Stand-alone Intelligent Voice Recognition System

Mohammed R.Saady[1],Hatem El-Borey[2],El-Sayed A. El-Dahshan[2, 3], and Ashraf Shamseldin Yahia[2]

[1] *Basic Science Department, Faculty of Computers and Information, Fayoum University, El Fayoum, Egypt.*

mrs00@fayoum.edu.eg

[2] ***Physics Department, Faculty of Science, Ain Shames University, Abbassia 11566, Cairo, Egypt.***

[3] ***Egyptian E-Learning University EELU-33 Elmesaha Str. El-Doki, 12611, El-Giza, Cairo*****seldahshan@eelu.edu.eg**

Abstract

In this paper,an expert system for security based on biometric human features that can be obtained without any contact with the registering sensor is presented. These features are extracted from human's voice, so the system is calledVoice Recognition System (VRS).The proposed system consists of a combination of three stages: signal pre-processing, features extraction by using Wavelet Packet Transform (WPT) and features matching by using Artificial Neural Networks (ANNs). The features vectors were formed after two steps: firstly, decomposing the speech signal at level 7 with Daubechies 20-tap (db20), secondly, the energy corresponding to each WPT nodeis calculated which collected to form a features vector. One hundred twenty eight features vector for each speaker was fed to the Feed Forward Back-propagation Neural Network (FFBPNN). The data used in this paper are drawn from theEnglish Language Speech Database for Speaker Recognition (ELSDSR) database which composes of audio files for training and other files for testing.The performance of the proposed system is evaluated by using the test files. Our results showed that therate of correct recognition of the proposed system is about 100% for training files and 95.7% for one testing file for each speakerfrom the ELSDSR database. The proposed method showed efficiency results better than the well-known Mel Frequency Cepstral Coefficient (MFCC) and the Zak transform.

Keywords: Voice Recognition, Wavelet Packet Transform, Feature extraction, Artificial Neural Network.

1. Introduction

Everywhere around us are signals that can be analyzed. Signals are time-varying quantities which carry a lot of information. They may be audio signals, images or video signals, sonar signals or ultrasound, biological signals such as the electrical pulses from the heart, communications signals, or many other types. Audio signals carry an enormous amount about you.

In the last several years, several numbers of security systems based on fingerprints, voice, iris, face images, the pattern of the hands, and the pattern of motion of the arms have been presented. The security system based on iris or fingerprint is not convenient in practice since it impose stringent

requirements on interaction with the user;theusers of modern fingerprintbecome anxious about the hygiene of the process, and the users of modern irisface stringent requirements on movements and visibility of the eyes**[1]**.So there is a strong motivation to work in security system based on voice.

The security systems based on voice have been increased rapidly because of its non-contact characteristic**[2]**,and no two individuals sound identical since their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different**[3]**. So the Voice Recognition System(VRS) became one of the most useful and popular biometric recognition systems in the world especially in the areas in which security is a major concern. It can be used for authentication, surveillance, forensic speaker recognition and a number of related activities.

Voice Recognition (VR) which also called Speaker Recognition (SR)is the process of automatically recognizing who is speaking depending on the basis of individual information included in his/her voice signal. SR can be classified into Speaker Identification (SI) and Speaker Verification (SV). Speaker Identification is the process of determining which registered speaker provides a given utterance (one-to-many matching), on the other hand, SVis the process of accepting or rejecting the identity claim of a speaker (one-to-one matching), and this decision depends on the degree of similarity between the claimed speaker and the enrolled speaker.

In VRS, two basic operations are carried out: feature extraction and classification. In feature extraction we capture the vital characteristicwhich enables us to distinguish a speaker from the other. Some of techniques were used for extracting features areFourier Transforms (FTs), and Short-Time Fourier Transforms (STFTs). But these techniques aren't suitable for representing non stationary signals such as voice signals. By using (FTs) the input signal is transformed from the time domain into the frequency domain so we can know the frequencies that contained in the signal but not when, and STFTs tried to handle the problem by using a local window to display the frequency and time relation but its accuracy depended on the used window. Wavelet Transforms (WTs)have handled some of these problems. Classification is executed at two steps: enrollment and matching which often referred to them as training and testing respectively. In enrollment step the speaker is introduced to the systemby using the extracted features from the training data,and the testing step is activated when a vector of data from unknown speaker is entered to the system. It matches the vector data to a model corresponding to a known speaker. Classifications techniques have been used in VRS include Hidden Markov Models (HMMs) and Artificial Neural Networks (ANNs).

Several studies based on WTs as well as ANNs have been presented to design a SI system**[2]**,**[4]**.A text-independent speaker identification system based on the Zak transform**[5]**has been presented where the speech database was drawn from the ELSDSR database**[6]**.In this study we have tried to enhance the efficiency of the SI System which has already existed**[5]**.This enhancement can be summarized as: Wavelet Packet Transform (WPT) has been applied at the stage of feature extraction but these data were not suitable for classification as the performance of ANNs is depended on the size of the training data. The critical choice of subset of features from a larger set is very important to improve the performance of speaker recognition **[7]**.In our study, the energy corresponding to WP nodes have been captured as features vectors. These features vectors have been fed to Feed Forward Back-propagation Neural Network (FFBPNN) to train it.Finally, we have tested the trained FFBPNN with a single test file for each speaker.

The rest of the paper is organized as follows:Section 2 gives in brief an overview on the structure of a general VRS.Section 3 introduces the proposed VRS. In section 4 the results and discussions were presented. Finally, section 5 presented a conclusion and the future work.

2. Generic Voice Recognition System

Each VRS includes two phases: training and testing phases as graphed in Figure (1). In the training phase the speaker's voice print is set up from the features vector. These voice prints are stored in a reference database. While in the testing phase,features are extracted fromunknownspeaker's voice, then a pattern matching algorithm computes the similarity score between the unknown speaker's features vector and that stored in the reference database. Both the training and the testing phases include feature extraction so it is often called the front-end of the system.

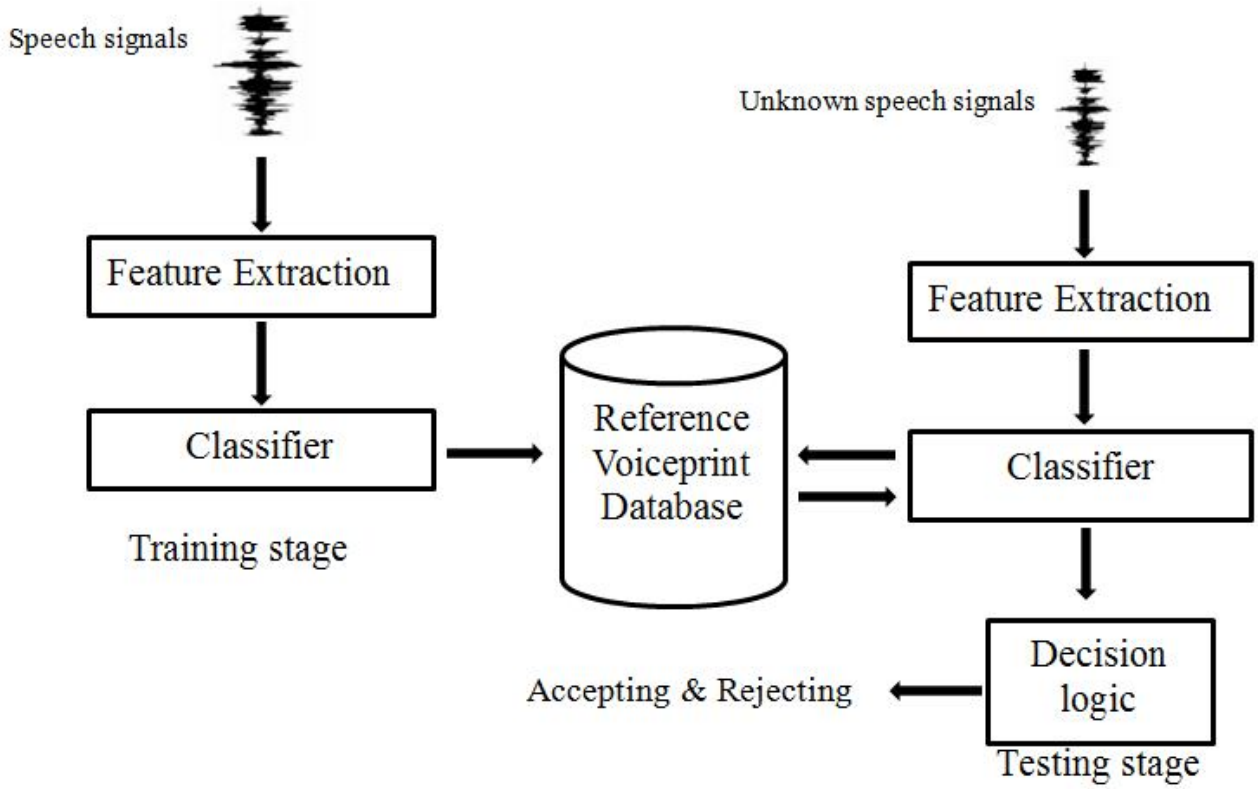

Fig1.*Typical voice recognition system.*

2.1 Feature extraction

Here we extract amount of features from the voice signal.The extracted features should be capable of separating the speakers from each other in its space. In traditional techniques speech features are extracted by FTs and (STFTs)**[2]**.But these techniques are not suitable for representing voice because they accept stationary signals within a given time frame therefore lack the ability of these techniques to analyze the non-stationary signals or signals in transient state **[8]**.One of the features obtained via FTs and used in the recognition taskwas the Mel Frequency Cepstral Coefficient (MFCC). But the recent researches showed that the identification rate with MFCCs can be as high as 99.5% for the noise free TIMIT database**[9]**.However, the identification accuracy reduced to 60% for the same data set that was transmitted over telephone channels. In 1984, The wavelet theory was proposed,Goupillaud et al. introduceda new transformation for the frequency analysis of the discretizedsignals. The transform is known as Wavelet Transform (WT)**[10],[11]**.It gives us the ability for decomposing any signal (analog or digital) of any field (medical, biometric, etc) into different resolution levels and displays two importantvariables (frequency & time)synchronously. Wavelets can be applied to many types of problems such as signal de-noising, compression and feature extraction.

The analysis by the Wavelet Transforms is performed by using a single wave is called a mother wavelet (often called window) Ψ(t) which can be considered as a band pass filter, has a limited duration of zero average, has irregular form, and often non-symmetrical unlike sinusoidal waves which extends from minus to plus infinity, and has a symmetrical form.The wavelet transform is defined as the inner product of a signal x(t) with the mother wavelet Ψ(t).

$$\Psi_{a,b}(t)=\psi\left(\frac{t-b}{a}\right)$$

$$W_{\Psi}\, x\,(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x\,(t)\,\psi^{*}\left(\frac{t-b}{a}\right)$$

Where a, b are the scale and translation parameters.

Depending on the scale factor a, the Wavelet Transforms have a large freedom degree to vary window size. This varying window size is wide for slow frequencies because the low frequency component completes a cycle in a large time interval, and it is narrow for the other high frequencies as the high frequency component completes a cycle in a much shorter interval**[12]**.Thus, an optimal time–frequency resolution is obtained in all frequency ranges.

2.2 Classification

After the features are extracted and saved for each speaker, these values are considered the threshold to enter the second operation of recognition, classification, whichmeans identification of the identity of unknown speaker by comparing the extracted features from his/her voice with the saved features (Reference Voiceprint Database), the reference which produces a maximum score of similarity is selected to yield a prediction for the speaker's identity.There are many classification tools such as Hidden Makrov Models (HMMs), Gaussian mixture model (GMM), andANNs which has an increasing interests in classification and used in this study. The advantages of ANNs for solving speech/speaker recognition problemsare their error tolerance and non-linear property**[13]**.

The ANNs is a mathematical model composed of a large number of simple processing units called neurons. The strategy of its work is inspired from the biological nervous system. Neurons are very lower than silicon logic gates, so the brain overcomes the relatively slow rate of operation by having a staggering number of neurons with massive interconnection between them. Neurons are connected to each other via synapses, and the strength of it is determined by the weight value it possess, each neuron as graphed in figure (2) has an adder for summing the weighted input signal, and the output of the neuron is limited by the neuron's activation function which the total input of it is controlled by an externally applied bias.

$$U_k = \sum_{j=1}^{m} w_{kj}\, x_j$$

Where$w_{kj}$is the synaptic weight directed from neuron j to neuron k, j is the number of inputs.

$$y_k = \Phi\,(U_k + b_k)$$

WhereΦ is the chosen activation function of the neuron, b_k is the bias.

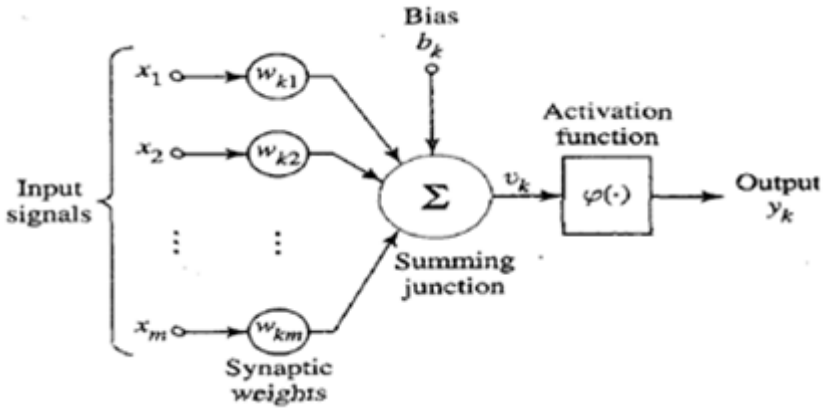

Fig 2.*Models of a neuron*

These neurons are arranged in layers one after the other, and the network architecture lies in one of three different classes; single-layer networks where we have an input layer of source nodes that projects onto an output layer of neurons, multilayer networks where there is one or more hidden layer, and the third class is the recurrent networks where there is at least one feedback loop.

Applications of ANNs included; pattern association, pattern recognition, function approximation, fitting, and beamforming. Our study belongs to pattern recognition task.

3. The proposed Voice Recognition System

The proposed system has been presented here is characterized by existing an operation before the feature extraction operation which included in any general VRS. This operation is called pre-processing, the purpose of it is to prevent the error estimation caused by speakers' volume changes, as see in figure (3).

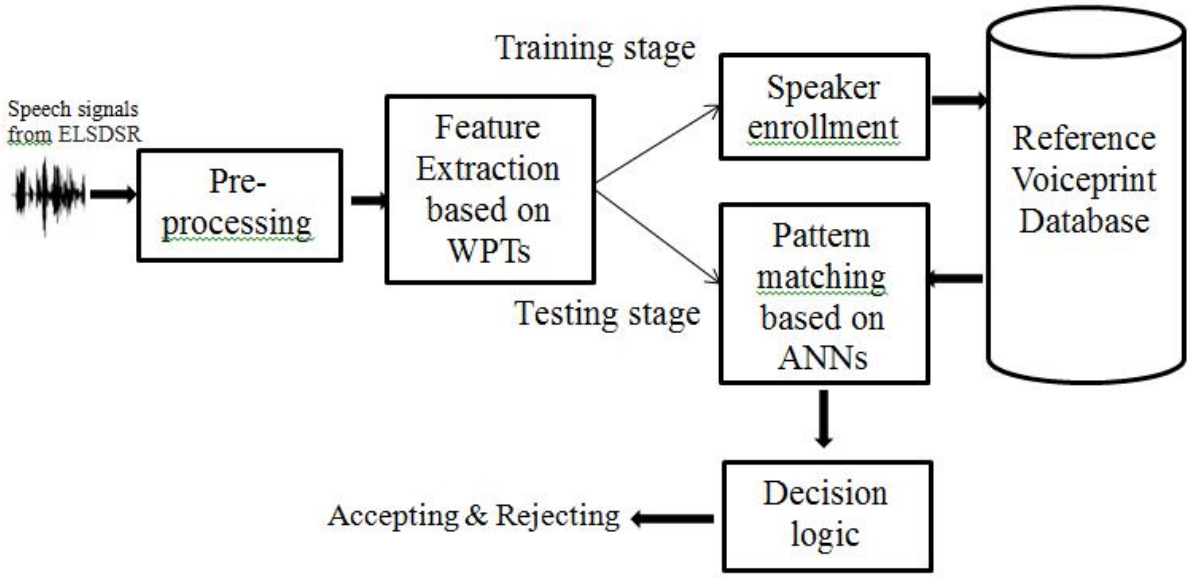

Fig3.*The proposed voice recognition system.*

3.1 database used in this study

The performance of the VRS which based on a standard speech database is comparable. There are 36 databases among free license such as the database which used here and proprietary bases that have been used in speaker recognition studies. The speech signals have been used in this study are drawn from the standard ELSDSR database. It is non-native speakers' database. It has been designed to provide speech data for the development and evaluation of automatic VRS. It has been designed by the Ph. D students and Master students from department of Informatics and Mathematical Modeling (IMM) at Technical University of Denmark. The speech language is English, and spoken by 21 Dane, one Islander and one Canadian. It contains voice messages from 23 speakers (13M/10F), and the age covered from 24 to 63. No a priori control of the speaker distribution by nationality and age has been done, except for the gender. Since the speakers were selected only in IMM, the speaker group exhibits relatively small variation in profession and educational background. The subjects of this database were from different countries and different places of one country, the dialect of reading English language in this database can probably be used as accent recognition. Part of the text, which is suggested as training subdivision, was made with the attempt to capture all the possible pronunciation of English language, which includes the vowels, consonants and diphthongs. With the suggested training and test subdivision, seven paragraphs of text are constructed and collected for training, which includes 11 sentences; and 46 sentences were collected for test text. In other word, for the training set, 161 (7*23) utterances were recorded; and for test set, 46 (2*23) utterances were provided. On average, the duration for reading the training data is:78.6s for male; 88.3s for female; and 83s for all. The duration for reading test data, on average, is: 16.1s (male); 19.6s (female); and 17.6s (for all) **[6]**.

3.2 Pre-processing of the speech signals

Signal pre-process have been carried out by performing normalization on the entered speech signals before feature extraction operation is carried to prevent the error estimation caused by speakers' volume changes. In other words, normalization makes the signals comparable regardless of differences in magnitude**[2]**.In this study, the signals are normalized by using the following equation**[14]**:

$$S_{pi} = \frac{Si - \mu}{\sigma},$$

Where$Si$is the ith element of the signalS, μand $\sigma$are the mean and standard deviation of the vector S, respectively;$S_{Pi}$is the ith element of the signal series S_P after normalization.

3.3 Feature extraction by Wavelet Transforms

In FTs, the signal has been converted from the Time-domain into the Frequency-domain. In feature extraction we need to analysis the signal several times by using variable filters to extract a vital feature. As there are several languages, it is hard to choose a feature which carries a lot of information about each speaker but we investigated to extract a vital feature. This can be done by using STFT which contribute to calculate the MFCCs which is the familiar parameter in the speech recognition task. But MFCCs failed in the in the background noise as mentioned before. The wavelet transforms proved that it is an effective signal processing technique for a variety of signal processing problems. Here we decomposing the signalby using a suitable wavelet transform version and then calculate the energy enclosed in each signal as a distinguished feature for each speaker.

3.3.1 the discrete Wavelet Transform

In Discrete Wavelet Transform (DWT), the signal is decomposed into Detail "D" and Approximation "A" coefficients in a dyadic form. This can be considered Multi-Resolution Analysis (MRA). In DWT, the scale and translation parameters are given by

$$a = a^{j}$$
$$b = kba^{j}$$

Where k and j are integers.The function family becomes

$$\Psi_{j,k}(t)=a^{-j/2}\Psi(a^{-j}t - kb)$$

We can imagine that signals transformed by DWT as putting data into a series of high-pass and low-pass filters g(n), h(n) respectively. The high frequency content of speech signals which passed through the g(n)filter is called "details", and the low frequency content which passed through the h(n) filter is called "approximations", and only the approximation coefficients can be decomposed further as the decomposition level grows. The wavelet decomposition of the signal S analyzed at level j has the following structure [cAj, cD1, . . . ,cDj]. DWT has been used at VRS but with limited success because of its left recursive nature**[2]**,**[7]**.

3.3.2 the Wavelet Packet Transform

In DWT, the high frequency components are removed, the voice will sound different but the speech can still be understood **[2]**. Wavelet packet analysis is an extension of the DWT **[15]**. Unlike the DWT, WavePT decomposes both the high and low frequency bands at each iteration. A pair of low pass and high pass filters is used to generate two sub-bands with different frequencies and this considered one level of decomposition. thesesub-bands are then down-sampled dyadically,. This process can be repeated to partition the frequency spectrum into smaller frequency bands for resolving different features while localizing the temporal information. The complete binary tree for 3-level decomposition is produced and shown in the Figure (4).

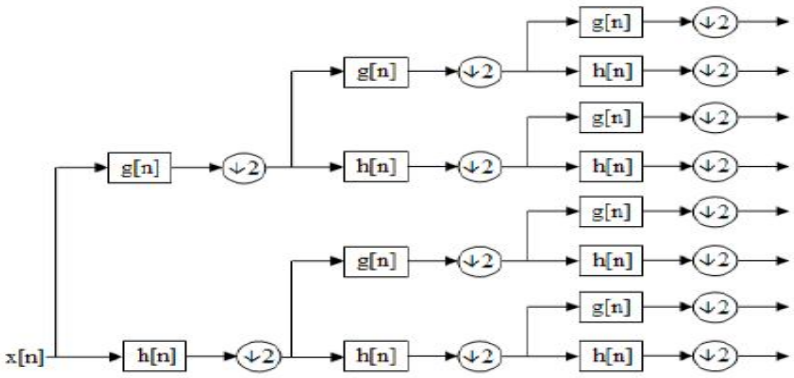

Fig 4.*3- levelWavelet Packet decomposition tree.*

In this study we have used WPT at level 7 with Daubechies 20-tap (db20), so we have divided the original signal into 128 sub-band signal. The Daubechies wavelet is a family of orthogonal wavelets, has a maximum number of vanishing moments, and conserves the energy of the signal and redistributed in a more compact form. The db20 is a member of the Daubechies family and is an estimation of the continuous form. It is asymmetric as shown in figure (5).

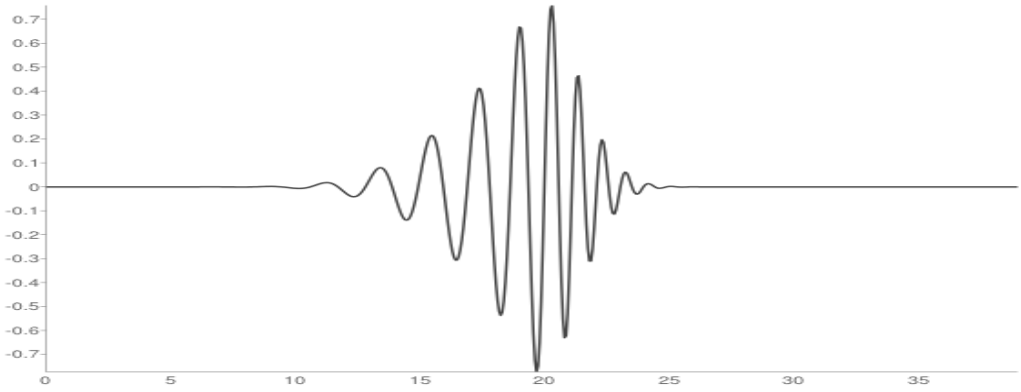

Fig 5.*The db20 wavelet.*

As mentioned above due to conservation of energy the energy corresponding to each WP node is calculated according to the following equation

$$\sum_{n=-\infty}^{\infty} (|x(n)|)2$$

to form the feature vector of size(1*128) which then used as input to the used Artificial Neural Network(ANN).

3.4 classification by Artificial Neural Networks

In the design of the Speaker Recognition System (SRS), a classification technique of the features which extracted by DWT and WPT was used to evaluate the effectiveness of the extracted features in differentiating between the speakers. In this study, the classifier has been used was ANNs. ANNs are adaptive model that can changes its structure depended on information passes through it during the training process. Here the FFBPNN is used which belongs to the multilayer networks. After several experiments with various network structures the highest accuracy was found with 128 neurons for the two hidden layers, and 23 neurons in the output layer. The number of training epochs was selected to be 1000. The learning rate was left to be variable. The error Back-Propagation learning algorithm was used. The FFBPNN is proposed in figure (6).

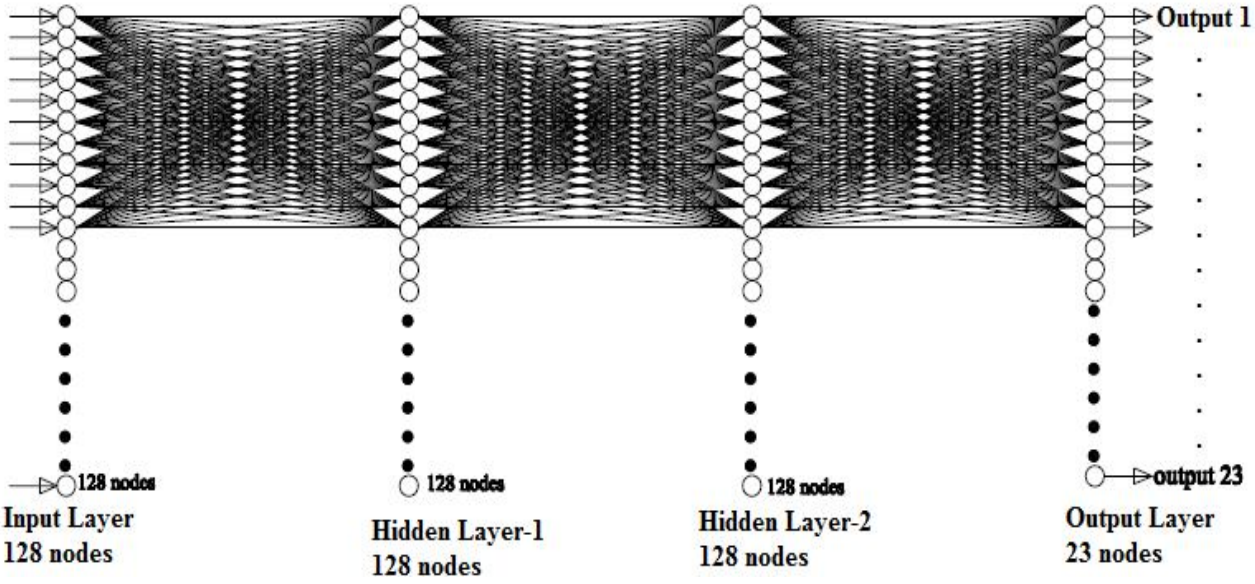

Fig 6.*The FFBPNN structure.*

The recognition rate is evaluated by plotting the confusion matrix. It contains information about actual and predicted classifications done by a classification system.it consists of columns and rows

where each column represents the predicted classes while each row represents the actual classes as proposed in figure (7).

		Predicted classes			
		Speaker #1	Speaker#2.....	Speaker #23	Accuracy%
Actual classes	Speaker #1	✓			
	Speaker #2		✓		
	:				
	:				
	:				
	Speaker #23			✓	
	Total				Average%

Fig 7.*The confusion matrix.*

4. Results and Discussion

Acomputer with specifications of processor Intel® Core ™ 2 Duo CPU 2 GHZ, RAM 2 GB, an operating system windows 7 32-bit, and MATLAB (R2009b) program were used for achieving the experiment. Astandard database ELSDSRis used where each registered speaker speaks a certain seven sentences and each sentence has been said for one time. Thus for each speaker we had seven speech samples, and 161 (23*7) speech samples for the whole speakers. The experiment is carried out on three stages ranged from stage 0 to stage 2 and can be summarized as follows.

Stage 0, we had made normalization on each speech sample to prevent the error estimation caused by speakers' volume changes as see in figure (8).

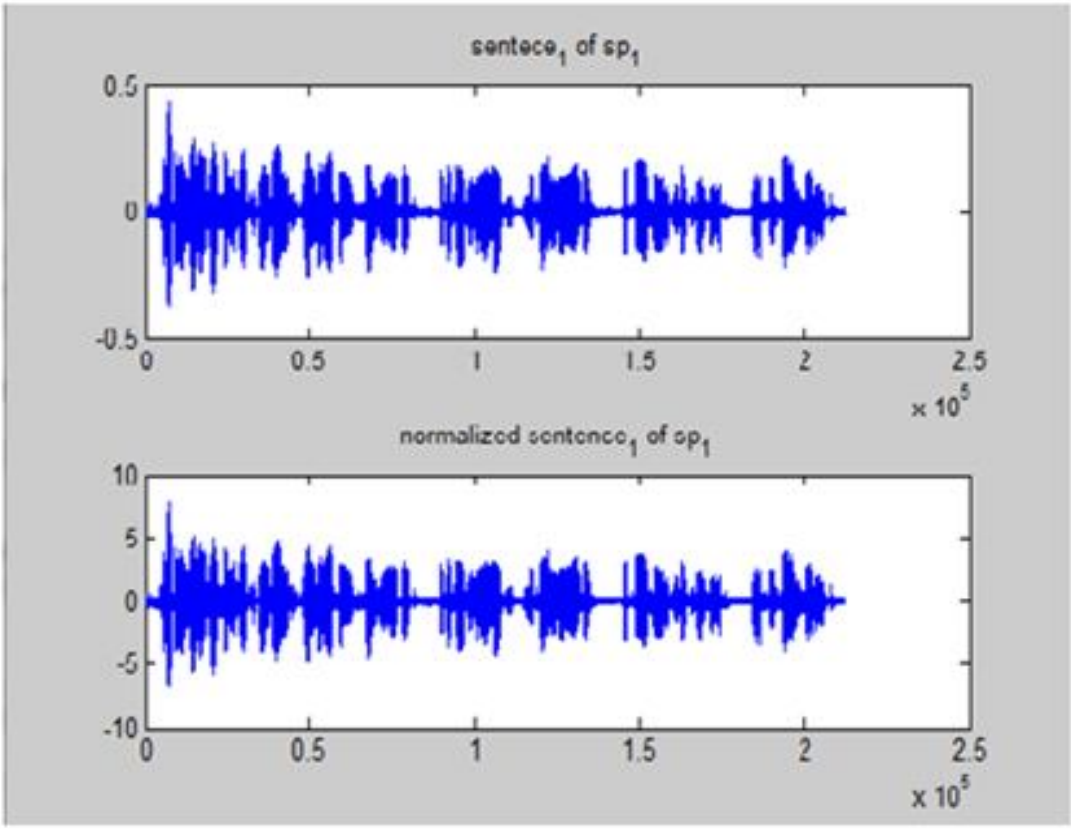

Fig 8.*The speech signal for the first speaker before and after normalization*

As noticed from figure (8) the signal values are redistributed in a wide range vertically at this instance the speech samples are ready to analysis them to extract the features vectors and enter the next stage.

Stage 1, we haveextracted the features vectors for each sentence for each speaker.In the features extraction step we have applied DWTand WPT but we have got a high performance rate when the speech samples were analyzed with WPT as it has characterized by the recursive analysis. We have used the Daubechies as mother wavelet and performed decomposition up to level 7 and the original speech sample is divided into 128 sub-band. We investigated on the better features that can represent each speaker, so calculation of the energy corresponding to each sub-band was the best choice as by it we can distinguish between the speakers' identities asproposed in figure (9).

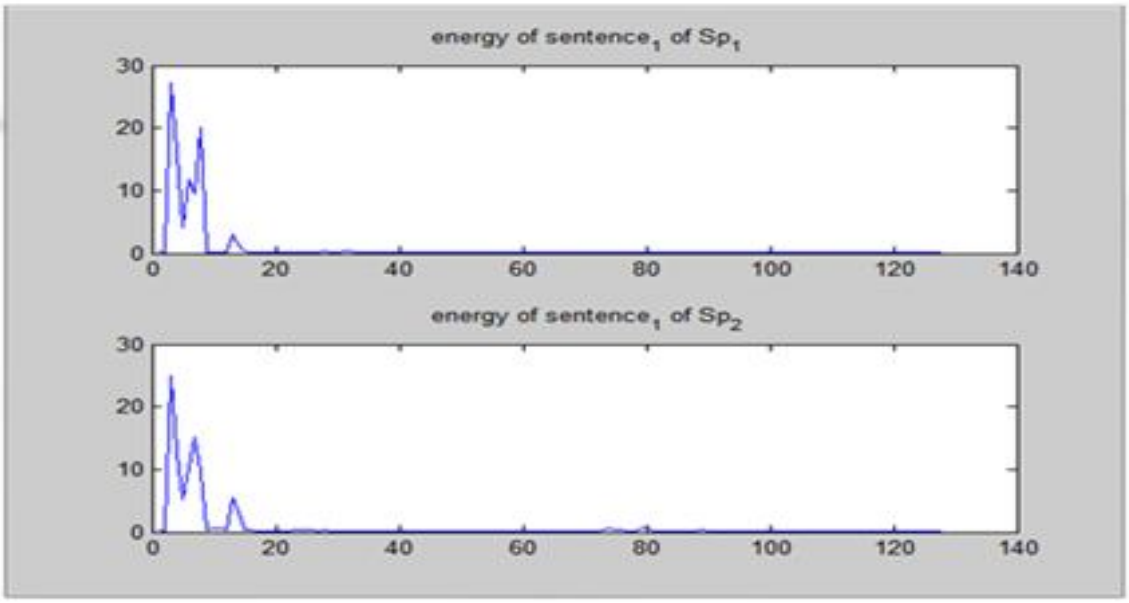

Fig9.*Energy distribution of sentence 1 for the two different speakers.*

As noticed from the preceding figure the energy feature can differentiate between the speakers. We have formed a (128*161) features matrix for the whole database's speakers. The features matrix was the threshold to enter the final stage in our experiment.

Stage 2,Classification is performed on two stages, in the training stage the features matrix and a suitable target matrix have been entered as inputsto the FFBPNN which specified in table (1) and is drawn in figure (6).

Table (1): description of the FFBPNN.

functions	Description
Network type	Feed Forward Back-Propagation
Number of Layers	Three layers: two hidden, one output
Number of neurons in each layer	128 neuron for each hidden, 23 neuron in the output, and 128 input node
Activation function	Tan-sigmoid
Training learning rule	Back-propagation
Number of epochs	1000
Learning rate	varied

The FFBPNN varies its weight depended on the input and the expected output which feed to it. The expected output which also called the target matrix has a size of (23*161). It is organized as for each speaker there is 7 columns (equals to the seven training sentences) with the first row has 1 values and the rest rows have 0 values, for the second speaker we bank the second seven columns with the second row in them has 1 values and the other rows have 0 values and so on until the twenty-three speaker for him the twenty-three row is filled with 1 values. Our target matrix can be summarized as in the table (2).

Table (2): the target matrix

Speaker #1							Speaker #2……..							Speaker #23							
1	1	1	1	1	1	1	0	0	0	0	0	0	0	…….	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	1	1	1	1	1	……..	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	……..	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	……..	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	……..	0	0	0	0	0	0	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.	……...	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	……...	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	……...	.	.	.	.	.	.	.
0	0	0	0	0	0	0	0	0	0	0	0	0	0	……..	1	1	1	1	1	1	1

Table (3) shows a summary of the confusion matrix which represents the accuracy of the FFBPNN for speakers individually as well as the system overall. Depending on the training database set, the system attempts to recognize 7 samples for each speaker where the total number of samples is 161.The overall system accuracy was 100%.

Table (3): recognition rate results at the Training stage

Speaker	**No. of signals**	**Recognized signals**	**Not Recognized signals**	**Recognition rate %**
Sp#1	7	7	0	100
Sp#2	7	7	0	100
Sp#3	7	7	0	100
Sp#4	7	7	0	100
Sp#5	7	7	0	100
Sp#6	7	7	0	100
Sp#7	7	7	0	100
Sp#8	7	7	0	100
Sp#9	7	7	0	100
Sp#10	7	7	0	100
Sp#11	7	7	0	100
Sp#12	7	7	0	100
Sp#13	7	7	0	100
Sp#14	7	7	0	100
Sp#15	7	7	0	100
Sp#16	7	7	0	100
Sp#17	7	7	0	100
Sp#18	7	7	0	100
Sp#19	7	7	0	100
Sp#20	7	7	0	100
Sp#21	7	7	0	100
Sp#22	7	7	0	100
Sp#23	7	7	0	100
Total	**161**	**161**	**0**	**100**

After getting on that accuracy in the training stage we saved the weights and biases of all layers of the trainedFFBPNN.the weight values from the second hidden layer to the output layer can be abbreviated in table (4).

Table (4): some of the weights values from the second hidden layer to the output layer

Column #1	**Column #2**	**Column #128**
-0.09268249182061265	-0.15119278911501377……..	0.20593362253011566
-0.06370861164394098	-0.21319245828066766……..	0.022286562111564632
-0.1717592648717832	-0.14805606367740934……..	-0.12241513724860374
-0.0431523394284899	-0.18449418630223063……..	-0.13357997743075695
0.09903762471988278	-0.05582782956379674……..	0.20524026699714845
.	……..	.
.	. ……..	.
.	. ……..	.
-0.13980827477698915	0.08471834362138755 ……..	0.1373001986105382

At the testing stage, we have created another feature matrix **(128*23)**where each speaker said another sentence differs from the seven sentences which used in the training stage and also differs from one speaker to another one, we have created the same FFBPNN with the saved weights and biases, and performed the test to get a recognition rate equal to 95.7%which recorded from the confusion matrix. Table (5) shows a summary of the confusion matrix in that stage. The system failed to recognize the six-teen speaker to become the overall recognized speakers equal to 22, and the overall accuracy of the system is 95.7%. Table (6) shows a comparison between MFCC, the Zak transform and the WPT techniques when all seven training files are used in the training phase. The WPTtechnique shows the same recognition rate 100% as the other two techniques if two test files are used and a recognition rate 95.7% if the first test file is used which better than the obtained by using the Zak transform 91.3%**[5]**.

Table (5): recognition rate results at the Testing stage

Speaker	**No. of signals**	**Recognized signals**	**Not Recognized signals**	**Recognition rate** %
Sp#1	1	1	0	100
Sp#2	1	1	0	100
Sp#3	1	1	0	100
Sp#4	1	1	0	100
Sp#5	1	1	0	100
Sp#6	1	1	0	100
Sp#7	1	1	0	100
Sp#8	1	1	0	100
Sp#9	1	1	0	100
Sp#10	1	1	0	100
Sp#11	1	1	0	100
Sp#12	1	1	0	100
Sp#13	1	1	0	100
Sp#14	1	1	0	100
Sp#15	1	1	0	100
Sp#16	1	**0**	**1**	**0**
Sp#17	1	1	0	100
Sp#18	1	1	0	100
Sp#19	1	1	0	100
Sp#20	1	1	0	100
Sp#21	1	1	0	100
Sp#22	1	1	0	100
Sp#23	1	1	0	100
Total	**23**	**22**	**0**	**95.7**

Table (6): Comparison with MFCC, Zak in terms recognition rate

Test Files	All Training Files	All Training Files	All Training Files
	MFFC	Zak	WPT
File 1	100%	91.30%	95.7%
Both	100%	100%	100%

5. Conclusion and Future Work.

In this paper an intelligent VRS using WPT and FFBPNN is presented. A new method for feature extraction based on energy distribution of WP tree nodes and classification based on the FFBPNN is presented. The obtained results showed that the proposed method can make an effective analysis. The average identification rate was 95.7 % better than another study 91.3%**[5]** on the same database where the recognition rate was 91.3% by using a single test file published before.In the future work we will try to reduce the number of features associated with each speaker, and try to improve the efficiency of the proposed system in the real-time where the noise exist such as the street by using a hardware circuit such as the FPGA circuit.

6. References

[1]Desyatchikov, A. A., Kovkov, D. V., Lobantsov, V. V., Makovkin, K. A, Matveev, I. A., Murynin A. B., and Chuchupal V. Ya.**,** (2006)."A System of Algorithms for Stable Human Recognition".*Journal of Computer and Systems Sciences International45: 958–969.*

[2]Jian-Da Wu, Bing-Fu Lin, (2009). "Speaker identification using discrete wavelet packet transform technique with irregular decomposition".*Expert systems with applications 36:3136-3143.*

[3]TomiKinnunen, Haizhou Li,(2010)."An overview of text-independent speaker recognition: From features to supervectors".*Speech Communication 52:12–40***.**

[4]Emad F. Khalaf, KhaledDaqrouq and Mohamed Sherif, (2011)."Wavelet Packet and Percent of Energy Distribution with Neural Networks Based Gender Identification System". *Journal of applied Sciences 11: 2940-2946.*

[5]AbdulnasirHossen, and Said Al-Rawahi, (2010)."A Text-Independent Speaker Identification System Based on the Zak Transform".*Signal Processing an International Journal 4:68-74.*

[6]ELSDSR database for speaker recognition, (2004), http://www.imm.dtu.dk/~lf/eLSDSR.htm

[7]Shung Y. Lung, (2006). "Wavelet feature selection based neural networks with application to the text independent speaker identification".*Pattern Recognition39:1518 – 1521*

[8]Avci, E., and Akpolat, Z. H., (2006)."Speech recognition using a wavelet packet adaptive network based fuzzy inference system". *Expert Systemswith Applications, 31, 495–503.*

[9]Sarikaya, R., Pellom, B. L., and Hansen, J. H. L., (1998)."Wavelet packet transform features with application to speaker identification". *In Proceedings of the IEEE Nordic signal processing symposium (pp. 81–84). Denmark.*

[10] Goupillaud, P., Grossman, A., and Morlet, J., (1984). "Cycle-octave and related transforms in seismic signal analysis".*Geoexploration, 23, 85–102.*

[11]Louis, A. K., Maass, D., &Rieder, A. (1997). "Wavelets-theory and applications". *Hoboken, NJ: Wiley.*

[12]Avci, E., (2007). "A new optimum feature extraction and classification method for speaker recognition: GWPNN". *Expert system with applications 32:485-498.*

[13]Haykin, S. (1999)."Neural networks: A comprehensive foundation" *(2nd ed.).Englewood Cliffs, NJ: Prentice-Hall.*

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

[15]Burrus, C. S., Gopinath, R. A., and Guo, H. (1998). "Introduction to wavelet and wavelet transforms". *New Jersey, USA: Prentice Hall.*