Speech Recognition System Based on Wavelet Transform and Artificial Neural Network

Engy R. Rady[1], Ashraf H. Yahia[2], El- Sayed A. El-Dahshan[3], and Hatem El-Borey[4]

[1] Basic Science Department, Faculty of Computers and Information,
Fayoum University, El Fayoum, Egypt
[2] Physics Department, Faculty of Science, Ain Shams University,Cairo, Egypt
[3] Egyptian E-Learning University EELU, Cairo, Egypt
era00@fayoum.edu.eg, ayahia@sci.asu.edu.eg,seldahshan@eelu.edu.eg

Abstract

For the past several decades, designers have processed speech for a wide variety of applications ranging from mobile communications to automatic reading machines. Speech recognition reduces the overhead caused by alternate communication methods. Speech has not been used much in the field of electronics and computers due to the complexity and variety of speech signals and sounds. However, with modern processes, algorithms, and methods we can process speech signals easily and recognize the text. This paper presents an expert speech recognition system for isolated words based on a developed model of Discrete Wavelet Transform (DWT) and Artificial Neural Network (ANN) techniques to improve the recognition rate. The data set was created by using English digits from zero to five and other nine words (spoken words) which was collected from four individuals in various time intervals. The feature vector was formed by using the parameters extracted by DWT. We have employed Daubechies 4-tap (db4) wavelet for the experiment. The feature vector was produced for all words and formed a training set for classification and recognition. Forty-four features were feed to feed forward backpropagation neural network (FFBPNN) for classification.The performance of the developed system was evaluated by using speech signals. The rate of correct classification was about 98.9 % for the sample speech signals.

Keywords: *Discrete Wavelet Transform, Speech Recognition, Feature Extraction, Artificial Neural Network*.

1. Introduction

The speech signal is the fastest and the most natural method of communication between humans. This fact has motivated researchers to think of speech as a fast and efficient method of interaction between human and machine [1]. The significance of speech recognition lies in its simplicity. This simplicity together with the ease of operating a device using speech, has lots of advantages. It can be used in many applications like, security devices, household appliances, cellular phones, ATM machines, and computers [2]. Automatic speech recognition methods, investigated for many years, have been principally aimed at realizing transcription and human computer interaction systems [3].

Speech features which are usually obtained via Fourier Transforms (FTs), Short Time Fourier transform (STFTs), or Linear Predictive Coding (LPC) techniques are, used for some kinds of Automatic Speech/Speaker recognition (ASR). They may not be suitable for representing speech/voice. These methods accept signal stationary within a given time frame and may therefore lack the ability to analyze localized events correctly [4]. Wavelet analysis has been proven as efficient signal processing techniques for a variety of signal processing problems [5]. It can be said that the benefits of using [6, 7, 8, 9] which are the new transforms are local; i.e. the event is connected to the time when it occurs. In studies wavelets used for speech/speaker recognition, it has been found that the original feature space can be augmented by the wavelet coefficients and will yield a smaller set of more robust features in the final classifier [7, 8, 9]. Artificial neural network is named from the network of nerve cells in the human brain [9]. ANNs have been investigated for many years in the hope of achieving human-like performance in automatic speech recognition [10]. These architectures are composed of many non-linear computational elements operating parallel in patterns similar to the biological neural networks [11]. Artificial neural networks have been used extensively in speech recognition during the past two decades. The most important advantages of ANNs for solving speech recognition problems are their error tolerance and non-linear property [12].

In this study, an expert speech recognition system was introduced. A combination of DWT and ANN was developed to efficiently extract the features from real speech signals and then recognize them. It will aid to increase the percentage of the correct speech recognition and enable further research of speech/voice recognition to be developed.

The organization of the paper is as follows. Section 2 reviews the wavelet transform. Section 3 demonstrates the neural network classifier. The proposed algorithm is presented in section 4. Section 5 describes the design and implementation of the system. Finally section 6 presents the conclusion.

2. Wavelet Transform

The wavelet transform was borne out of a need for further developments from Fourier transforms. Wavelet analysis represents a signal as a weighted sum of shifted and scaled versions of a characteristic wave-like function. Moreover, wavelets are often irregular and asymmetric and enable better representation of signals composed of fast changes [13]. A wavelet transform involves convolving the signal against particular instances of the wavelet at various time scales and positions. Since we can model changes in frequency by changing the time scale and model time changes by shifting the position of the wavelet, we can model both frequency and location of frequency. The wavelet transform becomes an emerging signal processing technique and it is used to decompose and reconstruct non-stationary signals efficiently. The wavelet transform can be used to represent speech signals by using the translated and scaled mother wavelets, which are capable to provide multi-resolution of the speech signal [14]. Wavelet transform is capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the speech signal. The wavelet analysis procedure is to adopt a wavelet prototype function, called an

analyzing wavelet or mother wavelet. Temporal analysis is performed with a contracted, high-frequency version of the prototype wavelet, while frequency analysis is performed with a dilated, low-frequency version of the same wavelet [15]. DWT is the most promising mathematical transformation which provides both the time and frequency information of the input signals. Wavelet transform is a technique to transform an array of N numbers from their actual numerical values to an array of N wavelet coefficients. DWT is any wavelet transform for which the wavelets are discretely sampled. It captures both frequency and location information. The digital speech signal X[n] is filtered by high pass filter H1(z) and low pass filter H0(z). The filtered results are down sampled by 2. Since most of the speech energy concentrates on the low frequency band, the low pass filtered signals need to be split again into sub bands by applying the H1 (z) and H0 (z) filters as in Figure 1. This procedure repeats until the desired decomposition level is reached. At high frequencies, the DWT provides good time resolution and poor frequency resolution. At low frequencies, DWT gives good frequency resolution and poor time resolution and vice versa.

Daubechies wavelets are compact orthogonal filter banks which satisfy the perfect reconstruction condition. In addition, Daubechies wavelets have maximum number of vanishing moments for a given order so that they can be used to provide the good approximation of the original signal. The Daubechies 4-tap (db4) filter bank was chosen for this design work.

The DWT is defined by the following equation:

$$W(j,k) = \sum_j \sum_k x(k) 2^{-j\backslash 2} \Psi(2^{-j} n-k) \qquad (1)$$

Where (Ψt) is a time function with finite energy and fast decay called the mother wavelet, j and k parameters refer to wavelet scale and translation factors. In wavelet analysis, we often speak of approximations and details. The approximations are the high- scale, low-frequency components of the signal (A). The details are the low-scale, high frequency components (D) [16].

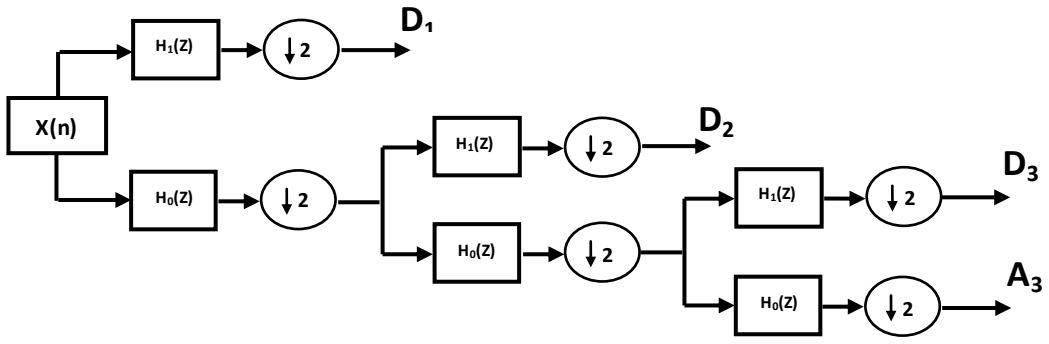

Figure 1: Discrete wavelet transform of a three stages analysis tree

3. Neural Network

Artificial Neural Network (ANN) is non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs. The ANN may be regarded as a massive parallel distributed processor that has a natural propensity for storing experimental knowledge and making it available for use [12]. The Multi Layer Perception (MLP) is a feed-forward network consisting of units arranged in layers with only forward connections to units in subsequent layers. The connections have weights associated with them. Each signal traveling along a link is multiplied by its weight. The input layer, being the first layer, has input units that distribute the inputs to units in subsequent layers. In the following (hidden) layer, each unit sums its inputs and adds a threshold to it and nonlinearly transforms the sum (called the net function) to produce the unit output (called the activation). The output layer units often have linear activations, so that output activations equal net function values [17]. MLP Neural Network is a good tool for classification purposes .It can approximate almost any regularity between its input and output [12]. The performance is measured by mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{1}^{n}(y_t - y_a)^2 \qquad (2)$$

Where y_t is the target value, y_a is the actual output, and n is the number of training data.

4. Proposed System

A complete speech recognition system based on DWT and ANN was developed in this paper to achieve the goal of the research (increasing the accuracy of recognition). Figure 2 depicts the speech recognition system developed in this study. It consists of three stages: (a) data acquisition and preprocessing and (b) features extraction and (c) classification.

In these studies, the developed system was successfully trained and tested in MATLAB version7.10 using a combination of the Signal Processing Toolbox, Wavelet Toolbox, and Neural Network Toolbox for MATLAB.

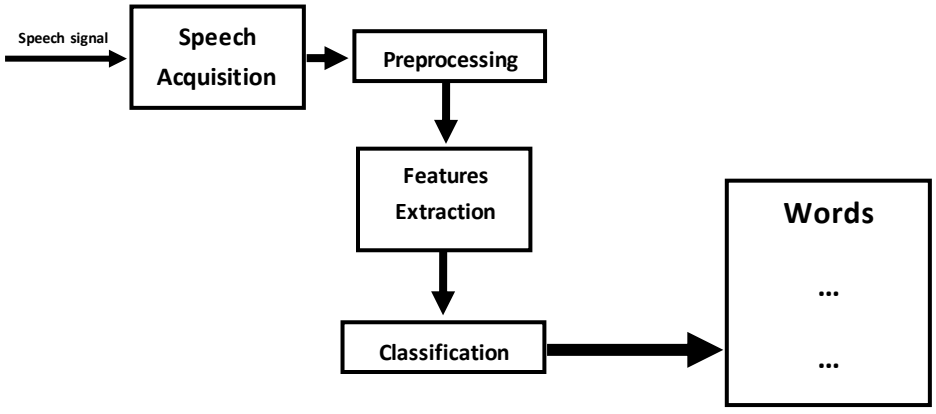

Figure 2: The structure of a speech recognition system

4.1. Data Acquisition and Preprocessing

Numbers from zero to five and the other nine words was uttered in English by 4 individuals, including 2 males and 2 females were transmitted to the computer by using a microphone and an audio card which had maximum 44 kHz sampling frequency and were recorded in a normal office environment by cool edit program version 2. Each individual uttered each word 40 times.

Recording parameters were chosen as:

Sampling rate was 16 kHz.
Bits per sample (bit rate) were 16 bits.
Number of channels was one channel (mono).
Audio format was wave.

The objective in the preprocessing stage was to modify the speech signal, so that it would be more suitable for the feature extraction analysis. A manual endpoint detection method was used to separate the word speech from the silent portions of the signal. The preprocessing stage includes removing the dc value of each signal to avoid dc offset problems and applying normalization on them to make the signals comparable as in Figure 3. The signals were normalized by using the formula

$$x_{ni} = \frac{x_i - \bar{x}}{\sigma} \qquad (3)$$

Where x_i is the *ith* element of the signal x, $\bar{x}$ and σ are the mean and the standard deviation of the vector x, respectively, x_{ni} is the *ith* element of the signal after normalization. (Lou & Loparo, 2004)

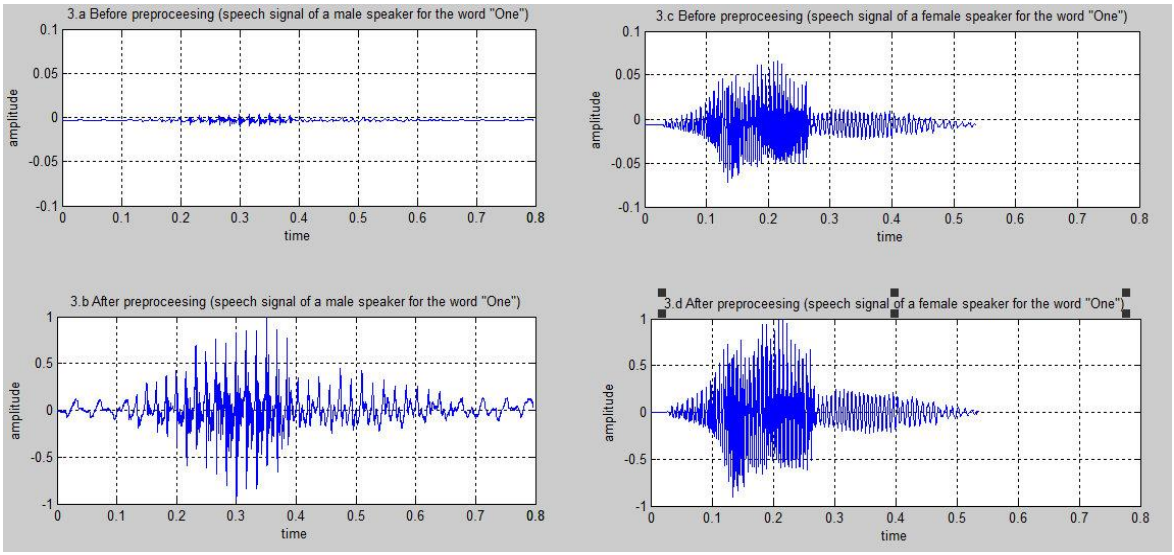

Figure 3: preprocessing stage , 3.a before preprocessing (speech signal of a male for the word "One"), 3.b after preprocessing (speech signal of a male for the word "One"), 3.c before preprocessing (speech signal of a female for the word "One"), and 3.d after preprocessing (speech signal of a female for the word "One")

4.2 Feature Extraction

In order to get an expert system based on speech recognition, features extracted from the speech signals must be chosen well since the best classifier will perform poorly if the features are not chosen correctly. Consequently Discrete Wavelet Transformation (DWT) was performed on the samples in the database. Daubechies wavelet of order 4 (db4) at level 10 was found to produce the best feature representation as in Figure 4.

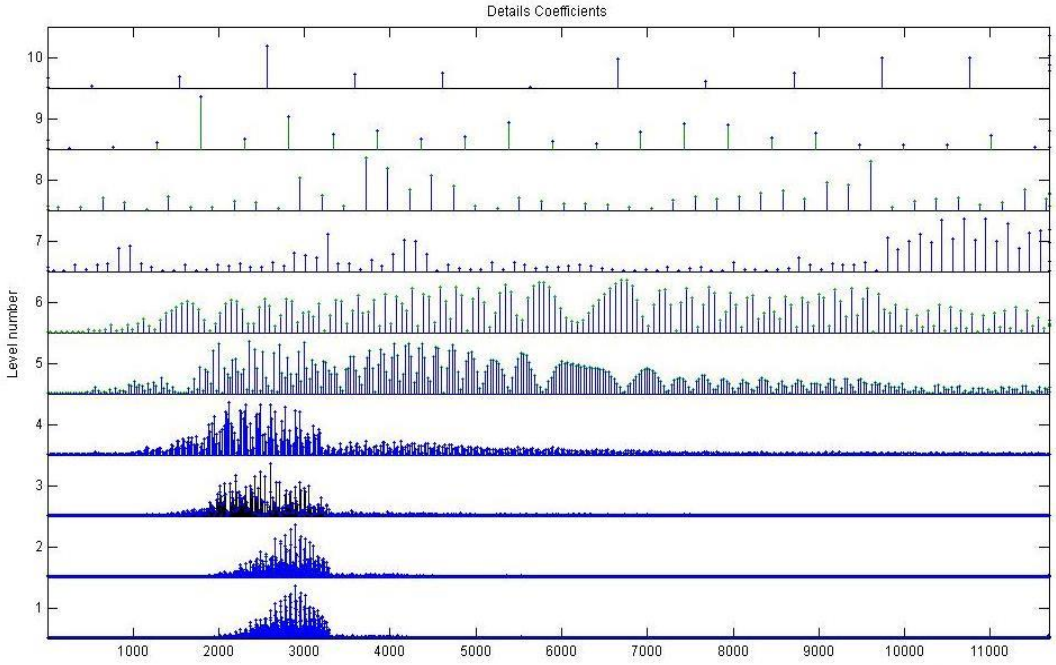

Figure 4: DWT coefficients.

The 10-level DWT resulted in 11 logarithmically spaced frequency bands. The decomposition generates a set of vectors which contain signal information at different frequency bands. After the ten-level wavelet transform, the wavelet norm, energy, maximum, and minimum for each subband were computed in order to extract the feature vector of size 44 elements per utterance. The feature extraction steps are illustrated in Figure 5.

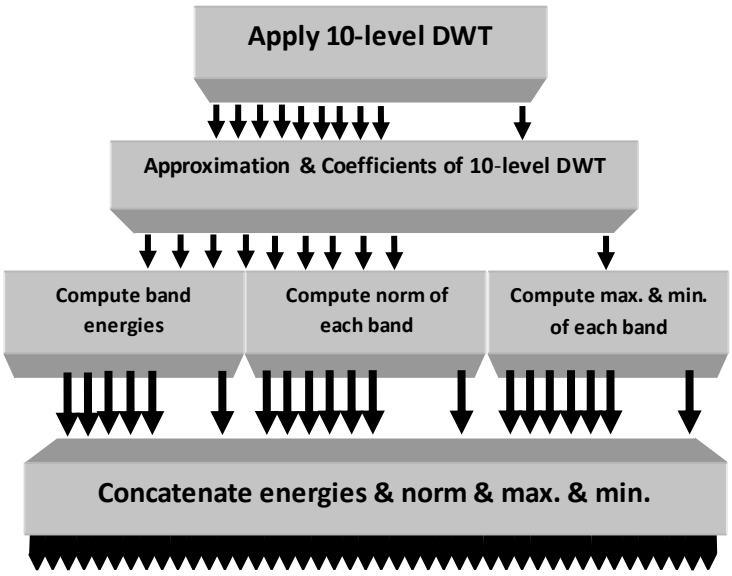

Figure 5: DWT-based feature extraction

4.3 Classification

A database of 2400 utterances was created from the English language. 60% of these utterances were used for training, 20% were used for validation and 20% were used for testing. We used the FFBPNN for training and testing of the neural network. The training parameters used in this research are illustrated in Table 1. The architecture of the network is 5-layer architecture, which is a 44-node input layer, hidden layer with 19 nodes, hidden layer with 17 nodes, and hidden layer with 15 nodes followed by the 15-node output layer.

These were selected for the best performance after several experiments. A feature matrix of size 44 × 2400 which was collected for all the words were applied to the input of the neural network as in Figure 6.

Each layer of the network (5 layers) had a weight coming from the previous layer. The first layer weights came from the inputs. The last layer, which is the network output, was designed as a 15 binary digits for each feature vector.

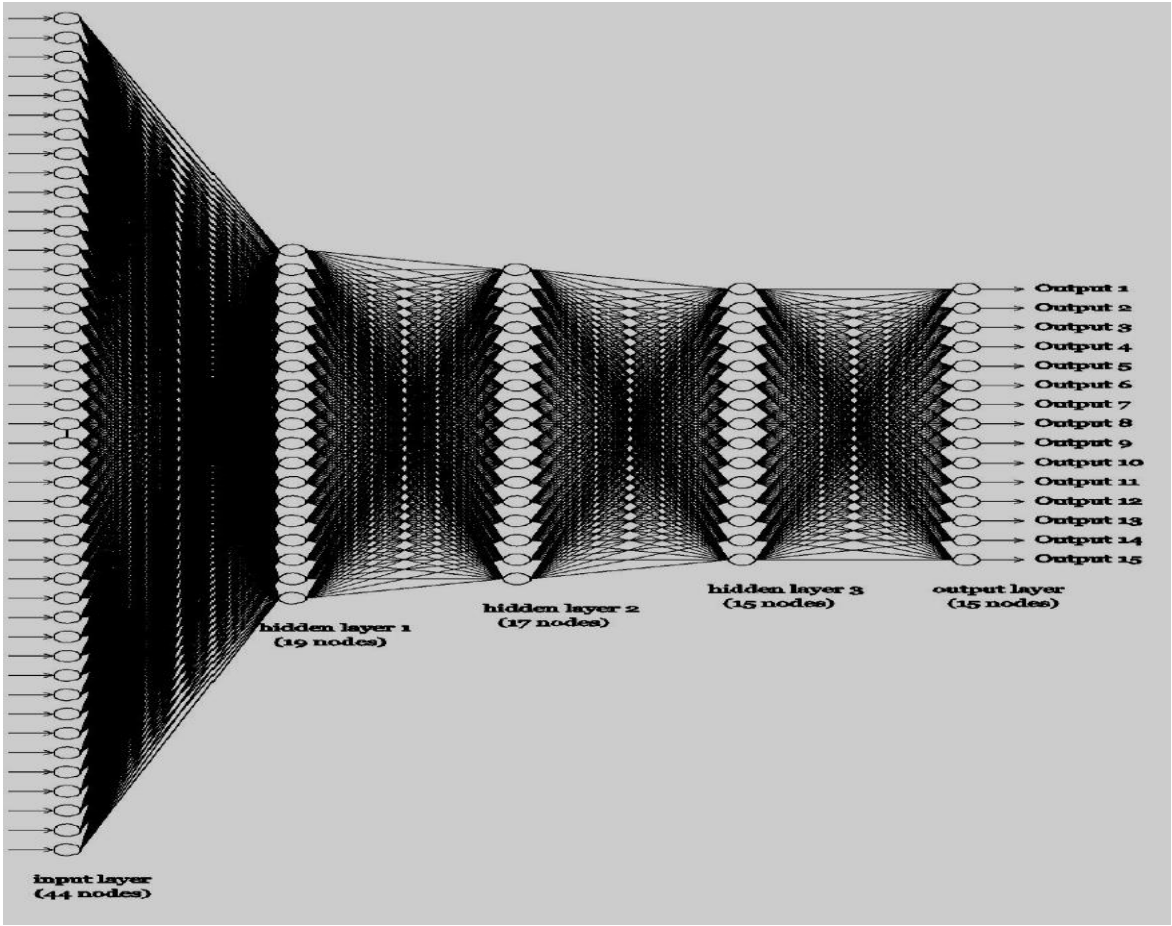

Figure 6: The feed forward backpropagation neural network

Table 1 Parameters used for the network

Architecture	
Network type	feed-forward backpropagation
No. of layers	five layers: input, three hidden and output Input:44 Hidden:19, 17, 15 Output:15
Activation function	sigmoid
Training algorithm	Levenberg-Marquardt backpropagation
performance function(mse)	1.0000e-005
No. of epochs	1000

5. Experiment and Result

The experiment was performed using a data base of 2400 English utterances for 15 words. Total of 4 individual speakers (2 males and 2 females) have spoken these 15 words. Each speaker speaks each word 40 times. Speech signals of a female and a male speaker for help word were shown in Figure 7 and 8, respectively. When the testing of the classifier was performed, an overall recognition accuracy of 98.9 % was achieved by means of FFBPNN. It indicated the effectiveness and the reliability of the proposed approach for extracting features from speech signals. Figure 9 shows a snap-shot for the GUI of the trained neural network. Testing results are tabulated in Table 2.

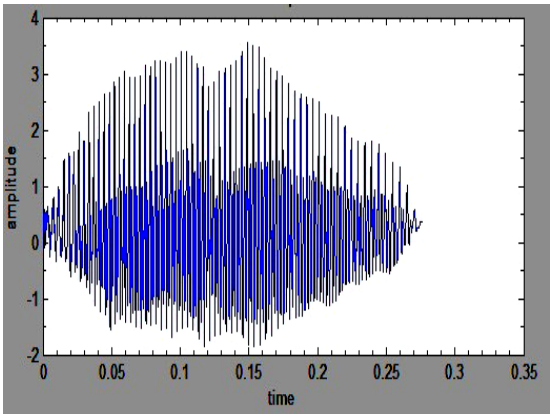

Figure 7: **Speech signal of a female speaker for help word**

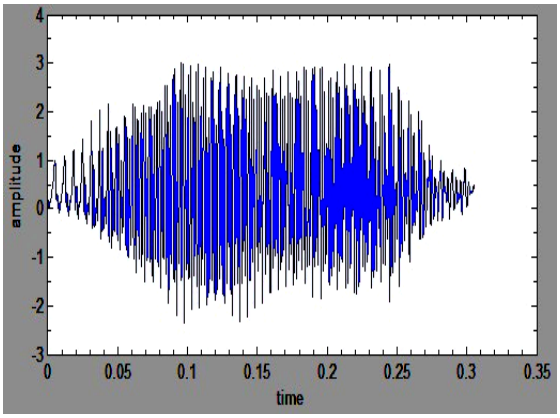

Figure 8: **Speech signal of a male speaker for help word**

A Receiver Operating Characteristic (ROC) curve is also added in Figure 10 to indicate the performance of the recognition accuracies. The ROC curve is a plot of the true positive rate versus the false positive rate. A plot of the training errors, validation errors, and test errors appears, as shown in Figure 11. The best validation performance occurred at iteration 18.

Table 2: FFBPNN recognition rate results

English Word	Total Number of Samples	Correct Classification	Incorrect Classification	The Average Recognition
Zero	160	156	4	97.5 %
One	160	154	6	96.3 %
Two	160	160	0	100 %
Three	160	154	6	96.3 %
Four	160	157	3	98.1 %
Five	160	158	2	98.8 %
Off	160	157	3	98.1 %
On	160	160	0	100 %
Play	160	158	2	98.8 %
Please	160	160	0	100 %
Sorry	160	160	0	100 %
Stop	160	160	0	100 %
Thanks	160	160	0	100 %
Ready	160	160	0	100 %
Help	160	160	0	100 %
Total	2400	2374	26	**98.9 %**

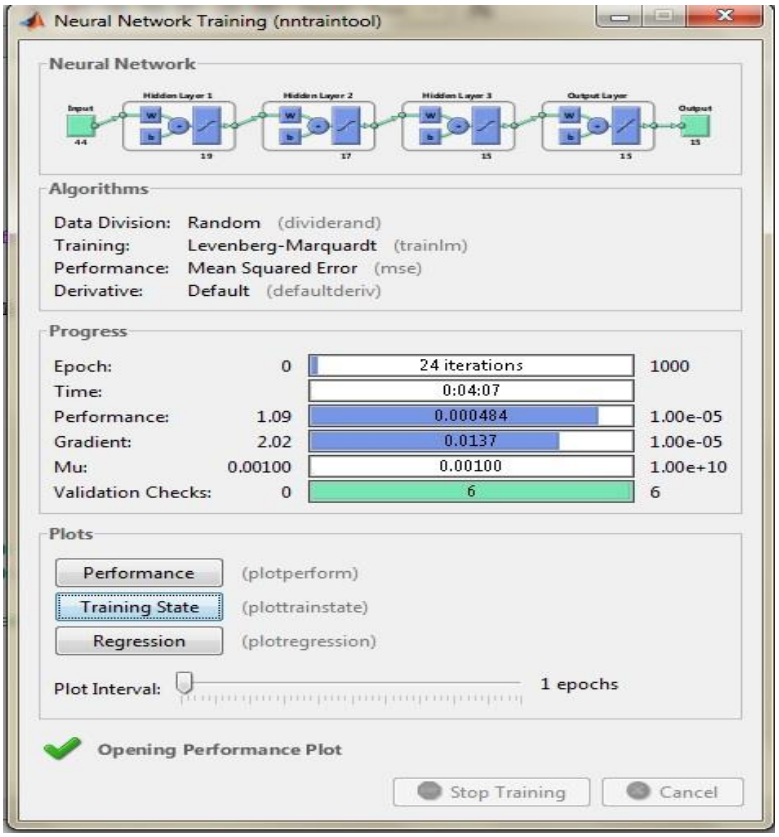

Figure 9. The GUI of the neural network training

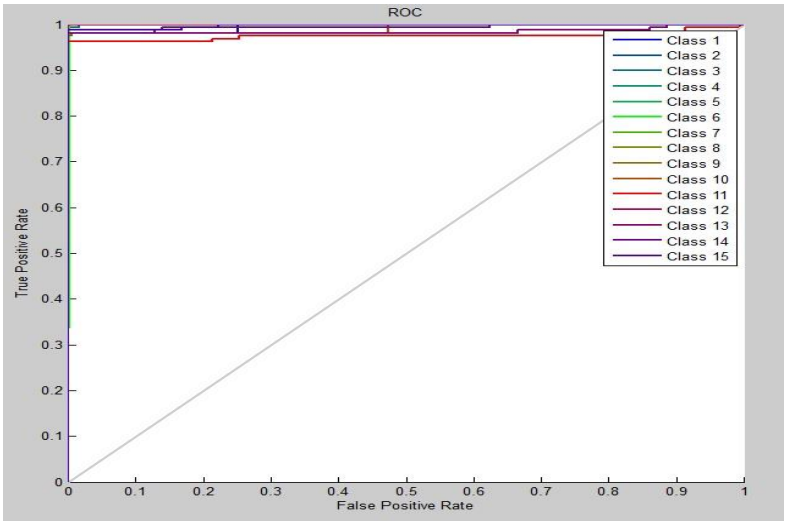

Figure 10. Operating Characteristic Curve for the proposed system

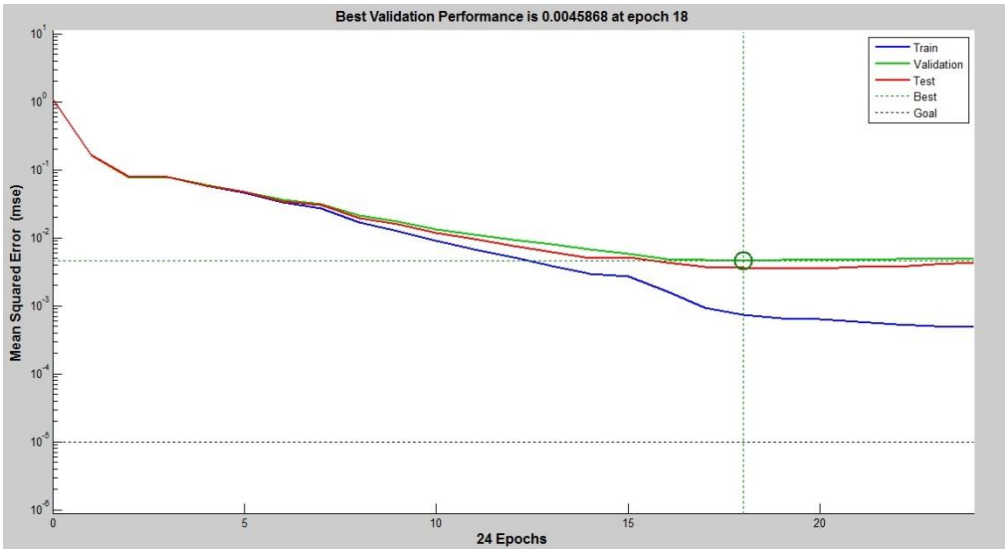

Figure 11. A plot of the training errors, validation errors, and test errors appears

6. Conclusion

In this study, an expert speech recognition system for isolated words based on a developed model of Discrete Wavelet Transform and Artificial Neural Network techniques was proposed. According to the experimental results, the proposed method can make an effectual analysis. The average identification rate of the system was 98.9 %. The stated results show that the proposed method can make an accurate and robust classifier.

Reference

[1] Ayadi, M.M.H.E., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition(2011) 572-587

[2] Patel, I., Dr. Srinivas Rao, Y., Speech Recognition using HMM with MFCC- An Analysis using Frequency Specral Decomposion Technique, Signal & Image Processing : An International Journal(SIPIJ) , 1(2) (December 2010).

[3] Trivedi, N., Dr. Kumar, V., Singh, S., Ahuja, S., and Chadha, R., Speech Recognition by Wavelet Analysis, International Journal of Computer Applications 15(8) (February 2011) 27–32.

[4] Avci, E., and Akpolat, Z.H., Speech recognition using a wavelet packet adaptive network based fuzzy inference system, SinceDirect, vol.31, no. 3, 2006, pp 495- 503.

[5] Siafarikas, M., Ganchev, T. & Fakotakis, Wavelet packets based speaker verification. In Proceedings of the ISCA speaker and language recognition workshop – Odyssey'2004, Toledo, Spain, May 31–June 3, (2004) 257–264.

[6] Saito, N. "Local feature extraction and its application using a library of bases." Phd thesis, Yale University (1994).

[7] Buckheit, J. B. and Donoho, D. L., WaveLab and Reproducible Research, Dept. of Statistics, Stanford University, Tech. Rep. 474 (1995).

[8] Wesfred, E., Wickerhauser, V., Adapted local trigonometric transforms and speech processing. IEEE trans. on Signal Proc. 41 N.12 (1993) 3596-3600.

[9] Visser, E., Otsuka, M. & Lee, A spatio-temporal speech enhancement scheme for robust speech recognition in noisy nvironments, Speech Communication. 41 (2003) 393–407.

[10] Alotaibi, Y.A., Investigation of spoken Arabic digits in speech recognition setting, Informatics and Computer Sciences 173 (1–3) (2005) 105–139.

[11] Lampinen, J., Oja, E., Fast self-organization by the probing algorithm, In Proceedings of the International Joint Conference on Neural Networks (IJCNN), volume II (1989) 503-507, Piscataway, NJ. IEEE Service Center.

[12] Haykin, S. Neural Networks: A comprehensive Foundation, Prentice Hall, 1999.

[13] Canal, M.R., "Comparison of Wavelet and Short Time Fourier Transform Methods in the Analysis of EMG Signals," Journal of Medical Systems, (2008)1- 4.

[14] Pang, J., Chauhan, S., FPGA Design of Speech Compression by Using Discrete Wavelet Transform, WCECS 2008, Francisco, USA, 22 - 24 October 2008, pp. 151 – 156.

[15] An Introduction to Wavelets, The original version of this work appears in IEEE Computational Science and Engineering, Summer 1995, vol. 2, num. 2, published by the IEEE Computer Society, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720, USA,

[16] Kadambe, S., Srinivasan, P., Application of Adaptive Wavelets for Speech, Optical Engineering 33(7) (July 1994) 2204-2211.

[17] Vimal Krishnan, V.R., Babu Anto, P., Feature Parameter Extraction from Wavelet Subband Analysis for the Recognition of Isolated Malayalam Spoken Words, (IJCNS) International Journal of Computer and Network Security, 1(1) (October 2009).