Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 32 (2007) 485–498
Expert Systems with Applications
www.elsevier.com/locate/eswa

A new optimum feature extraction and classification method for speaker recognition: GWPNN

Engin Avci *

Firat University, Technical Education Faculty, Department of Electronic and Computer Science, 23119 Elazig, Turkey

Abstract

Speech and speaker recognition is an important topic to be performed by a computer system. In this paper, an expert speaker recognition system based on optimum wavelet packet entropy is proposed for speaker recognition by using real speech/voice signal. This study contains both the combination of the new feature extraction and classification approach by using optimum wavelet packet entropy parameter values. These optimum wavelet packet entropy values are obtained from measured real English language speech/voice signal waveforms using speech experimental set. A genetic-wavelet packet-neural network (GWPNN) model is developed in this study. GWPNN includes three layers which are genetic algorithm, wavelet packet and multi-layer perception. The genetic algorithm layer of GWPNN is used for selecting the feature extraction method and obtaining the optimum wavelet entropy parameter values. In this study, one of the four different feature extraction methods is selected by using genetic algorithm. Alternative feature extraction methods are wavelet packet decomposition, wavelet packet decomposition – short-time Fourier transform, wavelet packet decomposition – Born–Jordan time–frequency representation, wavelet packet decomposition – Choi–Williams time–frequency representation. The wavelet packet layer is used for optimum feature extraction in the time–frequency domain and is composed of wavelet packet decomposition and wavelet packet entropies. The multi-layer perceptron of GWPNN, which is a feed-forward neural network, is used for evaluating the fitness function of the genetic algorithm and for classification speakers. The performance of the developed system has been evaluated by using noisy English speech/voice signals. The test results showed that this system was effective in detecting real speech signals. The correct classification rate was about 85% for speaker classification.

Keywords: English speech signal; Adaptive feature extraction; Wavelet packet decomposition; Entropy; Genetic algorithm; Wavelet packet-neural networks; Expert system

1. Introduction

Many speech processing tasks, like speech/voice and speaker recognition, reached satisfactory performance levels in speech recognition applications (Coifman & Wickerhauser, 1992; Siafarikas, Ganchev, & Fakotakis, 2004). Though a diversity of commercial yields were launched in recent years, many problems remain as an open research region, and perfect solutions are not found out, yet. For instance, such a problem, which is providing an adequate parameterization of the speech/voice signal for needs of speaker recognition.

Speech/speaker recognition systems commonly carry out some kind of classification/recognition based on speech features which are usually obtained via Fourier transforms (FTs), short-time Fourier transforms (STFTs) or linear predictive coding techniques. 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 analyse localized events correctly. Moreover, the LPC method accepts a particular linear (all-pole) model of speech/voice production which severely speaking is not the case. Other methods based on Cohens

* Tel.: +90 424 2370000x4257; fax: +90 424 2367064.
E-mail address: enginavci@firat.edu.tr

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.12.004

general class of time–frequency distributions such as the Born–Jordan, Cone–Kernel and Choi–Williams methods have also found use in speech/voice recognition applications but have the drawback of introducing unwanted cross-terms into the representation.

Over the recent years, wavelet and wavelet packet analysis in speech/voice and speaker recognition has been proven an effective signal processing technique for a variety of problems. In feature extraction stage of speech/voice and speaker recognition approach designed for the purpose of speech recognition, wavelets have been used twofold.

Automatic speech recognition (ASR) is generally supported by commercial corporations (Kadambe & Srinivasan, 1994). Opinions of the researchers, who study in speech/speaker recognition area, as summarized in Kadambe and Srinivasan (1994) and Mallat (1989) appear to ignore the benefits that can be gained by proper transformations of the input signal. The main task in automatic speaker recognition (ASR) is to separate various speaker classes (Evangelista, 1994; Kadambe & Boudreaux-Bartels, 1992; Kadambe & Srinivasan, 1994; Mallat, 1989). In the literature, some researchers have explored the use of wavelets to provide a richer feature space (Evangelista, 1993, 1994; Maes, 1994; Saito, 1994; Szu, Telfer, & Kadambe, 1992). Nevertheless, there is little evidence of widespread use of this technique (Kadambe & Srinivasan, 1994). In Saito (1994), it is claimed that pre-processing the data allows easier subsequent feature extraction and increased resolution. The signal was transformed from a time domain to a frequency domain using the Fourier transform by engineers (Buckheit & Donoho, 1995). Despite this being useful for some applications, this transform was not excess useful for automatic speaker recognition using real speech signals (Coifman & Wickerhauser, 1992). Because the Fourier transform tells us that a feature occurs somewhere in the signal, but does not specify where it occurs. Wavelets bring a new tool to the speech signal classification. It can be said that the benefits of using wavelets (Visser, Otsuka, & Lee, 2003; Wesfried & Wickerhauser, 1993) which are the new transforms are local; i.e., the event is connected to the time when it occurs. In studies in which wavelets are 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 (Coifman & Wickerhauser, 1992; Visser et al., 2003; Wesfried & Wickerhauser, 1993).

Artificial neural network is named from the network of nerve cells in the human brain (Visser et al., 2003). ANNs have been investigated for many years in the hope of achieving human-like performance in automatic speech/speaker recognition (Alotaibi, 2005). These architectures are composed of many non-linear computational elements operating parallel in patterns similar to the biological neural networks (Lippmann, 1989). Artificial neural networks have been used extensively in speech recognition during the past two decades. The most important advantages of ANNs for solving speech/speaker recognition problems are their error tolerance and non-linear property (Haykin, 1999).

In several studies, wavelet neural network (WNN) and wavelet packet-neural network (WPNN) were used for speech/voice and speaker recognition (Ha, Tran, & Dissanayake, 2005; Siafarikas et al., 2004; Wesfried & Wickerhauser, 1993). But in none of these studies, adaptive entropy method was used for reducing the dimension of the input feature vector.

In this paper a novel method, which is an expert system for speaker recognition, is introduced. It will aid in the automatic speaker recognition and enable further research of speaker recognition to be developed. It was constituted by a combination of genetic algorithm, wavelet packet transform and neural network to efficiently extract the features from pre-processed real English speech signals for the purpose of automatic speaker recognition among variety speakers. An algorithm called the expert system is developed, which processes the pattern recognition approximation.

In this study, an experiment set is used for obtaining the real speech/voice signal data sets. The speech experiment set is for educational purpose. English speech word signals are transmitted to computer media by using an audio card which has 44 kHz sampling frequencies.

The paper is organized as follows. In Section 2, is reviewed some basic properties of wavelet packet decomposition, 3-D time–frequency representations (short-time Fourier transform, Born–Jordan TFR, Choi–Williams TFR), genetic algorithm and wavelet packet-neural networks. An expert system is described in Section 3. This new method enables a large reduction of the speech signal data while retaining problem specific information, which facilitates an efficient speaker recognition process. The effectiveness of the proposed method for classification of English speech signals in automatic speaker recognition area is demonstrated in Section 4. Finally, Section 5 presents the conclusions.

2. Preliminaries

2.1. Wavelet packet transform

Wavelet packet transforms generalize the filter bank tree that relates wavelets. They conjugate mirror filters. In the decomposition of a speech signal by using the wavelet packet transform, only the lower frequency band is decomposed, giving a right recursive binary tree structure, where its right lobe represents the lower frequency band. Its left lobe represents the higher frequency band. In the corresponding decomposition by using wavelet packet transform, the lower, as well as the higher frequency bands are decomposed giving a balanced binary tree structure (Siafarikas et al., 2004). Such a tree is shown in Fig. 1.

Wavelet decomposition uses the fact that it is possible to resolve high frequency components within a small time window, while only low frequency components need large

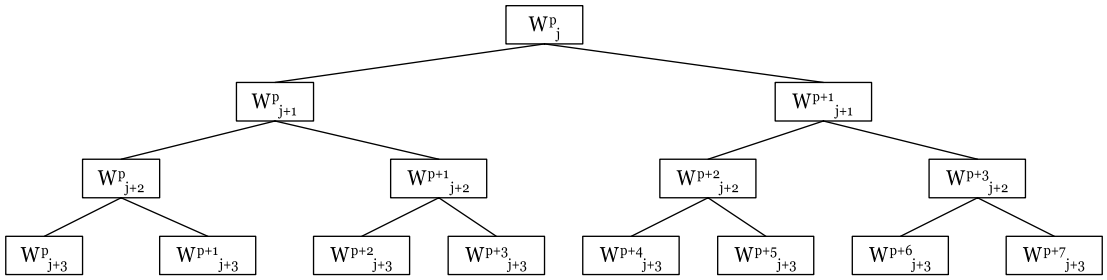


Fig. 1. Tree of wavelet packet transform (p: number of nodes at the same wavelet packet decomposition level, j: number of wavelet packet decomposition level).

time windows. This is because a low frequency component completes a cycle in a large time interval whereas a high frequency component completes a cycle in a much shorter interval. Therefore, slow varying components can only be identified over long-time intervals but fast varying components can be identified over short-time intervals. Wavelet decomposition can be regarded as a continuous time wavelet decomposition sampled at different frequencies at every level or scale. The wavelet decomposition function at level m and time location t_m can be expressed as

$$d_m(t_m) = x(t)\psi_m\left(\frac{t - t_m}{2^m}\right). \tag{1}$$

Wavelet packet analysis is an extension of the discrete wavelet transform (DWT) (Burrus, Gopinath, & Guo, 1998) and it turns out that the DWT is only one of the much possible decomposition that could be performed on the signal. The advantage of wavelet packet analysis is that it is possible to combine the different levels of decomposition in order to achieve the optimum time–frequency representation of the original (Mallat & Zhong, 1992).

2.2. Time–frequency representations and entropy types used in this study

In this paper, short-time Fourier transform (STFT), Born–Jordan TFR, Choi–Williams (CW) TFR (Boashash, 1992) were used for adaptive feature extraction. It can be looked at reference (Boashash, 1992) for more knowledge about these time–frequency representations.

Particularly, all TFR can be obtained from

$$C(t,w) = \frac{1}{4\pi^2}\int\int\int s\left(u - \frac{\tau}{2}\right)s^* \times \left(u + \frac{\tau}{2}\right)\phi(\theta,\tau)\mathrm{e}^{-j\theta t - j\theta\tau + j\theta u}\,\mathrm{d}u\,\mathrm{d}\tau\,\mathrm{d}\theta. \tag{2}$$

Here, θ is frequency variation, τ is time variation, and $\phi(\theta,\tau)$ is a two-dimensional function called the kernel, and $s(u - \tau/2)$ and $s^*(u + \tau/2)$ represent the real and complex conjugates of the signal, respectively. There are multiple frequencies in a given signal. The time–frequency spectrum will also show artificial frequencies in addition to the true ones present in the given signal. To reduce the problem of artificial frequencies in the case of a multi-component signal, many different kernels have been designed.

There are three main reasons as to why the kernel idea is particularly useful to study time–frequency distributions. First it is easy to generate a kernel function. The second reason is that distributions with certain properties can be extracted by constraining the kernel (Boashash, 1992). The third reason is that when a new distribution is considered its properties can readily be searched by examining its kernel (Boashash, 1992).

Entropy-based criteria describe information-related properties for an accurate representation of a given signal. Entropy is a common concept in many fields, mainly in signal processing (Boashash, 1992). A method for measuring the entropy appears as an ideal tool for quantifying the ordering of non-stationary signals. An ordered activity (i.e. a sinusoidal signal) is manifested as a narrow peak in the frequency domain, thus having low entropy. On the other hand, random activity has a wide band response in the frequency domain, reflected in a high entropy value (Quiroga, 1998). The types of entropy computing are threshold, norm, log energy, and sure which will be mentioned under Section 3.2 (Coifman & Wickerhauser, 1992; MATLAB 5.3 version Wavelet Toolbox).

2.3. Using of the genetic algorithms

An evolutionary process is used for solving a problem by genetic algorithms. The genetic algorithm begins with a set of solutions which are represented by individuals. These sets of solution are called as population. Solutions from one population are taken and used to form a new population. This iterative process is maintained during the new population and will be better than the old one. Solutions are then selected to form new solutions according to their fitness values. If fitness value of an individual is better than another individual's, this individual is more lucky than the other to be reproduced at the next population. This iterative process is repeated until some conditions (for example number of populations or improvement of the best solution) are satisfied.[1]

[1] http://cs.felk.cvut.cz/~xobitko/ga/, 2005.

Firstly, there is generated random population of n individuals which provide suitable solutions for the problem. Secondly, is evaluated the fitness $f(x)$ of each individual x in the population.[1] Thirdly, there is created a new population by repeating the following steps until the new population is complete. Fourthly, two parent individuals from a population are selected according to their fitness. The better fitness is interpreted as the bigger chance to be selected for the next population. Fifthly, the parents are crossed over with a crossover probability to form new individuals. If crossover is not performed, the individual is the exact copy of parents. Sixthly, new individuals are mutated with a mutation probability at each locus which is the position in an individual. Seventhly, new individuals are placed in the new population. Eighthly, the new generated population is used for a further run of the algorithm. Ninthly, the genetic algorithm is stopped, if the end condition is satisfied, and the best solution is returned in the current population. Tenthly, it is moved to the step second.

2.4. Using of wavelet packet-neural networks in speech/voice recognition area

An artificial neural network (ANN) is a mathematical model consisting of a number of highly interconnected processing elements organized into layers, the geometry and functionality of which have been likened to that of the human brain. The ANN may be regarded as processing learning capabilities. It has natural propensity for storing experimental knowledge.

Neural networks are systems that are constructed to make use of some organizational principles resembling those of the human brain (Haykin, 1994). They represent the promising new generation of information processing systems. Neural networks are good at tasks such as pattern matching and classification, function approximation, optimization and data clustering, while traditional computers, because of their architecture, are inefficient at these tasks, especially pattern-matching tasks (Bishop, 1996). The wavelet packet-neural networks try to combine aspects of the wavelet packet transformation for purpose of feature extraction and selection with the characteristic decision capabilities of neural network approaches (Zhang, Walter, Miao, & Lee, 1995). The wavelet packet-neural network (WPNN) is constructed based on the wavelet packet transform theory (Wang, Teo, & Lin, 2001; Zhang & Benveniste, 1992) and is an alternative to feed-forward neural network (Thuillard, 2000). Wavelet packet decomposition (Burrus et al., 1998) is a powerful tool for non-stationary signal analysis. Let $x(t)$ be a piecewise continuous function. Wavelet packet decomposition allows one to decompose $x(t)$ using a wavelet function $\psi : R^n \to R$. Based on the wavelet packet decomposition, wavelet network structure is defined by

$$y(x) = \sum_{i=1}^{N} w_i \psi[D_i(x - t_i)] + b, \tag{3}$$

where D_i are dilation vectors specifying the diagonal dilation matrices D_i, t_i are translation vectors, and the additional parameter b is introduced to help deal with non-zero mean functions on finite domains. An algorithm of the back-propagation type has been derived for adjusting the parameters of the WPNN (Zhang & Benveniste, 1992).

By virtue of its parallel distribution, an ANN is generally robust for tolerant of faults and noise, able to generalize well and capable of solving non-linear problems (Coifman & Wickerhauser, 1992). Applications of ANNs in the speech/speaker recognition field include applications of the automatic speech/speaker recognition by using speech/voice signals[1] (Avci & Turkoglu, 2003; Boashash, 1992; Coifman & Wickerhauser, 1992; Jakubiak, Arabas, & Grabczak, 1997; MATLAB 5.3 version Wavelet Toolbox; Quiroga, 1998; Turkoglu, Arslan, & Ilkay, 2003); however, genetic-wavelet-neural network analysis of speech/voice signals is a relatively new approach to date.

3. Experimental applications

Automatic speaker recognition (ASR) system developed in this study is shown in Fig. 2. It consists of two parts: (a) data acquisition and pre-processing and (b) optimum feature extraction and classification using a genetic-wavelet packet-neural network (GWPNN) structure.

3.1. English language database used in this study

An English language words database was created from 35 English words. These 35 English words are given in Table 1. A total of 40 individual speakers, 20 individual males and 20 individual females, spoke these 35 English words for training and testing phases and repeated all same 35 English words three times as noisy speech signals, which have different white-noise amplitudes (signal/noise rate (SNR) = −2 dB, −3 dB, and −5 dB) for the testing phase. Thus, the total number of tokens considered for training was 1400 (40 speakers × 35 English noiseless words). For the testing mode, all the 5600 (40 speakers × 35 Turkish words × 4 (for noiseless, noisy speech signals, which have different white-noise amplitudes (signal/noise rate (SNR) = −2 dB, −3 dB, and −5 dB))) tokens were used in the recognition phase (testing mode). This situation implies that the training data set is a subset of the testing data set. Table 2 shows some of the system parameters. English speech signals of male-10, male-12, female-2, and female-8 speakers for "zero" English word are given in Figs. 3–6, respectively.

All of the English language speech signals were acquired from the experimental set whose block diagram is shown in Fig. 2.

This system was partitioned into several stages according to their functionality as shown in Fig. 2. These stages can be explained as below:

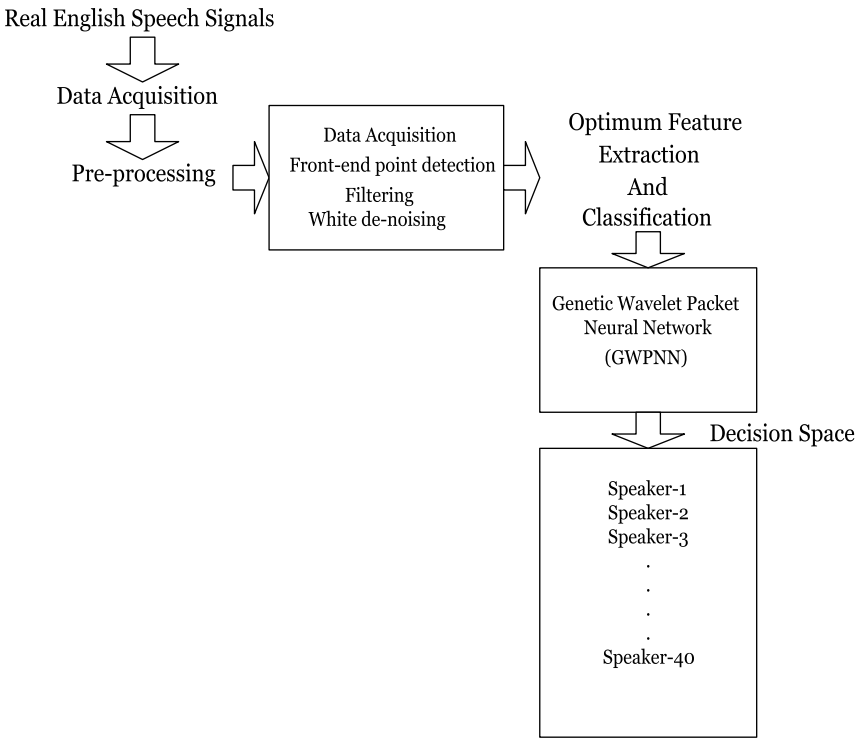


Fig. 2. The algorithm of the automatic speaker recognition (ASR) system.

3.1.1. Stage-1: data acquisition and pre-processing

Pre-processing to obtain the feature vector was performed on the digitized, which were received by the computer media by using audio card, in the following order:

(i) First is the digital signal processing front-end part, whose functions are speech/voice acquisition through a microphone, filtering, and sampling. Band-pass filter with cut-off frequencies of 100 Hz and 4.8 kHz was used to filter the speech signal before processing. The sampling rate was used as 10 kHz with 16-bit resolution for all recorded English language speech tokens. For separate speech from silent portions of the signal, a manual endpoint detection method was used. This method also detected the beginning and the end points of the spoken word (Avci, Turkoglu, & Poyraz, 2005a). A 256-point Hamming window was used to select the data points to be analyzed (Avci et al., 2005a).

(ii) White de-noising: White noise is a random signal that contains equal amounts of every possible frequency, i.e., its FFT has a flat spectrum (Avci, Turkoglu, & Poyraz, 2005b). The speech signals were filtered by removing the white noise by using wavelet. The white de-noising procedure contains three steps (Avci, Turkoglu, & Poyraz, 2005c):

- Computing the wavelet packet decomposition of the speech signal at level 7 and using the Daubechies wavelet of order 10.
- For each level from 1 to 7, soft thresholding is applied to the detail coefficients.
- Computing wavelet reconstruction based on the original approximation coefficients of level 7 and the modified detail coefficients of levels from 1 to 7.

3.1.2. Stage-2: feature extraction and classification

Fig. 7 shows the GWPNN algorithm for classification of English language speech signals waveform patterns from the speech experimental set. Feature extraction is the key for speaker recognition so that it is arguably the most important component of designing the expert system based on pattern recognition since the best classifier will perform poorly if the features are not chosen well. A feature extractor should reduce the pattern vector (i.e., the original waveform) to a lower dimension, which contains most of the useful information from the original vector.

The English language speech waveform patterns obtained from speech experimental set are rich in detail and highly non-stationary. After the data pre-processing has been realized, GWPNN algorithm is used for optimum feature extraction and classification.

3.2. Used entropy types for feature extraction and classification

The types of norm, log energy, and sure entropy are explained as below:

Table 1
English words used in this study

English words	Syllables	No. of syllables
Zero	CVCV	2
One	VCV	2
Two	CCV	1
Three	CCCVV	2
Four	CVVC	2
Five	CVCV	2
Six	CVC	1
Seven	CVCVC	2
Eight	VVCCC	2
Nine	CVCV	2
Mould	CVVCC	2
Pool	CVVC	2
Red	CVC	1
Apple	VCCCV	2
Stony	CCVCC	1
Leaf	CVVC	2
Vehicle	CVCVCCV	3
Article	VCCVCCV	3
Male	CVCV	2
Female	CVCVCV	3
Use	VCV	2
Father	CVCCVC	2
Above	VCVCV	3
Please	CCVVCV	3
Begin	CVCVC	2
Delete	CVCVCV	3
Different	CVCCVCVCC	3
Analysis	VCVCCCVC	3
Sword	CCVCC	1
Letter	CVCCVC	2
Notebook	CVCVCVVC	4
Pencil	CVCCVC	2
Inquiry	VCCVVCC	3
Dress	CCVCC	1
Cliff	CCVCC	1

Table 2
Some of the ASR system parameters

Parameter	Value
Sampling rate	10 kHz, 16 bits
Database	Isolated 35 English words
Speakers	40 (20 male + 20 female)
Magnitude of the white noise	−2 dB, −3 dB, −5 dB
Filter cut-off frequencies	100 Hz and 4.8 kHz
Window type and size	Hamming, 256

The Norm entropy $E(s)$ is

$$E(s) = \sum_{i=0} |s_i|^p \quad \text{for } (1 \leqslant p < 2), \tag{4}$$

where s is signal and s_i is the ith coefficient of the signal (Coifman & Wickerhauser, 1992; MATLAB 5.3 version Wavelet Toolbox).

The Sure entropy $E(s)$ is

$$|s_i| \leqslant \varepsilon \Rightarrow E(s) = \sum_{i=0} \min(s_i^2, \varepsilon^2). \tag{5}$$

Here, ε is a positive threshold value (Coifman & Wickerhauser, 1992; MATLAB 5.3 version Wavelet Toolbox), s is signal and s_i is the ith coefficient of the signal.

The logarithmic energy entropy $E(s)$ is

$$E(s) = \sum_i \log_2(s_i^2), \tag{6}$$

where s is signal and s_i is the ith coefficient of the signal (Coifman & Wickerhauser, 1992; MATLAB 5.3 version Wavelet Toolbox).

3.3. Using of the GWPNN for optimum feature extraction and classification

The feature extraction methods, which can be chosen by using genetic algorithm in this study, are ordered as below:

1. *Wavelet packet decomposition (WPD).* For wavelet packet decomposition of the English language speech waveforms, the decomposition structure and reconstruction tree are at level 7. Wavelet packet decomposition was applied to the English speech signal by using the Daubechies-10 wavelet packet decomposition filters. Thus, there was obtained $2^7 = 128$ wavelet packet coefficients for each of the English speech waveforms.
2. *Wavelet packet decomposition and short-time Fourier transform (WPD-STFT).* In this method, same wavelet packet decomposition process was applied to English speech signal as in feature extraction method-1. Afterwards, the STFT was applied to each of the obtained wavelet packet decomposition coefficients. The STFT is understood as the most robust of the various time–frequency representations. The STFTs of waveforms of terminal nodes were computed.
3. *Wavelet packet decomposition and Born–Jordan time–frequency representation (WPD-BJTFR).* In this method, same wavelet packet decomposition process was applied to English speech signal as in feature extraction method-1. Afterwards, the BJ TFR was applied to each of the obtained wavelet packet decomposition coefficients. The BJ TFRs of waveforms of terminal nodes were computed.
4. *Wavelet packet decomposition and Choi–Williams time–frequency representation (WPD-CWTFR).* In this method, same wavelet packet decomposition process was applied to English speech signal as in feature extraction method-1. Afterwards, the CW TFR was applied to each of the obtained wavelet packet decomposition coefficients. The CW TFRs of waveforms of terminal nodes were computed.

3.4. GWPNN algorithm

GWPNN algorithm was developed for determining the most efficient method of four different feature extraction methods which are stated in Section 3.3, optimum P

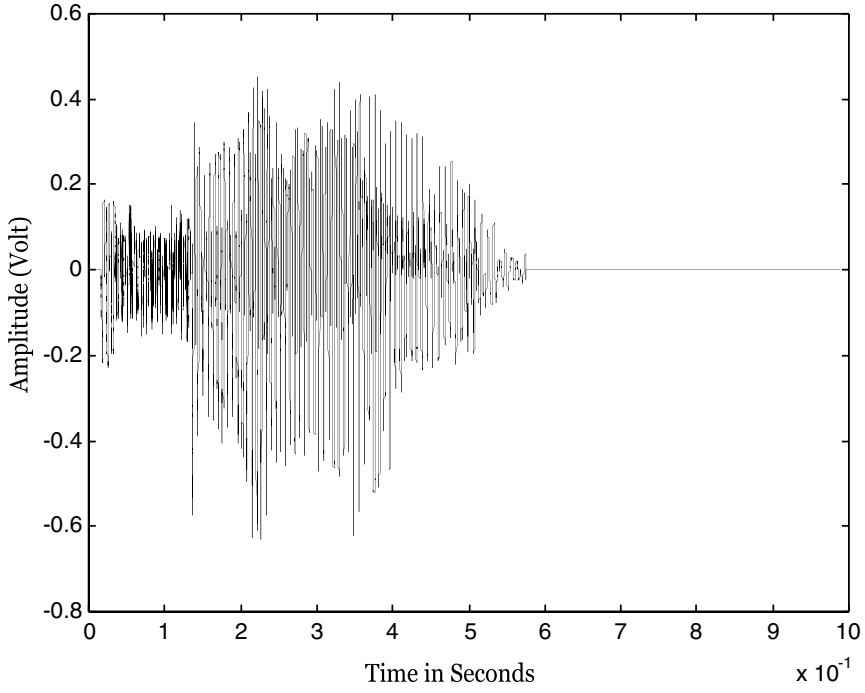


Fig. 3. English speech signal of male-10 speaker for "zero" English word.

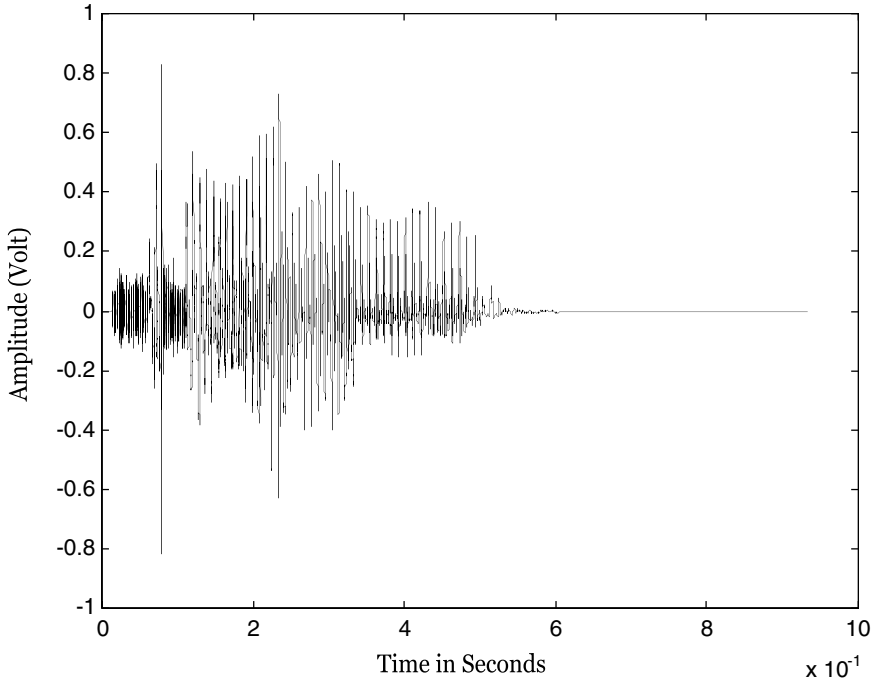


Fig. 4. English speech signal of male-12 speaker for "zero" English word.

parameter value of the norm entropy and optimum ε parameter value of the sure entropy at automatic speaker recognition. The realized operations at the GWPNN can be ordered as below:

• *Initial population:* Twenty random individuals formed from total 8 bits were chosen as initial population. First and second bits of each of the individuals represent one of the four different feature extraction methods which are mentioned above. Third, fourth, and fifth bits of each of the individuals represent P parameter value of the norm entropy which is mentioned under Section 3.2. Sixth, seventh, and eighth bits of each of the individuals

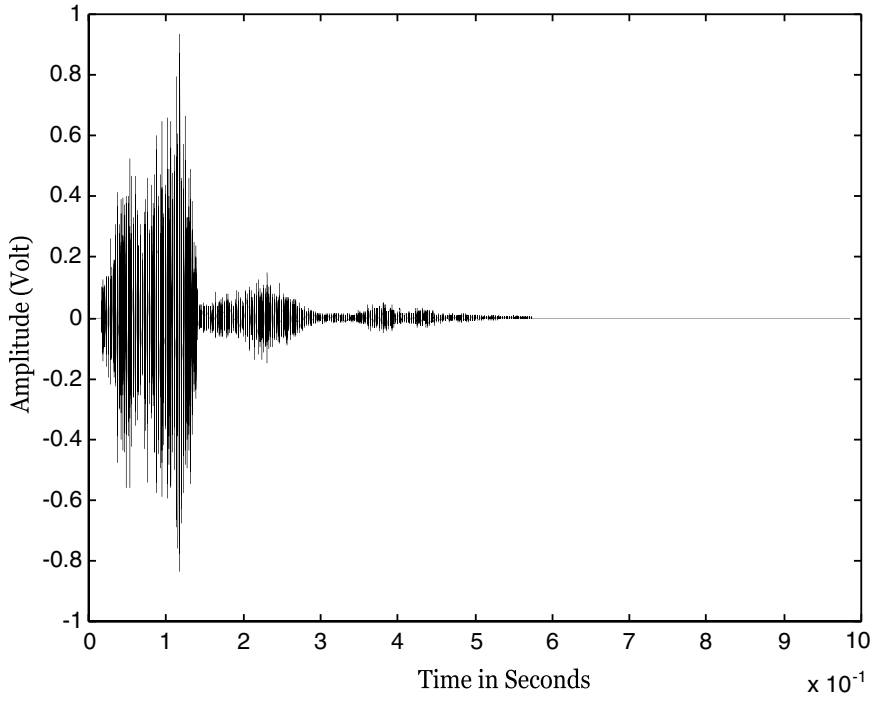


Fig. 5. English speech signal of female-2 speaker for "zero" English word.

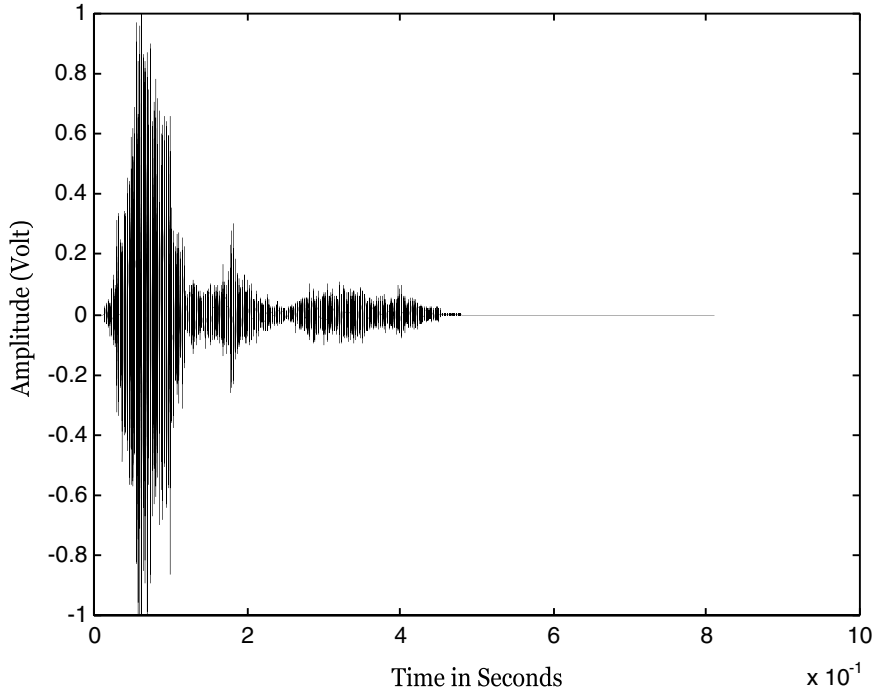


Fig. 6. English speech signal of female-8 speaker for "zero" English word.

represent ε parameter value of the sure entropy which is mentioned in Section 3.2. In norm entropy, P is the power and must be such that $1 \leqslant P < 2$. In sure entropy, ε is the threshold and must be such that $1 \leqslant \varepsilon \leqslant 8$. Where, sensitive value for P parameter is 1/7 due to it being represented to P parameter by using three bites at each of the individuals of the population. The values which P parameter can get are 1, 1.142, 1.285, 1.426, 1.568, 1.71, 1.852, and 1.994. It is represented to ε parameter by using three bites for each of the individuals

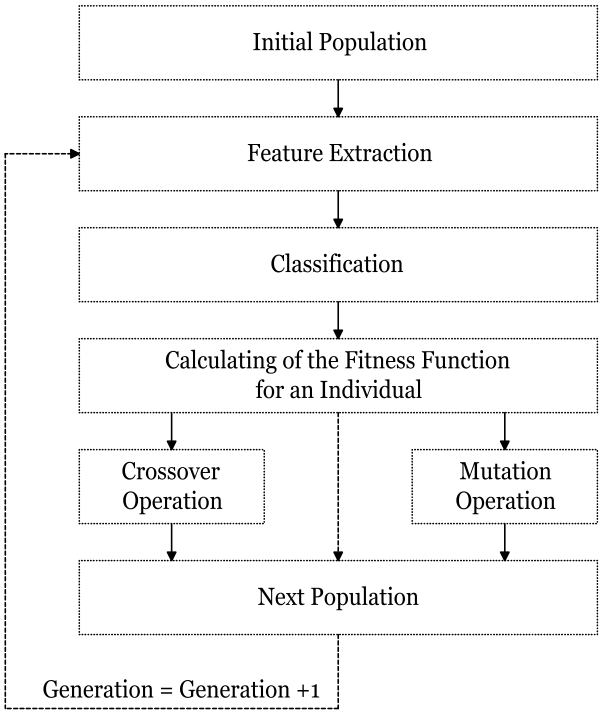


Fig. 7. Algorithm of the GWPNN.

of the population. The values which ε parameter can take are 1–8. According to this, these values are converted to binary form as below:

For selecting feature extraction methods:

Feature extraction method-1 → 0 0,	Feature extraction method-2 → 0 1,
Feature extraction method-3 → 1 0,	Feature extraction method-4 → 1 1.

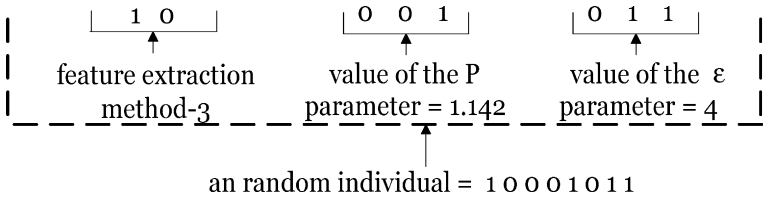
For selecting optimum P parameter value of the norm entropy:

1 → 0 0 0,	1.142 → 0 0 1,	1.285 → 0 1 0,	1.426 → 0 1 1,
1.568 → 1 0 0,	1.71 → 1 0 1,	1.852 → 1 1 0,	1.994 → 1 1 1

For selecting optimum ε parameter value of the sure entropy:

1 → 0 0 0,	2 → 0 0 1,	3 → 0 1 0,	4→0 1 1,
5 → 1 0 0,	6 → 1 0 1,	7 → 1 1 0,	8 → 1 1 1

For example, one individual of the population may be shown as below:



• *Feature extraction:* The realized operations in this stage are ordered as below:

1. Each of the individuals of the population represent one of the feature extraction methods which are mentioned in Section 3.3, a P parameter value of the norm entropy which is mentioned in Section 3.2, and a ε parameter value of the sure entropy which is mentioned in Section 3.2. When a random individual of the population is selected, the feature extraction method, value of the P parameter, and value of the ε parameter which are represented by this individual are sent to the feature extraction stage.
2. The feature extraction process is realized as appropriate to feature extraction method, value of the P parameter, and value of the ε parameter, which are represented by the relevant individual in this feature extraction mechanism. Thirty-five English language words signals are used for feature extraction mechanism to each of the 40 speakers. Thus, total 1400 English speech words signals are processed in this feature extraction mechanism. Each of these 1400 signals has 128 wavelet decomposition coefficients. Each of these 128 wavelet packet coefficients has values of the norm entropy, sure entropy, and logarithmic energy entropy. These wavelet packet entropy values were obtained by using the feature extraction method, value of the P parameter, and value of the ε parameter of the relevant individual. The norm entropy values of the waveforms are calculated at the terminal node signals obtained from wavelet packet decomposition as defined in Eq. (4). Where, the wavelet entropy E is a real number, s is the terminal node signal and (s_i) is the waveform of terminal node signals. In norm entropy, P is the power and must be such that $1 \leqslant P < 2$. In the same way, is calculated the sure entropy and logarithmic energy entropy as defined in Eqs. (5) and (6) of the waveforms at the terminal node signals obtained from wavelet packet decomposition. In sure entropy, ε is the threshold and must be such that $1 \leqslant \varepsilon \leqslant 8$. All obtained entropy values are normalized by dividing by $N = 50$. Thus, total normalized 384 entropy values are found for each of these 1400 signals. At the same time, each of these entropy values is called as wavelet packet entropy. As a result, 1400×384 wavelet packet entropy feature vector is obtained in this feature extraction mechanism.
3. The input numbers were reduced from 128 to 6 for making easier training of multi-layer perception (MLP) artificial neural network to each of the inputs of the English

Table 3
MLP architecture and training parameters

Architecture	
The number of layers	3
The number of neurons on the layers	Input: 18, hidden: 60, output: 40
The initial weights and biases	The Nguyen–Widrow method
Activation functions	Log-sigmoid
Training parameters learning rule	Back-propagation
Adaptive learning rate	Initial: 0.0001, increase: 1.05, decrease: 0.7
Momentum constant	0.98
Sum-squared error	0.000001

language speech signals. When input numbers increase, network structure of this MLP increases. This large value of the input and hidden layers of this MLP causes more complex network structure and more difficult training of MLP. Therefore, there is used reduced numbers of the inputs whose numbers are 6 for training of MLP in this study. These six inputs are maximum values, minimum values, arithmetic average values, geometric average values, standard deviation values, and variance (Boashash, 1992) values, respectively of the 128 wavelet packet entropy values for each of the input English language speech signals. Total $3 \times 6 = 18$ wavelet packet entropy values calculated for each of these input speech signals due to three entropy types (norm, sure and logarithmic energy) are taken care for an input speech signal.

• *Classification:* This mechanism realizes the intelligent classification by using features obtained from feature extraction stage. The training parameters and the structure of the MLP used in this study are as shown in Table 3. These were selected for the best performance, after several different experiments, such as the number of hidden layers, the size of the hidden layers, value of the moment constant and learning rate, and type of the activation functions.

The realized operations in this stage are ordered as below:

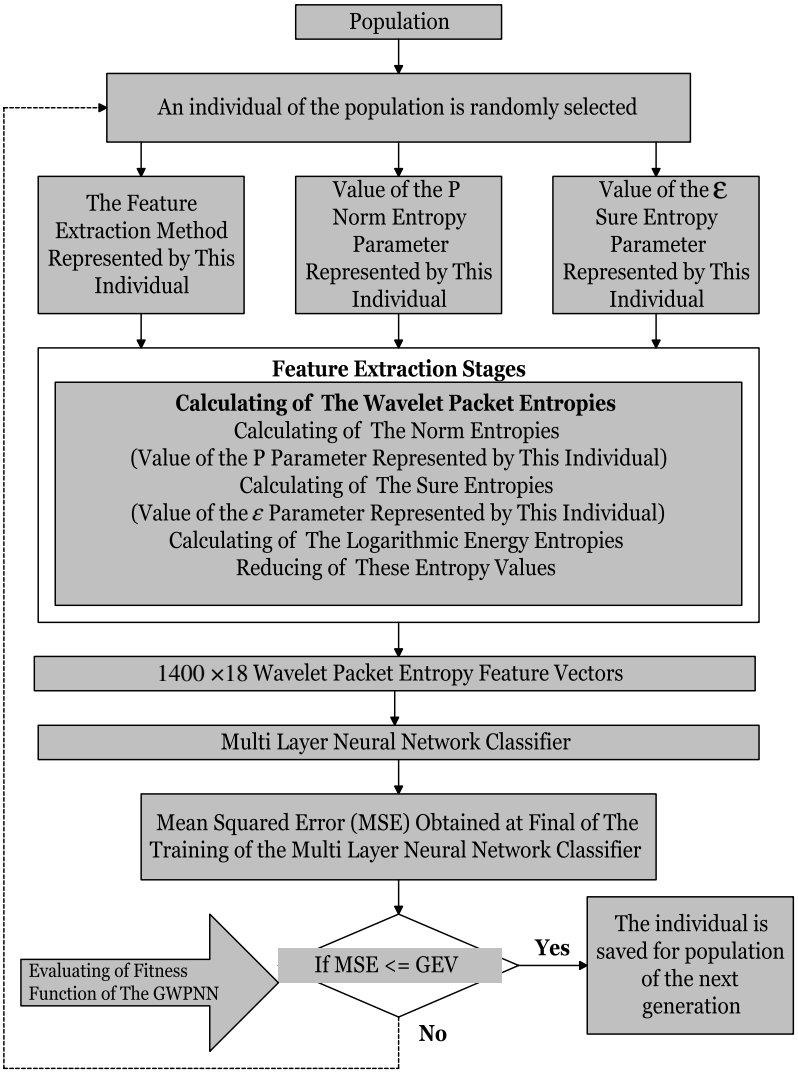


Fig. 8. Feature extraction and classification flow diagram of the GWPNN.

Table 4
The obtained optimum values by using GWPNN algorithm and classification performances

Obtained optimum method no.	Generation number	The selected feature extraction method	The obtained optimum P parameter value	The obtained optimum ε parameter value	The obtained MSE value of the MLP	The classification performance of the GWPNN (%)
1	8	WPD-STFT	1.142	4	0.000000014	90.51
2	5	WPD	1.285	2	0.00000059	91.72
3	4	WPD-BJTFR	1	3	0.00000017	90.64

1. 1400×18 feature vector which is obtained in feature extraction mechanism is given to the input of the MLP classification. The decision space at the output of MLP classifier is formed from 40 used individual speakers.
2. The mean square error (MSE) of the ANN is obtained at the final of the training of the ANN classifier.
3. There is a goal error value (GEV) of the MLP classifier. It is wanted as the MSE (mean squared error) value which is obtained from the final training of the MLP classifier made equal to GEV or less than GEV. This comparison of the MSE–GEV is used for genetic algorithm as fitness function of an individual of the population.

• *Calculation of the fitness function for an individual:* If MSE value which is obtained from the final of the MLP classifier is made equal to GEV or less than GEV for a random individual, fitness value of this individual is high. If else fitness value of this individual is low. In this application, fitness values of all the individuals of the population are calculated in the same way. Later, obtained fitness values of all the individuals at the population are ordered from 1 to 20. If fitness value of an individual is less than 11, fitness value of this individual is evaluated as low. If else, fitness value of that individual is evaluated as high. The individuals which have high fitness values at the current population are saved for the population in the next generation. The individuals which have low fitness values at the current population are eliminated. If there are individuals which have same fitness value, one of these individuals is selected as random for the population in the next generation. As a result, optimum 10 individuals of current population are saved for the population in the next generation.

• *Crossover operation:* Crossover portion used in this study is 30% portion of the optimum 10 individuals (three individuals) obtained in calculation of the fitness function for an individual stage. These three individuals are randomly selected and subjected to crossover operator. Two bits of each of the random two individuals are randomly selected and replaced with each other for crossover operations. At the final of the crossover operations, six new individuals are obtained.

• *Mutation operation:* In there, the bit inversion is used as mutation operator.[1] The mutation operation is realized by using 0.3% portion of the total bits numbers of left over seven individuals. If a bit is equal to 1, it is changed to 0. If it is equal to 0, it is changed to 1. At the final of the mutation operation, seven new individuals are obtained. There are total 20 individuals in the population of the next generation at the final of these stages. The processes in these stages are realized respectively for each of the 20 individuals

Table 5
Classification performance of the GWPNN for obtained optimum method no. 1

Speakers	Total number of samples	Correct classification #	Incorrect classification #	The average recognition (%)
Sp # 1	140	123	17	87.85
Sp # 2	140	112	28	80
Sp # 3	140	117	23	83.57
Sp # 4	140	129	11	92.14
Sp # 5	140	137	3	97.85
Sp # 6	140	118	22	84.28
Sp # 7	140	126	14	90
Sp # 8	140	132	8	94.28
Sp # 9	140	111	29	79.28
Sp # 10	140	113	27	80.71
Sp # 11	140	125	15	89.28
Sp # 12	140	132	8	94.28
Sp # 13	140	114	26	81.42
Sp # 14	140	122	18	87.14
Sp # 15	140	126	14	90
Sp # 16	140	116	24	82.85
Sp # 17	140	121	19	86.42
Sp # 18	140	135	5	96.42
Sp # 19	140	121	19	86.42
Sp # 20	140	136	4	97.14
Sp # 21	140	123	17	87.85
Sp # 22	140	117	23	83.57
Sp # 23	140	108	32	77.14
Sp # 24	140	113	27	80.71
Sp # 25	140	119	21	85
Sp # 26	140	121	19	86.42
Sp # 27	140	124	16	88.57
Sp # 28	140	116	24	82.85
Sp # 29	140	123	17	87.85
Sp # 30	140	107	33	76.42
Sp # 31	140	125	15	89.28
Sp # 32	140	115	25	82.14
Sp # 33	140	119	21	85
Sp # 34	140	121	19	86.42
Sp # 35	140	128	12	91.42
Sp # 36	140	131	9	93.57
Sp # 37	140	129	11	92.14
Sp # 38	140	104	36	74.28
Sp # 39	140	115	25	82.14
Sp # 40	140	118	22	84.28
Total	5600	4842	758	86.46

in the population for a generation. Flow diagram of feature extraction and classification mechanism of the GWPNN is given in Fig. 8.

Main goal of using GWPNN is selecting the most appropriate feature extraction method from among four different feature extraction methods, optimum value of P norm entropy parameter, and optimum value of ε sure entropy parameter. The combination of selecting these values are given in feature extraction stage.

4. Experimental results

The architecture and training parameters of used MLP are given in Table 3. The experiments are performed using total 5600 English language word signals of 40 individual speakers. For each of these speakers, 140 English speech signals were used in which 35 of these signals are noiseless English language words and other 105 signals are noisy English language words signals, which have different white-noise amplitudes (signal/noise rate (SNR) = −2 dB, −3 dB, and −5 dB). One thousand four hundred of these 5600 signals were used for optimum feature extraction and classification stages on the GWPNN. The left over of these 4200 signals were used for testing performance of the GWPNN. In these experiments, 100% correct classification was obtained at the GWPNN training among the 40 different speakers signal classes. It clearly indicates the effectiveness and the reliability of the proposed approach

Table 6
Classification performance of the GWPNN for obtained optimum method no. 2

Speakers	Total number of samples	Correct classification #	Incorrect classification #	The average recognition (%)
Sp # 1	140	114	26	81.42
Sp # 2	140	129	11	92.14
Sp # 3	140	113	27	80.71
Sp # 4	140	101	39	72.14
Sp # 5	140	123	17	87.85
Sp # 6	140	132	8	94.28
Sp # 7	140	117	23	83.57
Sp # 8	140	119	21	85
Sp # 9	140	107	33	76.42
Sp # 10	140	121	19	86.42
Sp # 11	140	126	14	90
Sp # 12	140	129	11	92.14
Sp # 13	140	116	24	82.85
Sp # 14	140	119	21	85
Sp # 15	140	123	17	87.85
Sp # 16	140	125	15	89.28
Sp # 17	140	111	29	79.28
Sp # 18	140	127	13	90.71
Sp # 19	140	103	37	73.57
Sp # 20	140	124	16	88.57
Sp # 21	140	128	12	91.42
Sp # 22	140	112	28	80
Sp # 23	140	127	13	90.71
Sp # 24	140	132	8	94.28
Sp # 25	140	115	25	82.14
Sp # 26	140	119	21	85
Sp # 27	140	124	16	88.57
Sp # 28	140	133	7	95
Sp # 29	140	129	11	92.14
Sp # 30	140	121	29	86.42
Sp # 31	140	135	5	96.42
Sp # 32	140	124	16	88.57
Sp # 33	140	121	19	86.42
Sp # 34	140	127	13	90.71
Sp # 35	140	113	27	80.71
Sp # 36	140	104	36	74.28
Sp # 37	140	127	13	90.71
Sp # 38	140	114	26	81.42
Sp # 39	140	128	12	91.42
Sp # 40	140	132	8	94.28
Total	5600	4844	756	86.50

Table 7
Classification performance of the GWPNN for obtained optimum method no. 3

Speakers	Total number of samples	Correct classification #	Incorrect classification #	The average recognition (%)
Sp # 1	140	126	14	90
Sp # 2	140	113	27	80.71
Sp # 3	140	121	19	86.42
Sp # 4	140	107	33	76.42
Sp # 5	140	115	25	82.14
Sp # 6	140	129	11	92.14
Sp # 7	140	111	29	79.28
Sp # 8	140	123	17	87.85
Sp # 9	140	105	35	75
Sp # 10	140	132	8	94.28
Sp # 11	140	124	16	88.57
Sp # 12	140	127	13	90.71
Sp # 13	140	109	31	77.85
Sp # 14	140	130	10	92.85
Sp # 15	140	123	17	87.85
Sp # 16	140	112	28	80
Sp # 17	140	123	17	87.85
Sp # 18	140	103	37	73.57
Sp # 19	140	132	8	94.28
Sp # 20	140	112	28	80
Sp # 21	140	115	25	82.14
Sp # 22	140	127	13	90.71
Sp # 23	140	123	17	87.85
Sp # 24	140	109	31	77.85
Sp # 25	140	113	27	80.71
Sp # 26	140	128	12	91.42
Sp # 27	140	117	23	83.57
Sp # 28	140	112	28	80
Sp # 29	140	121	19	86.42
Sp # 30	140	111	29	79.28
Sp # 31	140	129	11	92.14
Sp # 32	140	131	9	93.57
Sp # 33	140	104	36	74.28
Sp # 34	140	121	19	86.42
Sp # 35	140	123	17	87.85
Sp # 36	140	109	31	77.85
Sp # 37	140	131	9	93.57
Sp # 38	140	113	27	80.71
Sp # 39	140	121	19	86.42
Sp # 40	140	117	23	83.57
Total	5600	4752	848	84.85

for extracting optimum features obtained from English language speech signals. The most appropriates of feature extraction methods obtained are mentioned in Section 3.3, optimum values of P parameter, and optimum values of ε parameter by using GWPNN algorithm are given in Table 4.

At the same time, Table 4 shows classification performance of the GWPNN algorithm. Classification performances of the GWPNN for obtained optimum method nos. 1–3, are given in Tables 5–7, respectively.

5. Conclusions

In speech/speaker recognition area, the presented novelties of this study can be summarized as follow:

First novelty: For selecting optimum feature extraction method, classification was performed using the GWPNN which has an effectively adaptive feature extraction and classification method that increases percentage of the speaker recognition.

Second novelty: The presented second novelty in this paper is using of the genetic-wavelet packet-neural network model for selecting the feature extraction method and finding the optimum wavelet packet entropy parameter values which are used for obtaining the wavelet packet entropy values in wavelet packet layer as a new and efficient method in automatic speaker recognition area.

The wavelet packet decomposition has been demonstrated to be an effective tool for extracting information from the real English speech signals. The proposed feature extraction methods which are mentioned in Section 3.3 are robust against to noise in the speech signals.

In this study, was developed an expert system for the interpretation of the English language speech signals using pattern recognition and the speaker recognition performance of this method demonstrated on the 40 individual speakers. The stated results show that the proposed method can make an effective interpretation. The performance of the expert system is given in Table 4. Classification performances of the GWPNN for obtained optimum method nos. 1–3. are given in Tables 5–7, respectively.

In this study, the application of the wavelet packet entropy in the feature extraction mechanism of GWPNN to the optimum feature extraction from speech signals is shown. Wavelet packet entropy proved to be a very useful feature for characterizing the speech signal, furthermore the information obtained from the wavelet packet entropy is related to the energy and consequently with the amplitude of the signal. This means that with this method, new information can be accessed with an approach different from the traditional analysis of amplitude of speech signal.

The recognition performances of this paper show the advantages of this system: it is rapid, easy to operate, and not expensive. This system offers advantage in commercial and security applications. The most important aspect of the expert system is the ability of self-organization of the GWPNN without requirements of programming and the immediate response of a trained net during real-time applications. These features make the expert system suitable for automatic classification in interpretation of the English language speech signals. These results point out the ability of designing of a new expert speaker recognition assistance system.

Even though this expert speaker system was carried out on the English language speech signals, similar results for the other languages speech signals recognition studies can be expected. Besides the feasibility of a real-time implementation of the expert system, by increasing the variety and number of speech signals additional information (i.e., quantification of the data length) can be provided for speaker recognition.

References

Alotaibi, Y. A. (2005). Investigating spoken Arabic digits in speech recognition setting. *Information Sciences, 173*(1–3), 115–139.

Avci, E., & Turkoglu, I. (2003). Modelling of tunnel diode by adaptive-network-based fuzzy inference system. *International Journal of Computational Intelligence, 1*(1), 231–233.

Avci, E., Turkoglu, I., & Poyraz, M. (2005a). Intelligent target recognition based on wavelet packet neural network. *Experts Systems with Applications, 29*(1).

Avci, E., Turkoglu, I., & Poyraz, M. (2005b). A new approach based on scalogram for automatic target recognition with X-band Doppler radar. *Asian Journal of Information Technology, 4*(1), 133–140.

Avci, E., Turkoglu, I., & Poyraz, M. (2005c). Intelligent target recognition based on wavelet adaptive network based fuzzy inference system. *Lecture notes in computer science* (vol. 3522/2005, pp. 594–601). Springer-Verlag.

Bishop, C. M. (1996). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Boashash, B. (1992). *Time–frequency signal analysis methods and applications*. Cheshire: Longman.

Buckheit, J. B., & Donoho, D. L. (1995). Wavelab and reproducible research. In *Wavelets and statistics*. New York: Springer-Verlag.

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

Coifman, R. R., & Wickerhauser, M. V. (1992). Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory, 38*(2), 713–718.

Evangelista, G. (1993). Pitch-synchronous wavelet representations of speech and music signals. *IEEE Transactions on Signal Processing, 41*(12).

Evangelista, G. (1994). Comb and multiplexed wavelet transforms and their application to speech processing. *IEEE Transactions on Signal Processing, 42*(2).

Ha, Q. P., Tran, T. H., & Dissanayake, G. (2005). A wavelet and neural network based voice interface system for wheelchair control. *International Journal of Intelligent Systems Technologies and Applications, 1*(1–2).

Haykin, S. (1994). *Neural networks, a comprehensive foundation*. New York: Macmillan College Publishing Company Inc.

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

Jakubiak, A., Arabas, J., Grabczak, K., et al. (1997). Radar clutter classification using Kohonen neural network. Radar 97 (Conf. Publ. no. 449), Edinburgh, UK, pp. 185–188.

Kadambe, S., & Boudreaux-Bartels, G. F. (1992). Application of the wavelet transform for pitch detection of speech signals. *IEEE Transactions on Information Theory, 32*, 712–718.

Kadambe, S., & Srinivasan, P. (1994). Applications of adaptive wavelets for speech. *Optical Engineering, 33*(7), 2204–2211.

Lippmann, R. (1989). *Review of neural networks for speech recognition, neural computation* (pp. 1–38). MIT press.

Maes, S. (1994). Nonlinear techniques for parameter extraction from quasi-continuous wavelet transform with application to speech. In *Proceedings of SPIE – The International Society for Optical Engineering* (Vol. 2093, pp. 8–19).

Mallat, S. A. (1989). Theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*, 674–693.

Mallat, S., & Zhong, S. (1992). Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14*, 710–732.

MATLAB 5.3 version Wavelet Toolbox, MathWorks Company.

Quiroga, R. Q. (1998). *Quantitative analysis of EEG signals: Time–frequency methods and Chaos theory*. Lübeck: Institute of Physiology, Medical University.

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

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

Szu, H., Telfer, B., & Kadambe, S. (1992). Neural network adaptive wavelets for signal representation and classification. *Optical Engineering, 31*(9), 1907–1916.

Thuillard, M. (2000). A review of wavelet networks, wavenets, fuzzy wavenets and their applications. In *ESIT'2000, Aachen, Germany, September 14–15* (pp. 5–16).

Turkoglu, I., Arslan, A., & Ilkay, E. (2003). An intelligent system for diagnosis of the heart valve diseases with wavelet packet neural networks. *Computer in Biology and Medicine, 33*, 319–331.

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

Wang, L., Teo, K. K., & Lin, Z. (2001). Predicting time with wavelet packet neural networks. In *International joint conference on neural networks, proceedings of the IJCNN'01, INNS-IEEE, Washington, DC* (Vol. 3, pp. 1593–1597).

Wesfried, E., & Wickerhauser, M. V. (1993). Adapted local trigonometric transforms and speech processing. *IEEE SP, 41*, 3597–3600.

Zhang, Q., & Benveniste, A. (1992). Wavelet network. *IEEE Transactions on Neural Networks, 3*(6), 889–898.

Zhang, J., Walter, G. G., Miao, Y., & Lee, W. N. W. (1995). Wavelet neural networks for function learning. *IEEE Transactions on Signal Processing, 43*(6).