1974.

[5] M. A. Al-Alaoui, A New Weighted Generalized Inverse Algorithm for Pattern Recognition, IEEE Transactions on Computers, Vol. C-26, No. 10, pp. 1009-1017, October 1977.

[6] M. A. Al-Alaoui, Application of Constrained Generalized Inverse to Pattern Recognition, Pattern Recognition, Pergamum Press, Vol. 8, pp. 277-281, 1976.

[7] M. A. Al-Alaoui, J. El Achkar, M. Hijazi, T. Zeineddine and M. Khuri, Application of Artificial Neural Networks to QRS Detection and LVH Diagnosis, Proceedings of ICECS'95, pp. 377-380, Amman-Jordan, 17-21 December 1995.

[8] J. D. Ptterson and B. F. Womack, An Adaptive Pattern Classification System, IEEE Transactions Syst. Man. Cybern. , Vol. SSC-2, pp. 62-67, August 1966.

[9] W. G. Wee, Generalized Inverse Approach to Adaptive Multiclass Pattern Classification, IEEE Transactions on Computers, Vol. C-17, pp. 1157-1164,December 1968.

[10] Joe Campbell, Special Issue on Speaker Recognition, Digital Signal Processing, Science Direct, vol. 10, January 2000.

[11] R. Teunen, B. Shahshahani, and L. Heck, A Model-based Transformational Approach to Robust Speaker Recognition, ICSLP October 2000.

[12] A. Higgins, L. Bahler, and J. Porter, Speaker Verification using Randomized Phrase Prompting, Digital Signal Processing, vol. 1, pp. 89-106,1991

[13] Chakroborty, S., Roy, A. and Saha, G., Improved Closed set Text-Independent Speaker Identification by Combining MFCC with voidance from Flipped Filter Banks, International Journal of Signal Processing, Vol. 4, No. 2, Page(s):114-122, 2007.

[14] Sandipan Chakroborty and Goutam Saha, Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter., International Journal of Signal Processing 5, Winter 2009.

[15] Sanderson S. Automatic Person Verification Using Speech and Face Information, Griffith University. 2002.

[16] Petry A. and Barone D. A. C. Fractal Dimension Applied to Speaker Identification, ICASSP (Salt Lake City). May 7-11. 405-408, 2001.

[17] Liu C. H., Chen O. T. C. A Text-Independent Speaker Identification System Using PARCOR and AR Model, MWSCAS. Vol 3, 332-335. 2002.

[18] J. M. Naik, L. P. Netsch, and G. R. Doddington. Speaker verification over long distance telephone lines, IEEE Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing, Glasgow, Scotland, May 1989, pages 524--527.

[19] C. Griffin, T. Matsui, and S. Furui. Distance measures for text-independent speaker recognition based on MAR model, IEEE Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, April 1994, pages 309--312.

[20] D. Gabor, Theory of communication, Journal of I.E.E. 93 pp 429-441, 1946.

[21] P. Goupillaud, A. Grossmann, J. Morlet, Cycle-octave and related transforms in seismic signal analysis, Geoexploration, 23, 85-102, 1984-1985.

[22] A. Grossmann and J. Morlet, Decomposition of Hardy functions into square integrable wavelets of constant shape, SIAM J. Math. Anal, Vol. 15, pp. 723-736, 1984.

[23] Meyer, Wavelets, Ed. J.M. Combes et al., Springer Verlag, Berlin, p. 21, 1989.

[24] Dr. Joseph Picone, Fundamentals of speech recognition, Mississippi State University, MAY 15-17, 1996.

[25] S. Jothilakshmi, V. Ramalingam, S. Palanivel, Unsupervised speaker segmentation with residual phase and MFCC features, Expert Systems with Applications 36 (2009) 9799–9804.

[26] K. Daqrouq, Emad Khalaf, A. Al-Qawasmi, and T. Abu Hilal Wavelet Formants Speaker Identification Based System via Neural Network International Journal of Recent Trends in Engineering Vol. 2, No. 1, Nov 2009.