Speech Compression Using Wavelet Packet Best Tree Encoding (BTE)

Amr M. Gody, Safey A. Abdelwahab, Tamer M. Barakat and Mohamed Y. Mohamed

Abstract---Speech compression is one area of digital signal processing that focusing on reducing the bit rate of the speech signal for transmission or storage without significant loss of quality. This paper presents a new design feature for speech Compression using Wavelet Packet Transform and Linear prediction Coding. The proposed algorithm uses best tree decomposition of wavelet packets, which develop features afterward compressed signal is compressed by Linear prediction Coding. The 4 points encoded vector is a full of information just like the original best tree's structure. The implied scoring system makes BTE suitable for compression problems and we reach at 12 compression ratio. The performance of speech signal is measure on the basis of signal to noise ratio (SNR), mean square error (MSE) and peak signal to noise ratio (PSNR).

Keywords---Speech Compression; Wavelet Packet Transform; BTE; LPC

I. Introduction

SPEECH compression has been and still is a major issue in the area of digital speech processing. Speech compression is the act of transforming the speech signal at hand, to a more compact form, which can then be transmitted with a considerably smaller memory. The motivation behind this is the fact that access to unlimited amount of bandwidth is not possible. Therefore, there is a need to code and compress speech signals. Speech compression is required in long distance communication, high quality speech storage, and message encryption.

For example, in digital cellular technology many users need to share the same frequency bandwidth. Utilizing speech compression makes it possible for more users to share the available system. Another example where speech compression is needed is in digital voice storage. For a fixed amount of available memory, compression makes it possible to store longer messages [1-4].

Speech compression is a lossy type of coding, which means that the output signal does not exactly sound like the input. The input and the output signal could be distinguished to be different. Speech compression tries to code the speech in a perceptually lossless way. This means that even though the input and output signals are not mathematically equivalent, the sound at the output is the same as the input. This type of compression is used in applications for audio storage, broadcasting, and Internet streaming [5]. Several techniques of speech compression such as Linear Predictive Coding (LPC), Waveform Coding and Sub-hand Coding are existed.

The objective of this paper is to introduce new features for speech signal. Features are developed from the wavelet packets best tree decomposition of speech signal. This research aims to explain the proposed features in details. Also it targets to introduce the benefits of using the proposed feature in speech compression problems.

This paper is outlined as follows: In section II, the concept and motivation of wavelet transform are reviewed. Section III presents feature extraction. Section IV gives linear prediction coding. Finally, the results and the concluding remarks are given in section V and VI.

II. Wavelet Transform

Wavelets are short duration waveforms that can express any function by scaling and shifting of certain mother signal that is called mother wavelet [6]. Wavelet algorithm is acting as a filter banks on the input signal. The output of the filter banks are the wavelet signal's amplitudes.

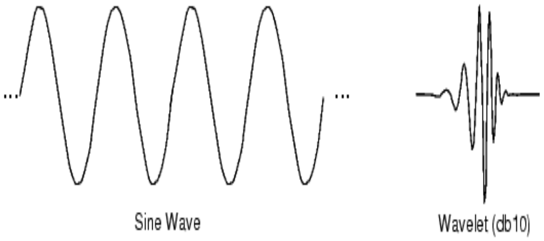

Fig 1. Sin Wave is used for Fourier Representation of the Signal While Wavelet Function is used in Wavelet Representation for Daubechies 10 Pointes Filter. Sin Wav is Infinite in Time but Finite in Frequency Domain While Wavelet is Finite in Both Time and Frequency Domains.

Figure 1 indicates a very important property of wavelet function. Wavelet function is a finite in time. It is also finite in frequency [7]. This is not the case of "Sine" basis functions (harmonic functions) used for Fourier analysis. All derived wavelets are orthogonal. This makes each wavelet acts as an identifier of the signal in a certain band. Figure 2 gives a brief

Manuscript received on August 30, 2013, review completed on September 25, 2013 and revised on October 11, 2013.

Amr M. Gody, Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt. (phone: 084-6337580; fax: 084-6334031; E-Mail: amg00@fayoum.edu.eg).

S. A. S. Abdelwahab, Engineering Department, Nuclear Research Center, Atomic Energy Authority, Egypt. E-Mail: safeyash@yahoo.com.

T. M. Barakat, Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt. E-Mail: tmb00@fayoum.edu.eg.

M. Y. Mohamed, Engineering Department, Nuclear Research Center, Atomic Energy Authority, Egypt E-Mail: myehiaa@yahoo.com.

Digital Object Identifier: DSP092013004.

0974-9705/CIIT–IJ-4643/06/$20/$100 © 2013 CiiT Published by the Coimbatore Institute of Information Technology

comparison between different possible spaces to express certain function [8].

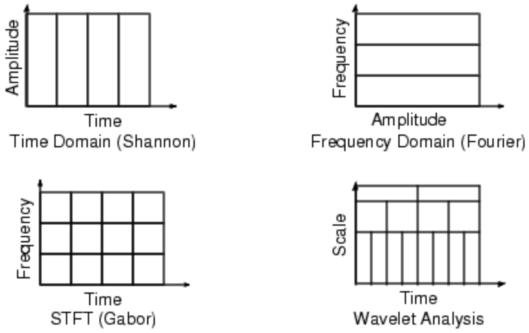

Fig 2. Comparison between different signal spaces

Wavelet packets are an extension to wavelet transform. It includes the high frequency parts in the analysis for more signal resolution of the frequency spectrum as shown in figure 3.

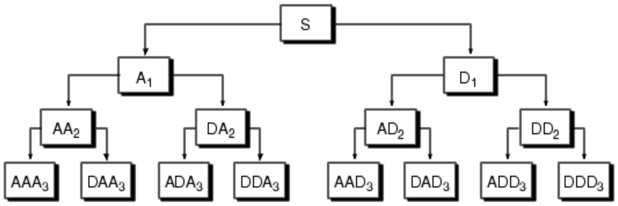

Fig 3. Signal Decomposition using Wavelet Packets.

To simplify the subject, let us discuss Fourier series as a signal representation tool.

$$f(x) = a_0 + \sum_{n=1}^{\infty}(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L} \tag{1}$$

Equation 1 indicates the Fourier series representation of function **f(x).** By the same approach, **f(x)** may be expressed using wavelet packets as in equation 2.

$$f(x) = \sum_j \sum_{n=0}^{2j-1} b_{jn} W_{j,n}(x) \tag{2}$$

Where b is wavelet coefficients and *W* is wavelet packet. Let us start with the two filters of length 2N, where h(n) and g(n), corresponding to the wavelet filters.

$$W_{2n}(x) = \sqrt{2}\sum_{k=0}^{2N-1} h(k) W_n(2x-k) \tag{3}$$

$$W_{2n+1}(x) = \sqrt{2}\sum_{k=0}^{2N-1} g(k) W_n(2x-k) \tag{4}$$

g(k) and h(k) are filter banks. Where:
$W_0(x) = \phi(x)$ is called the scaling function.

$W_1(x) = \psi(x)$ is called the wavelet function.

Where:

$$W_{jnk}(x) = w_n(2^{-j}x - k) \tag{5}$$
$$n \quad \varepsilon \quad N \quad and\,(j,k)\varepsilon \quad Z$$

K is not a dynamic parameter after the decomposition of the signal rather it is a constant value for each wavelet packet *W*. This makes it much better to abstract (5) as:

$$W_{j,n} = w_n(2^{-j}x - k) \qquad k\varepsilon Z \tag{6}$$

Hence:

$$W_{0,0}(x) = \Phi(x-k) \tag{7}$$

$$W_{1,1}(x) = \Psi(\frac{x}{2} - k) \tag{8}$$

The idea is explained by figure 4. Scaling "ϕ" and wavelet " Ψ " functions are used to generate *W* functions that cover all the frequency-scale space. The parameter k is used to indicate the time location of certain *W* function. *K* is chosen to best fit the original function to be expressed by wavelet packets while the scaling and wavelet functions are designed such that all W functions to be orthogonal.

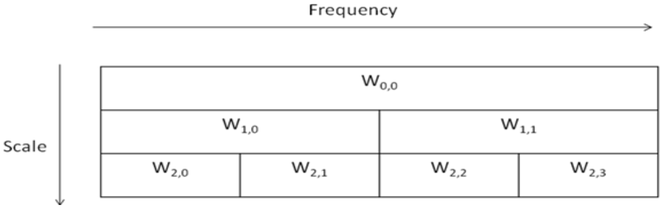

Fig 4. Frequency-Scale Space for Wavelet Packets.

Many researchers deal with the best way to optimize the full binary tree in such that best describe the contained information [9]. Different entropy functions may be used in such optimization [10]-[11].

III. FEATURE EXTRACTION

In this section the process of feature extraction using Best Tree 4 point Encoded features (BTE) will be explained. Wavelet packets process is very similar to filter banks. Both of them are filter banks in nature. The wavelet packets method is a generalization of wavelet decomposition that offers a richer signal analysis. Wavelet packet atoms are waveforms indexed by three naturally interpreted parameters: position, scale (as in wavelet decomposition), and frequency. For a given orthogonal wavelet function, we generate a library of bases called wavelet packet bases. Each of these bases offers a particular way of coding signals, preserving global energy, and reconstructing exact features. The wavelet packets can be used for numerous expansions of a given signal. We then select the most suitable decomposition of a given signal with respect to an entropy-based criterion [12].

The first step in BTE is to align the neighboring bands. This is very important for a good scoring process. Scoring process tries to score adjacent bands in such that minimizing the distance. For our case of best tree by Matlab, adjacent bands are indexed not in sequence.

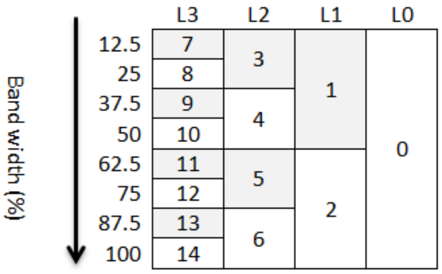

Fig 5. Wavelet Packet Tree Analysis Chart to Figure Out Adjacent Bands.

The objective is to remap node indices in such that adjacent node indices lay in adjacent frequency bands. To explain this subject considers the following table that represents the indices in a typical wavelet packet tree for 4-levels decomposition. Figure 5 represents band indexes in Matlab wavelet packets for 3 levels decomposition. Node indices are written inside the boxes that represent the nodes in the wavelet tree decomposition. As shown in figure 5 that node 7 and node 6 are too far in frequency while they are subsequent nodes as wavelet packets indexing system. This problem needs to be altered in such that adjacent frequency bands are listed as contiguous numbers. This way we will ensure that indexing system reflects frequency scale. This property may be used in the scoring system. Information in figure 5 is tabulated in table I to make it simple to figure out adjacent bands. Traversing tree as Left → Right → Center will be very logical to make good criteria for adjacency. Figure 6 explains the new indexing system.

Now we are ready to apply the best tree algorithm to optimize the full binary tree shown in figure 6. The optimization minimizes the number of tree nodes such that it best fit the information included in the speech signal. The entropy is used in the optimization algorithm.

TABLE I
BANDWIDTH DISTRIBUTION OVER WAVELET PACKET DECOMPOSITION BANDS.

Filter bank's Upper Limit with respect to total bandwidth (%)	Filter Bank's Node- index according to wavelet packet indexing system
100	0
50	1
100	2
25	3
50	4
75	5
100	6
12.5	7
25	8
37.5	9
50	10
62.5	11
75	12
87.5	13
100	14

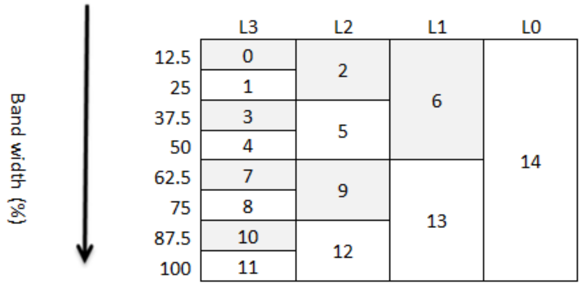

Fig 6. Proposed Indexing to Solve the Adjacency Problem due to Wavelet Packet's Indexing System.

The number is formed such that it reflects the tree structure within the cluster. Trees that cover the same bands will be almost adjacent trees. This property will be utilized in the scoring system. By considering all clusters, a vector of 4 components will be formed. Each vector's component represents a certain cluster. And each cluster covers a certain area in the total bandwidth. This is the 4 point encoded method that construct BTE features vector.

Figure 7 introduces a simple example to explain features encoding for a frame of speech signal. Circles in figure 7 represent leave nodes in the best tree decomposition.

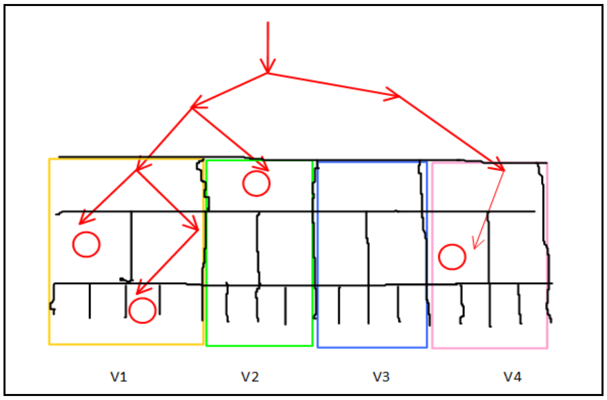

Fig 7. Proposed Indexing to Solve the Adjacency Problem due to Wavelet Packet's Indexing System.

The indicated tree structure in figure 7 will be encoded into features vector of 4 elements as shown in table II.

TABLE II
BEST TREE 4 POINT ENCODING EVALUATION.

Element	Binary Value	Decimal value	Frequency Band
V1	0001100	12	0 - 25 %
V2	1000000	64	25% - 50%
V3	0000000	0	50%-75%
V4	0000100	4	75%- 100%

IV. LINEAR PREDICTION CODING

In modern signal processing, the analysis procedure extracts useful information from the structure of a signal. LPC technique is a developed algorithm used in speech analysis for many years. Its basic idea comes from a model representing the resonances of the human vocal tract. In general, speech sounds are produced by acoustic excitation of the vocal tract. During the production of voiced sounds, the vocal tract is excited by a series of nearly periodic pulses generated by the vocal cords. With unvoiced sounds, the excitation is provided by air passing turbulently through constrictions in the tract [13-16]. Fig. 1 shows the speech signal production model in which the speech synthesizer strongly depends on the estimation of a_p. Here, a_p are autoregressive parameters obtained from the linear prediction method, and provide better results to characterize human speech. The use of these parameters assumes the speech signal can be represented as the output signal of an all pole digital filter in which the excitation is an impulse sequence with a frequency equal to the pitch of the speech signal under analysis when the segment is voiced, or with noise when the segment is unvoiced [17-19].

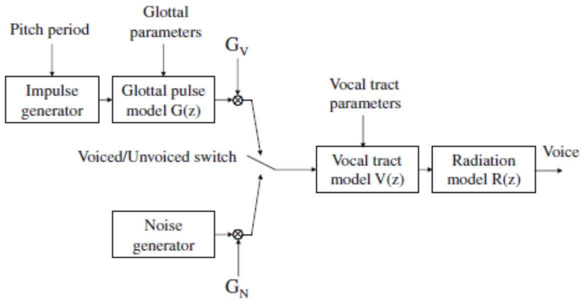

Fig 8. Flow Chart of Speech Synthesis.

After the speech signal is segmented, the *p* autocorrelation coefficients are estimated, where *p* is the linear predictor order. The autocorrelation function can be estimated using the biased or unbiased autocorrelation algorithms [20-23]. Once autocorrelation coefficients *p* are evaluated from each segment, the signal at time n can be rewritten as a linear combination from the pass samples of the input signal:

$$\hat{s}(n) = -(a_1 s(n-1) + a_2 s(n-2+\ldots+a_p(n-p)) \qquad (9)$$

Or

$$\hat{s}(n) = -\sum_{k=1}^{r} a_k s(n-k), \qquad k = 1,2,\ldots.. p. \qquad (10)$$

Therefore, it is affirmed a filter can be designed to estimate the data at time n only using the previous data at time n-1:

$$\hat{s}(n) = -a_e s(n-1), \qquad (11)$$

V. RESULT

The various speech signal sample is simulated using MATLAB. The speech signal is segmented in to packets with 640 samples. The Wavelet Packet Transform decomposition is computed for each packet. The transformed coefficients are extracted for further processing and the energy of the input signal is computed.

The results obtained for SNR, PSNR, NRMSE are calculated using the formulas:-

1) Signal to Noise Ratio :

This value gives the quality of reconstructed signal. Higher the value, the better:

$$\mathbf{SNR} = \mathbf{10}\log_{\mathbf{10}}\left(\frac{\sigma x^2}{\sigma e^2}\right) \qquad (12)$$

σx^2 is the mean square of the speech signal and σe^2 is the mean square difference between the original and reconstructed signals.

2) Mean Square Error :

$$\mathbf{MSE} = \frac{\mathbf{1}}{\mathbf{N}}\sum_{i=0}^{N-1}(x_i - r_i)^2 \qquad (13)$$

Where x_i is the original speech signal data and r_i is the reconstructed signal and N is the length of the reconstructed signal.

3) Peak Signal to Noise Ratio :

$$\mathbf{PSNR} = \mathbf{10}\log_{\mathbf{10}}\left[\frac{\mathbf{NX^2}}{\|\mathbf{x - r}\|^2}\right] \qquad (14)$$

Where X is the maximum absolute square value of the signal x

4) Compression Ratio (CR) :

It is the ratio of length of the original signal to the compressed signal.

$$CR = {}^{length(x(n))}\Big/_{Length(cr)} \qquad (15)$$

Where cr is the length of the compressed vector.

TABLE III
PERFORMANCE FOR MALE SPOKEN.

CR	SNR	MSE	PSNR
2.5515	-0.9807	0.0352	14.3086
3.1862	-0.9632	0.0350	14.3262
4.2412	-0.9399	0.0348	14.3495
6.3409	-0.8993	0.0345	14.3901
12.5574	-0.8284	0.0339	14.4610

TABLE IV
PERFORMANCE FOR FEMALE SPOKEN.

CR	SNR	MSE	PSNR
2.5515	-4.6337	0.0203	16.8416
3.1862	-4.5778	0.0200	16.8975
4.2412	-4.5078	0.0197	16.9676
6.3409	-4.4042	0.0192	17.0711
12.5574	-4.1998	0.0183	17.2755

TABLE V
PERFORMACE FOR MUSIC.

CR	SNR	MSE	PSNR
2.5515	-1.8229	0.0232	15.1822
3.1862	-1.7755	0.0230	15.2296
4.2412	-1.6934	0.0225	15.3117
6.3409	-1.6274	0.0222	15.3777
12.5574	-1.4804	0.0215	15.5247

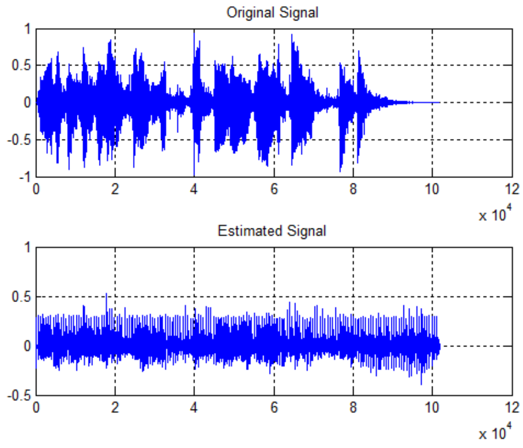

Fig 9. Output Waveform of Male Spoken.

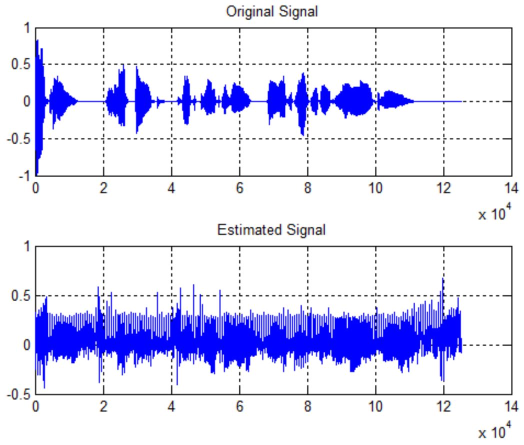

Fig 10. Output Waveform of Female Spoken.

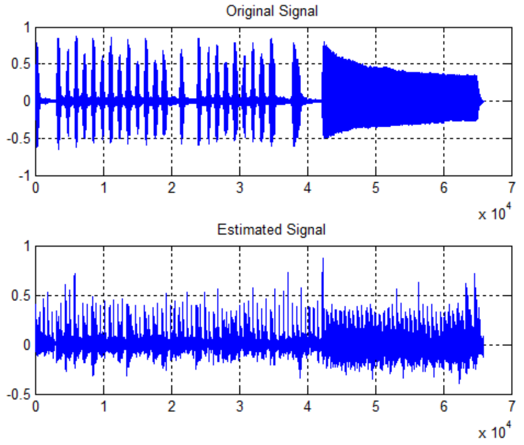

Fig 11. Output Waveform of Music.

VI. CONCLUSION

Wavelet packets make a similar processing on speech signal as the Filter banks method. It is much smarter than filter banks in that the number of filters is adapted by considering signal entropy to find the best tree. The problem of having dynamic size feature vectors is solved by considering the 4 points encoding algorithm. The proposed encoding system grantees that minimizing distance between feature vectors based on adjacency in frequency domain. This adjacency based on frequency domain of feature vectors distance calculation makes (BTE) features are highly promising in speech compression systems.

REFERENCES

[1] S. M. Joseph, "Spoken digit compression using wavelet packet," International conference on signal and image processing, PP: 255-259, 2010.

[2] M. A. Osman and N. Al, "Speech compression using LPC and wavelet," 2nd International conference on computer engineering and technology, PP: V7-92 – V7-99, 2010.

[3] R. W. Yeung, "A First Course in Information Theory," New York: Kluwer Academic/Plenum Publishers, 2002.

[4] J. Karam, "End point detection for wavelet based speech compression," International journal of biological and life sciences, PP. 167-170, 2008.

[5] R.V. Cox and P. Kroon, “Low bit-rate speech coders for multimedia communication”, IEEE Communications Magazine, pages 34-40, 1996. http://www.bell-labs.com

[6] MatLab,http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/ch06_a11.html.

[7] Gilbert Strang, "Wavelets and filter banks", Wellesley-Cambridge Press, ISBN: 0-9614088-7-1, pp. 37-86, ©1996.

[8] “A Tutorial of the Wavelet Transform” by Chun-Lin, Liu in February 23,2010.

[9] Coifman, R.R.; M.V. Wickerhauser (1992), "Entropy-based algorithms for best basis selection," IEEE Trans. on Inf. Theory, vol. 38, 2, pp. 713-718.

[10] Hai Jiang, Meng Joo Er and Yang Gao ," Feature Extraction Using Wavelet Packets Strategy", Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, Hawaii USA, December 2003

[11] http://en.wikipedia.org/wiki/Information_entropy.

[12] Coifman, R.R.; M.V. Wickerhauser (1992), "Entropy-based algorithms for best basis selection," IEEE Trans. on Inf. Theory, vol. 38, 2, pp. 713 718.

[13] Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. The Journal of the Acoustical Society of America, 50, 637–655.

[14] Sroka, J. J., & Braida, L. D. (2005). Human and machine consonant recognition. Speech Communication, 45, 401–423.

[15] Wu, J. D., & Lin, B. Fu. (2009). Speaker identification using discrete wavelet packet transform technique with irregular decomposition. Expert Systems with Applications, 36, 3136–3143.

[16] Wu, J. D., & Ye, S. H. (2009). Driver identification based on voice signal using continuous wavelet transform and artificial neural network techniques. Expert Systems with Applications, 36, 1061–1069.

[17] Perez-Meana, H. (2007). Advances in audio and speech signal processing: Technologies and applications. Hershey: IGI Global.

[18] Edler, B., “Coding of Audio Signals with Overlapping Block Transform and Adaptive Window Functions,” (in German), Frequenz, vol.43, pp.252-256, 1989.

[19] Q. Memon, T. Kasparis, “Transform Coding of Signals Using Approximate Trigonometric Expansions”. Journal of Electronic Imaging, Vol. 6, No. 4, October 1997, pp. 494-503.

[20] Childers, D. G. (2000). Speech processing and synthesis toolboxes. New York: John Wiley & Sons.

[21] Haydar, A., Demirekler, M., & Yurtseven, M. K. (1998). Speaker identification through use of features selected using genetic algorithm. Electronics Letters, 34, 39–40.

[22] Lou, X., & Loparo, K. A. (2004). Bearing fault diagnosis on wavelet transform and fuzzy inference. Mechanical System and Signal Processing, 18, 1077–1095.

[23] Lung, S. Y. (2006). Wavelet feature selection based neural networks with application to the text independent speaker identification. Pattern Recognition, 39, 1518–1521.

Amr M. Gody received the B.Sc. M.Sc., and PhD. from the Faculty of Engineering, Cairo University. Egypt, in 1991, 1995 and 1999 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 1994. He is a co-author of about 50 papers in national and international conference proceedings and journals. He is the Acting chief of Electrical Engineering department, Fayoum University in 2010, 2012 till now. His current research areas of interest include speech processing, speech recognition and speech compression.

Safey Ahmed Shehata Abdelwahab received his B.Sc. in Electronics and Communications form Engineering Faculty – Cairo University in 1992. He received his M.Sc. and Ph.D in Systems & Computers Engineering form Engineering Faculty – Al-Azhar University in 1998, 2003. His interests are: Digital Image and Digital Signal Processing, Fuzzy Logic, Design of Microcontroller based instruments, Design of radiation measurement instruments, Software programming for interfacing and data acquisition, Embedded Systems, Developing ICT- Based Materials, Design of FPGA based instruments, Computers.

Tamer M. Baraket received his BSc in communications and computers engineering from Helwan University, Cairo; Egypt in 2000. Received his MSc in Cryptography and Network security systems from Helwan University in 2004 and received his PhD in Cryptography and Network security systems from Cairo University in 2008. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt, in 2009. His main interests are: Cryptography and network security, Digital Image, and Digital Signal Processing. More specially, he is working on the design of efficient and secure cryptographic algorithms, in particular, security in the wireless sensor networks. Other things that interest him are number theory and the investigation of mathematics for designing secure and efficient cryptographic schemes.

Mohamed Y. Mohamed graduated from the Faculty of Engineering; He is now an MSc student. His interest is in signal processing and satellite system.