Computers and Electrical Engineering 38 (2012) 1467–1479

Contents lists available at SciVerse ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Average framing linear prediction coding with wavelet transform for text-independent speaker identification system ☆

Khaled Daqrouq [a,*], Khalooq Y. Al Azzawi [b]

[a] Electrical & Computer Eng. Department, King Abdulaziz University, Jeddah, Saudi Arabia
[b] Electromechanical Engineering Dept., Univ. of Technology, Baghdad, Iraq

ARTICLE INFO

Article history:
Received 19 June 2011
Received in revised form 22 April 2012
Accepted 23 April 2012
Available online 17 May 2012

ABSTRACT

In this work, an average framing linear prediction coding (AFLPC) technique for text-independent speaker identification systems is presented. Conventionally, linear prediction coding (LPC) has been applied in speech recognition applications. However, in this study the combination of modified LPC with wavelet transform (WT), termed AFLPC, is proposed for speaker identification. The investigation procedure is based on feature extraction and voice classification. In the phase of feature extraction, the distinguished speaker's vocal tract characteristics were extracted using the AFLPC technique. The size of a speaker's feature vector can be optimized in term of an acceptable recognition rate by means of genetic algorithm (GA). Hence, an LPC order of 30 is found to be the best according to the system performance. In the phase of classification, probabilistic neural network (PNN) is applied because of its rapid response and ease in implementation. In the practical investigation, performances of different wavelet transforms in conjunction with AFLPC were compared with one another. In addition, the capability analysis on the proposed system was examined by comparing it with other systems proposed in literature. Consequently, the PNN classifier achieves a better recognition rate (97.36%) with the wavelet packet (WP) and AFLPC termed WPLPCF feature extraction method. It is also suggested to analyze the proposed system in additive white Gaussian noise (AWGN) and real noise environments; 58.56% for 0 dB and 70.52% for 5 dB. The recognition rates for the whole database of the Gaussian mixture model (GMM) reached the lowest value in case of small number of training samples.

1. Introduction

Automatic speech recognition (ASR) has been studied by a large number of researchers for about four decades [1]. From a commercial viewpoint, ASR is a tool with a potentially large market due to its wide range of application from the automation of operator-assisted service to speech-to-text aiding systems for hearing-impaired individuals [2].

A commonly used technique for feature extraction is based on the Karhunen–Loeve transform (KLT) [10]. These models have been applied to text-independent speaker recognition cases [3] with exceptional results. Karhunen–Loeve transform is the optimal transform according to minimum mean square error (MMSE) and maximal energy packing. Most of the suggested speaker identification systems use Mel frequency cepstral coefficient (MFCC) [5] and linear predictive cepstral

☆ Reviews processed and approved for publication by Editor-in-Chief Dr. Manu Malek.
* Corresponding author. Address: P.O. Box 80204, Jeddah 21589, Saudi Arabia. Tel.: +966 5 66 980400; fax: +966 5 695 268.
E-mail address: haleddaq@yahoo.com (K. Daqrouq).

0045-7906/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compeleceng.2012.04.014

coefficient (LPCC) [6] as features. Although MFCC and LPCC have proved to be two very good features in speech recognition, the disadvantage of the MFCC is that it uses short time Fourier transform, which has a weak time–frequency resolution and an assumption that the signal is stationary. Therefore it is relatively difficult to recognize plosive phonemes by these features. Currently, some researches [7–9] are focusing on the wavelet transform for speaker feature extraction.

Wavelet transform [4,3,11] has been extensively considered in the last two decades and has been widely utilized in various areas of science and engineering. The wavelet analysis process is implemented with dilated and translated versions of a mother wavelet. Since signals of interest can generally be expressed using wavelet decompositions, signal processing algorithms can be implemented by adjusting only the corresponding wavelet coefficients. From a mathematical point of view, the scale parameter of a wavelet can be a positive real value and the translation can be an arbitrary real number [1]. From a practical point of view, however, in order to improve computation efficiency, the values of the shift and scale parameters are often limited to some discrete lattices [12,13].

Wavelet and WP analysis have been proven as effectual signal processing techniques for a variety of digital signal processing problems. Wavelets have been used in two different methods in feature extraction plans designed for the task of speech/voice recognition. Discrete wavelet transform in place of discrete cosine transform is utilized for the feature extraction period in the first method [16]. In the second method, wavelet transform is used directly on the speech/voice signals and either wavelet coefficients containing high energy are extracted as features [8] but suffer from shift variance, or sub band energies are used instead of the Mel filter-bank sub band energies proposed in [17]. Particularly, WP bases are used in [18] as close approximations of the Mel-frequency division using Daubechies orthogonal filters. In [19], a feature extraction method based on the wavelet Eigen function was proposed. Wavelets can offer a significant computational benefit by reducing the dimensionality of the Eigen value problem. A text-independent speaker identification system based on improved wavelet transform is proposed in [9], where learning of the correlation between the wavelet transform and the expression vector is performed by kernel canonical correlation analysis.

The wavelet packets transform (WPT) performs the recursive decomposition of the speech signal obtained by the recursive binary tree. Basically, the WPT is very similar to discrete wavelet transform (DWT). However, WPT decomposes both details and approximations instead of only performing the decomposition process on approximations. WPT features have superior presentation than those of the DWT [19]. Nevertheless, as the number of wavelet packet bases grows, the time required to appropriately classify the database will become nonlinear. Consequently, dimensionality decreasing becomes a significant issue. Selecting a beneficial and relevant subset of features from a larger set is crucial to enhance the performance of speaker recognition [20,21]. A feature selection scheme is, therefore, needed to choose the most valuable information from the complete feature space to form a feature vector in a lower-dimensionality, and take away any redundant information that may have disadvantageous effects on the classification quality. To select an appropriate set of features, a criterion function can be used to provide the discriminatory power of the individual features.

The wavelet packet perceptual decomposition tree was first proposed by Sarikaya [22] and yields the wavelet packet parameters (WPP). In [24], the energy indexes of DWT or WPT were proposed for speaker identification, where WPT was superior in terms of recognition rate. Sure entropy was calculated for the waveforms at the terminal node signals obtained from DWT [25,60] for speaker identification.

Neural network applications for classification have been considered in recent years [30,15]. They are widely applied in data analysis and speaker identification. The advantage of the artificial neural network is that the transfer function between the input vectors and the target matrix (output) does not have to be predicted in advance. Artificial neural network performance depends mainly on the size and quality of training samples [28,29]. When the number of training data is small, not representative of the possible space, standard neural network results are poor. Fuzzy theory has been used successfully in many applications to reduce the dimensionality of feature vector [31]. There are many kinds of artificial neural network models, among which the back-propagation neural network (BPNN) model is the most widely used [32]. The generalized regression neural network (GRN) was introduced by [32]. Ganchev et al. [35] proposed a probabilistic neural network for speaker identification.

In fact, LPC is popular and widely used because its coefficients representing a speaker by modeling vocal tract parameters and the data size are very suitable for speaker and speech recognition. Many algorithms were developed to find a better representation of a speaker by means of a linear predictive coding technique [37,38,23]. The predictor coefficients themselves are rarely utilized as features, but they are transformed into robust and less correlated features such as linear predictive cepstral coefficients (LPCCs) [39], line spectral frequencies (LSFs) [40], and perceptual linear prediction (PLP) coefficients [41]. PLP is known as a state of the art for speech recognition task. Other, somewhat less effective features include partial correlation coefficients (PARCORs), log area ratios (LARs) and formant frequencies and bandwidths [42,56]. In the present work, the focus will be on modifying LPC coefficients and reducing the dimensionality of feature vectors.

In this research, the authors improve an effectual and a novel feature extraction method for text-independent systems, taking in consideration that the size of neural network input is a very crucial issue. This affects quality of the training set. For this reason, the presented features extraction method offers a reduction in the dimensionality of speech signals. The proposed method is based on average framing LPC in conjunction with WT upon suitable level with an appropriate wavelet function (Daubechies-type1, which is known as Haar function). For classification, PNN is proposed to accomplish online operations in a speedy manner.

2. Wavelet transform using in speaker feature extraction

2.1. Wavelet packet transform feature extraction method

To decompose the speech signal into wavelet packet transform (WPT), we start from the common form of the equivalent low pass of discrete time speech signal

$$u(t) = \sum_m X_m p(t - mT), \tag{1}$$

where X_m is a sequence of discrete speech signal values, which are obtained by a data acquisition stage; the signal $p(t)$ is a pulse, whose figure represents an important signal design problem when there is a bandwidth restriction on the channel; and T is the sampling time. Considering that $\phi(t - mT)$ is a scaling function of a wavelet packet, i.e., $\phi \in W_{2^N}^0$, then a finite set of orthogonal subspaces can be constructed as [47,48].

$$W_{2^N}^0 = \bigoplus_{(l,n)\in\rho N} W_{2^l}^0, \tag{2}$$

where $W_{2^N}^0 \subset L^2(R)$, $\rho N = \{(l,n)\}$ is a dyadic interval that forms a disjoint covering of $[0, 2^N]$, $W_{2^l}^n$, denoting the closed linear span of process $\sqrt{2^l}\psi_n(2^l t - m)$, $m \in Z$, and $\{\psi_n(t)\}_{n\in N}$ is called the wavelet packet, considered by the scaling function ϕ. Therefore, the speech signal model in (1) is customized as

$$u(t) = \sum_m \sum_{(l,n)\in\rho N} X_m \sqrt{2^l}\psi_n(2^l t - m). \tag{3}$$

The speech signal model in (3) is the basic form of wavelet packet transform, which is used in signal decomposition. The signal is carried by orthogonal functions, which shape a wavelet packet composition in $W_{2^N}^0$ space. We may use the discrete wavelet packet transforms (DWPT) procedure as

$$\phi_{l+1}^{2n}(i) = \sum_{k\in Z} h(k - 2i)\phi_l^n(k), \tag{4}$$

$$\phi_{l+1}^{2n+1}(i) = \sum_{k\in Z} g(k - 2i)\phi_l^n(k), \tag{5}$$

where $\phi_{l+1}^n \in W_{2^{l+1}}^n$ and $\phi_l^n \in W_{2^l}^n$. These two processes can be carried out recursively by proceeding through the binary tree structure, with $0(N \log N)$ computational complexity. Using (3)–(5), the coefficients of the linear combination may be shown to be the reversed versions of the decomposition sequences $h[k]$ and $g[k]$ (with zero padding), respectively. Continuously, we can reconstruct $\phi_0^1(i)$ via the terminal functions of an arbitrary tree-structured decomposition:

$$\phi_0^1(i) = \sum_{l\in L, n\in C_l} \sum_{k\in Z} f_{ln}(i - 2^l k)\phi_l^n(k), \tag{6}$$

where L is the set of levels having the terminals of a given tree; C_l is the set of indices of the terminals at the lth level; and $f_{ln}[i]$ is the equivalent sequence generated from the combination of $h[k]$, $g[k]$ and decimation operation, which leads from the root to the (l, n)th terminal, i.e.,

$$\phi_l^n(i) = \sum_{k\in Z} f_{ln}(k - 2^l i)\phi_0^1(k). \tag{7}$$

For a certain tree structure, the function ϕ_l^n in (7) is called the constituent terminal function of ϕ_0^1. In this work, the tree consists of two stages, and therefore has three high pass nodes and three low pass nodes.

The wavelet packet is used to extract additional features to guarantee a higher recognition rate. In this study, WPT is applied at the stage of feature extraction, but these data are not proper for classification due to a great amount of data length (for example, a speech signal with a number of 35,582 samples will reach 71,166 after WPT decomposition at level two). Thus, we have to seek for a better representation of the speech features. Avci et al. [27] proposed a method to calculate the entropy value of the wavelet norm in digital modulation recognition. In the biomedical field, Behroozmand and Almasganj [49] presented a combination of genetic algorithm and wavelet packet transform used in the pathological evaluation, and the energy features are determined from a group of wavelet packet coefficients. Sarikaya and Hansen [50] proposed a robust speech recognition scheme in a noisy environment by using wavelet-based energy as a threshold for denoising estimation. In [24], the energy indexes of WP were proposed for speaker identification. Sure entropy is calculated for the waveforms at the terminal node signals obtained from DWT [25] for speaker identification. Avci [26] proposed a features extraction method for speaker recognition based on a combination of three entropies types (sure, logarithmic energy and norm). In this paper, we use LPCC obtained from WP tree nodes for speaker feature vector constructing to be used for speaker identification.

2.2. Discrete wavelet transform feature extraction method

The DWT indicates an arbitrary square integrable function as a superposition of a family of basic functions. These functions are wavelet functions. A family of wavelet basis functions can be produced by translating and dilating the mother wavelet [14]. The DWT coefficients can be generated by taking the inner product between the original signal and the wavelet functions. Since the wavelet functions are translated and dilated versions of each other, a simpler algorithm, known as Mallat's pyramid tree algorithm, has been proposed (see Fig. 2) [14].

The DWT can be utilized as the multi-resolution decomposition of a sequence. It takes a length N sequence $a(n)$ as the input and produces a length N sequence as the output. The output $N/2$ has values at the highest resolution (level 1) and $N/4$ values at the next resolution (level 2), and so on. Let $N = 2^m$, and let the number of frequencies, or resolutions, be m, while

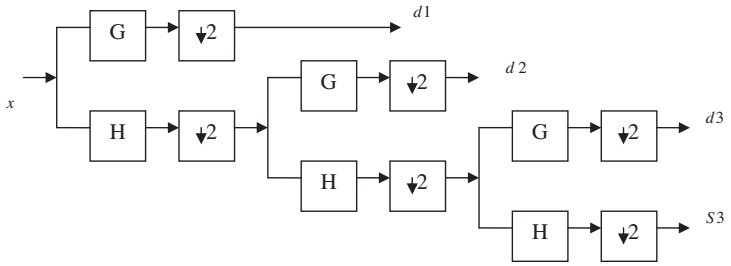

Fig. 1. DWT-tree by Mallat's algorithm.

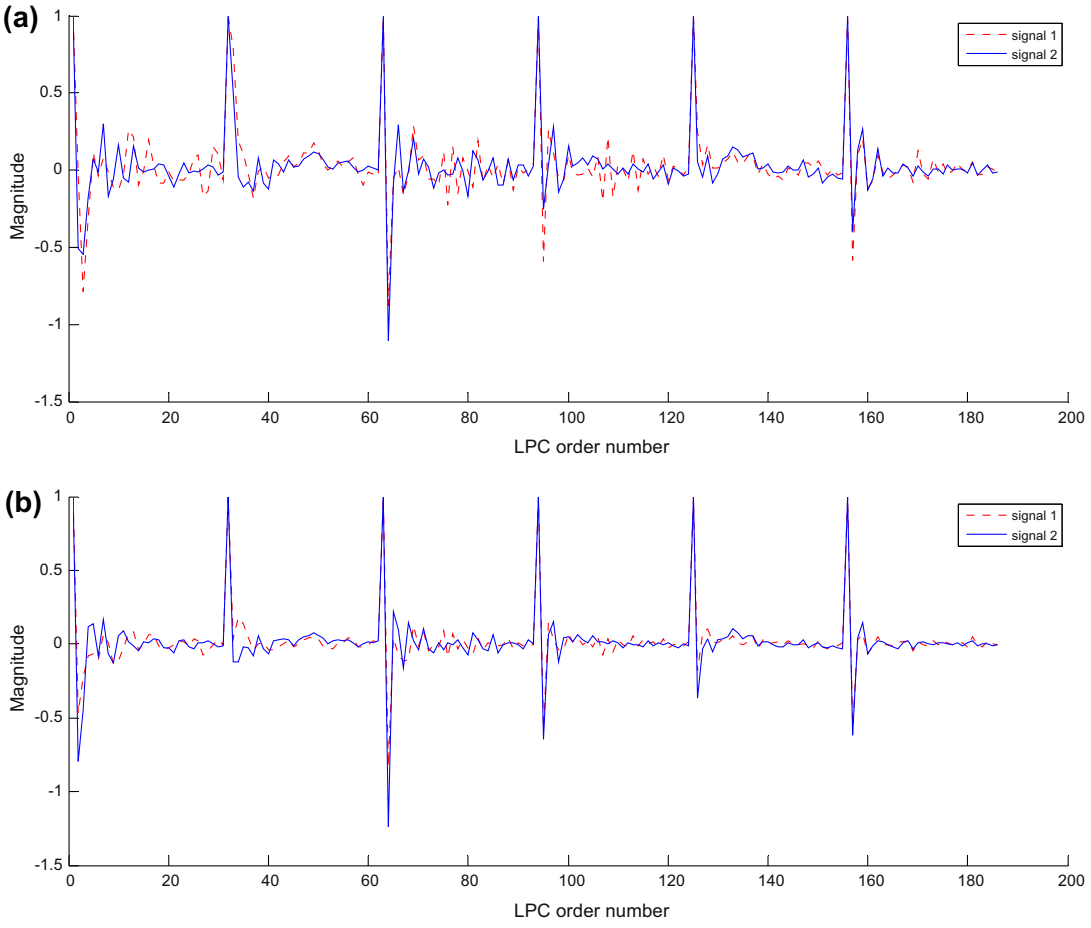

Fig. 2. Two feature vectors taken for a single speaker: (a) illustrates feature vectors using LPC from WP at level two, and (b) illustrates feature vectors using AFLPC from WP at level two.

bearing in mind that $m = \log N$ octaves. So the frequency index k varies as $1, 2, \ldots, m$ corresponds to the scales $2^1, 2^2, \ldots, 2^m$. As described by the Mallat pyramid algorithm (Fig. 1), the DWT coefficients of the previous stage are expressed as follows [51]:

$$W_L(n,k) = \sum_i W_L(i,k-1)h(i-2n), \tag{8a}$$

$$W_H(n,k) = \sum_i W_L(i,k-1)g(i-2n), \tag{8b}$$

where $W_L(p,j)$ is the pth scaling coefficient at the jth stage, $W_H(p,j)$ is the pth wavelet coefficient at the jth stage, and $h(n)$, $g(n)$ are the dilation coefficients relating to the scaling and wavelet functions, respectively.

In the last decade, there has been an enormous increase in the applications of wavelets in various scientific fields [59]. Typical applications of wavelets include signal processing, image processing, security systems, numerical analysis, statistics, biomedicine, etc. Wavelet transform tenders a wide variety of useful features, on the contrary to other transforms, such as Fourier transform or cosine transform. Some of these are as follows:

- Adaptive time–frequency windows.
- Lower aliasing distortion for signal processing applications.
- Computational complexity of $O(N)$, where N is the length of data.
- Inherent scalability.

Delac et al. [52] proposed DWT for face recognition. In [16,31], the use of DWT for speech recognition, which has a good time and frequency resolution, is proposed instead of the discrete cosine transform (DCT) to solve the problem of high frequency artifacts being introduced due to abrupt changes at window boundaries. The features based on DWT and WPT were chosen to evaluate the effectiveness of the selected feature for speaker identification [23]. Daqrouq [28] stated that the use of a DWT approximation sub signal via several levels instead of the original imposter had good performance on AWGN facing, particularly on levels 3 and 4 in the text-independent speaker identification system. Therefore, we use LPCC obtained from DWT tree nodes for speaker feature vector constructing to be used for text-independent speaker identification.

Modified DWT (MDWT) is proposed in this text for comparison with the proposed method, which is achieved by applying the same Mallat operation to the high frequency sub signal (d_1) as well as the low frequency. This assists greatly in expanding the utility of DWT via a high pass band of frequency.

2.3. Average framing LPC feature extraction method

Before the stage of features extraction, the speech data are processed by a silence removing algorithm followed by the application of a pre-processing, which is achieved by applying the normalization on speech signals to make the signals comparable regardless of differences in magnitude, because the distribution of these magnitudes is closely related to the volume of the speakers. The signals are normalized by using the following formula [23]:

$$S_{Ni} = \frac{S_i - \dddot{S}}{\sigma}, \tag{9}$$

where S_i is the ith element of the signal S, $\dddot{S}$ and σ are the mean and standard deviation of the vector S, respectively, and S_{Ni} is the ith element of the signal series S_N after normalization.

LPC is not a new technique. It was developed in the 1960s [53] but is admired and widely used to this day because the LPC coefficients representing a speaker by modeling vocal tract parameters and the data size are very suitable for speech compression throughout the digital channel [23]. In the proposed study, the focus will be on modifying LPC coefficients for reducing the size of feature vectors. In our work, we propose the AFLPC to extract features from Z frames of each WT speech sub signal:

$$\{u_q(t)\} = \{u_{q1}(t), u_{q2}(t), \ldots, u_{qZ}(t)\}, \tag{10}$$

where Z is the number of considered frames (each frame of 20 ms duration) for the qth WT sub signal $u_q(t)$. The average of LPC coefficients calculated for Z frames of $u_q(t)$ is utilized to extract a wavelet sub signal feature vector as follows:

$$aflpc_q = \sum_{z=1}^{Z} LPC(u_{qz}(t))\frac{1}{Z}. \tag{11}$$

The feature vector of the whole given speech signal is

$$\text{AFLPC} = \{aflpc_1, aflpc_2, \ldots, aflpc_Q\}. \tag{12}$$

The superiority of the proposed feature extraction method over a conventional one is shown in Fig. 2, where Fig. 2a illustrates two feature vectors taken for a single speaker using LPC from WP at level two. It can be seen that the LPC coefficients have similar shape but are dispersedly distributed. Fig. 2b illustrates two feature vectors taken for the same speaker using AFLPC from WP at level two. This Figure shows these coefficients distributed very well after using AFLPC.

2.4. Genetic algorithm for LPC orders number optimization

The fundamental purpose of genetic algorithms (GAs) is optimization. Given that optimization problems arise frequently, this makes GAs quite useful for a great variety of tasks. As in all optimization problems, we are faced with the problem of maximizing/minimizing an objective function (fitness function) over a given space of arbitrary dimension. Genetic algorithm (GA) is a searching method based on the laws of natural selection and genetics [54,61]. It emulates the individuals in the natural environment, staging that the natural selection mechanism makes the stronger individuals probable winners in the competing environment. Earlier, [55] had applied the GA training for HMM-based speech recognition and gave a better quality of results than the traditional Baum–Welch algorithm [42]. Kwong et al. [43] jointed the GA with the Baum–Welch algorithm to form a hybrid-GA, such that the quality of the results and the runtime performance of the GA were further improved. As a result, GA can be used to specify the number of LPC orders of each speaker feature. The GA works on a population using a set of operators that are applied to the population. A population is a set of points in the design space. The initial population is generated randomly by default. The next generation of the population is computed using the fitness of the individuals in the current generation.

To use the GA to find out the best number of LPC orders for speaker identification, we need to provide at least two input arguments, a fitness function and the number of variables in the problem. The fitness function is the function that can evaluate the feature vector for each number of the LPC orders. For this purpose, the percentage root mean square difference similarity score (PRMS) is proposed:

$$\mathrm{PRDS} = 100 - \left[100 * \sqrt{\left(\sum (V1 - V2)^2 / \sum V1^2\right)}\right], \tag{13}$$

where $V1$ is the LPC feature vector obtained for the first utterances of a speaker and $V2$ is the LPC feature vector obtained for the second utterances of the same speaker of a given number of orders.

The first two output arguments returned by GA are the best point (number of LPC orders) found, and the function (PRDS) value at the best point. A third output argument tells you the reason why GA stopped. It is an integer number which determines whether the optimization has terminated successfully. A value of this integer more than zero indicates success, and less or equal to zero indicates failure. The proposed GA parameters are shown in Table 1.

3. Classification

Next to the introduction of the probabilistic neural network by Specht [33], several enhancements, extensions, and generalizations of the original copy have been proposed. These efforts aim at improving either the learning capability [44], or the classification accuracy of PNNs; or on the other hand, at optimizing network size, thus reducing memory requirements and the resulting complexity of the model, as well as achieving lower operational times [34].

In this work, we use the PNN for the speaker feature vectors classification. The essential motivation of such choice is the possibility of working in an unsupervised training mode that makes the system work online, which is easier for implementation as well as giving PNN the ability to provide the confidence in its decision that follows directly from the Bayes' theorem [57,35].

Although this process does not affect the system performance, in addition, it will offer a speedy process as well as performing in real timely manner.

In Fig. 3, the basic configuration of a PNN for classification in K classes is demonstrated. As seen in the figure, the first layer of the PNN, denoted as an input layer, accepts the input vectors to be classified. The nodes of the second layer, which is chosen as a pattern layer, are grouped in K groups according to the class k_i to which they belong. These nodes, also referred to as pattern units or kernels, are connected to all the inputs of the first layer. Several probability density function estimators are possible [35]. Here we suppose that every pattern unit can be defined as having an activation function, the Gaussian basis function:

Table 1
Parameters used for the GA.

Functions	Description
Generation	100
Population size	50
Fitness function	PRDS
Crossover	Scattered
Mutation	Gaussian
Crossover fraction	0.8
Number of variables	One
Optimized variable	Number of LPC orders
Number of runs	100

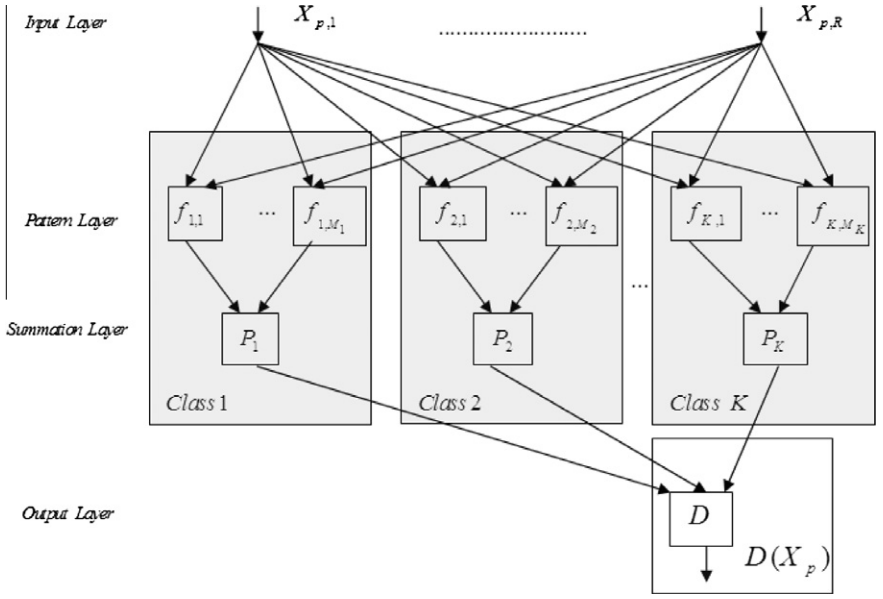

Fig. 3. Structure of the original probabilistic neural network.

$$f_{ij}(x; c_{ij}, \sigma) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{1}{2\sigma^2}(x - c_{ij})^T (x - c_{ij})\right), \tag{14}$$

where $i = 1, \ldots, K$, $j = 1, \ldots, M_i$, and M_i is the number of pattern units in a given class k_i. σ is the standard deviation, also known as the spread or smoothing factor. It regulates the receptive field of the kernel. The input vector x and the centers $c_{ij} \in R^d$ of the kernel are of the dimensionality d.

Finally, exp stands for the exponential function, and the superscript T indicates the transpose of the vector [35].

Clearly, the total number of the second-layer nodes is given as a sum of the pattern units for all classes:

$$M = \sum_{i=1}^{K} M_i. \tag{15}$$

Next, the weighted outputs of the pattern units from the second layer that belong to the group k_i are connected to the third layer which has been chosen as a summation layer corresponding to that specific class k_i. The weights are determined by the decision cost process and the a priori class distribution. In general, the positive weight coefficients ω_{ij} for weighing the member functions of class k_i have to satisfy the requirement

$$\sum_{j=1}^{M_i} \omega_{ij} = 1 \text{ for every given class } k_i, \quad i = 1, \ldots, K. \tag{16}$$

3.1. Proposed probabilistic neural networks algorithm

Ganchev et al. [36] proposed PNN with Mel-frequency cepstral coefficients for text-independence. Although numerous enhanced versions of the original PNN exist, which are either more economical or exhibit an appreciably better performance, for simplicity of exposition, we adopt the original PNN for classification task. The proposed algorithm is denoted by PNN and depends on the following construction:

$$\text{Net} = \text{PNN}(X, P, \text{SP}),$$

where X is a 180x 24 matrix of 24 input speaker feature vectors (pattern) of 180 average framing LPC coefficients, a method that was denoted above by AFLPC, taken from DWT or WP sub signals for net training:

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{124} \\ x_{21} & x_{22} & \ldots & x_{224} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{1801} & x_{1802} & \ldots & x_{18024} \end{bmatrix}, \tag{17}$$

P is the target class vector

$$P = [1, 2, 3, \ldots, 24], \tag{18}$$

The SP parameter is a spread of radial basis functions. We use an SP value of one because that is a typical distance between the input vectors. If the SP approaches zero, the network will act as the nearest neighbor classifier. As the SP becomes larger, the designed network will take into account several nearby design vectors. We create a two layer network. The first layer has radial basis transfer function (RB) neurons (as shown in Fig. 4):

$$\mathrm{RB}(n) = \exp(-n^2), \tag{19}$$

and calculates its weighted inputs with Euclidean distance (ED);

$$\mathrm{ED} = \sum \sqrt{((x-y)^2)}, \tag{20}$$

and its net input with net product functions, which calculate a layer's net input by combining its weighted inputs and biases. The second layer has competitive transfer function (see Fig. 5) neurons, and calculates its weighted input with a dot product weight function. It's weight function applies weights to an input to get weighted inputs. The proposed net calculates its net input functions (called NETSUM) that calculate a layer's net input by combining its weighted inputs and biases. Only the first layer has biases. PNN sets the first layer weights to X', and the first layer biases are all set to 0.8326/SP, resulting in radial basis functions that cross 0.5 at weighted inputs of ±SP. The second layer weights are set to P.

Now, to test the network on a new feature vector (outsider imposter) for identification, simulation with network results is performed.

4. Results and discussion

To examine the presented text-independent speaker identification system, a testing database was created from the Arabic language. The recording environment is a normal office setting via PC-sound card, with original frequency of 4 kHz and a sampling frequency of 16 kHz. These utterances are Arabic spoken digits from 0 to 14. Each speaker also distinctly reads 30 s worth of different Arabic texts ten separate times. A total of 47 individual speakers (19–40 years old), of whom are 31 individual males and 16 individual females, spoke these Arabic words and texts for training and testing modes. The total number of tokens considered for training and testing was 1128.

Some experiments were performed using all of the 1128 Arabic utterances from these 47 individual speakers. For each of these speakers, 24 speech signals were used, of which 6 were used for the training mode and 18 for testing. The proposed system was tested by utilizing all of these speakers.

In the first experiment, GA is applied to reveal the correlation between the LPC order and the recognition rate. We examined the LPC orders with an upper limit of 50 in order to determine the feature vector of lower dimensionality. Based on the results of GA, four LPC orders were determined: 16, 22, 30 and 50 in term of the fitness function PRDS (presents identification accuracy). Table 2 gives the results of the recognition rate by means of the proposed method for the four optimized LPC orders. In all cases it was found that the recognition rate was proportional to LPC orders except LPC coefficients (50). With more coefficients, the higher recognition rate was acquired, and, the increase of LPC coefficients did not tremendously burden the system load. However, the use of these parameters still has its limitation since the number of parameters slightly affects the recognition rate. When the recognition rate reached over 90%, it did not produce any improvement in the performance even though double the amount of LPC coefficients (50) were used. Moreover, for PNN, the increase in parameters also affects the training time.

Based on stated results, an LPC order of 30 for each frame will be used. It was determined based on the GA and empirically as a tradeoff between the recognition rate and the feature vector length.

The complete analysis flowchart is shown in Fig. 6, pointing out that the speech signals are processed by a silence removing algorithm. This process is followed by the application of a pre-process by applying the normalization on speech signals. This stage makes the signals comparable regardless of differences in magnitude before extracting the feature vector. The

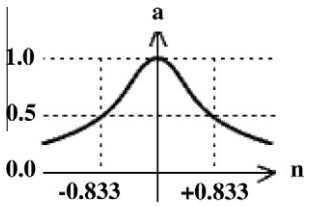

Fig. 4. Radial basis transfer function.

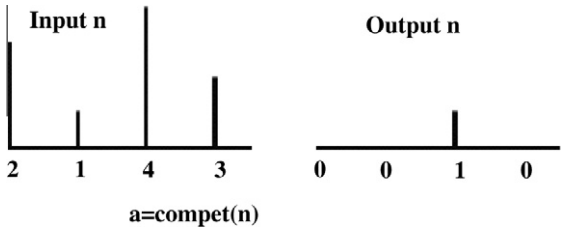

Fig. 5. Competitive transfer function.

Table 2
GA results.

No. of speakers	Optimized LPC order	16	22	30	50
47	Recognition rate	86.61	86.86	97.36	94.32

performance of the AFLPC method was evaluated by PNN classifier, which is not only rapid in the training procedure, but also has the potential for real-time applications.

In the experiments, several feature extraction methods were analyzed to expose the efficacy of the proposed system. The following experiment investigates the proposed method in terms of the recognition rate. This can be concluded after interpretation of the results in Table 3, where the results of DWT with conventional LPC (DWTLPC), DWT with AFLPC (DWTLPCF), WP with conventional LPC (WPLPC) and WP with AFLPC (WPLPCF) are tabulated. DWT was processed at level 5 with 6 sub signals while WP was processed at level two with 6 sub signals. It was found that the recognition rates of WP methods are superior (96.54 and 97.36) when compared with DWT methods (93.26 and 94.36). On the other hand, methods of AFLPC are superior (94.36 and 97.36) when compared with the conventional methods (93.26 and 96.54). The same conclusion was derived by means of the correlation coefficient method being taken for 150 different signals of 15 speakers instead of the PNN classifier (Fig. 7).

A comparative study of the proposed feature extraction method with other feature extraction methods was performed. The Wavelet packet energy index distribution method (WPID) [23], genetic wavelet packet neural network (GWPNN) [45], Modified DWT with conventional LPC (MDWTLPC), Eigen vector with conventional LPC [46] in conjunction with WP (EWPLPC) or with DWT (EDWTLPC), Shannon [28], sure [25], MFCC with Gaussian mixture model (GMM) (MFGMM) [30,36,62] and log energy [45] entropies methods taken for WP are employed for comparison. The results are presented in Table 5. To choose the optimal WP level used for entropies and energy index methods to be used in comparison, investigation results are presented in Table 4. For all these methods, PNN classifier is utilized. The best recognition rate selection obtained was 97.36 for WPLPCF (Table 5).

GMM is extensively used for classification tasks, especially in speaker identification [46]. Because GMM can smoothly estimate the density distribution of the data clusters, its accuracy and performance are outstanding.

To further test the proposed method, the probabilistic neural network was replaced with a GMM [58]. Then AFLPC coefficients were extracted from LPC and used as a feature vector. The classification stage includes PNN and GMM. Experimental results showed that both PNN and GMM are rapid in the single training procedure. However, PNN spent slightly more time than GMM. The experimental results revealed that the proposed AFLPC technique with GMM can accomplish the speaker

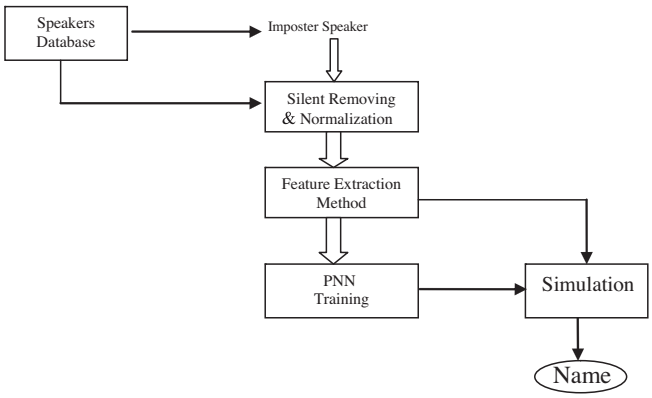

Fig. 6. Flowchart of the proposed system.

Table 3
Recognition rates of wavelet transform based feature extraction methods.

Speaker	No. of signals	Recognition rate (%)			
		DWTLPC	DWTLPCF	WPLPC	WPLPCF
Sp.1	24	100	100	100	100
Sp.2	24	88.88	91.58	87.50	92.50
Sp.3	24	100	100	100	100
Sp.4	24	88.88	91.51	100	100
Sp.5	24	100	100	100	100
Sp.6	24	66.66	68,91	87.50	91
Sp.7	24	100	100	100	100
Sp.8	24	100	100	100	100
Sp.9	24	100	100	100	100
Sp.10	24	100	100	100	100
Sp.11	24	88.88	91.45	100	100
Sp.12	24	66.66	86.91	87.50	91.50
Sp.13	24	100	100	100	100
Sp.14	24	100	100	100	100
Sp.15	24	100	100	100	100
Sp.16	24	100	100	100	100
Sp.17	24	87.5	88.75	87.50	91.50
Sp.18	24	100	100	100	100
Sp.19	24	87.5	89.75	100	100
Sp.20	24	100	100	100	100
Sp.21	24	100	100	100	100
Sp.22	24	100	100	100	100
Sp.23	24	100	100	100	100
Sp.24	24	100	100	100	100
Sp.25	24	100	100	100	100
Sp.26	24	100	100	100	100
Sp.27	24	100	100	87.50	90.20
Sp.28	24	62.5	65.75	100	100
Sp.29	24	87.5	88.75	100	100
Sp.30	24	100	100	100	100
Sp.31	24	100	100	100	100
Sp.32	24	100	100	100	100
Sp.33	24	100	100	100	100
Sp.34	24	87.5	89.75	75	78.7
Sp.35	24	87.5	89.50	87.50	89.50
Sp.36	24	100	100	100	100
Sp.37	24	100	100	100	100
Sp.38	24	100	100	75	79
Sp.39	24	100	100	87.50	90.50
Sp.40	24	100	100	100	100
Sp.41	24	87.5	89.25	87.50	90.50
Sp.42	24	87.5	89.75	100	100
Sp.43	24	100	100	87.50	90.90
Sp.44	24	87.5	89.35	100	100
Sp.45	24	100	100	100	100
Sp.46	24	75	78.25	100	100
Sp.47	24	62.5	63.75	100	100
Total	1128	93.26	94.36	96.54	97.36

identification in a short period of time and achieve a satisfactory recognition rate. Table 6 shows the experimental results of the two classification approaches within AFLPC. The recognition rates of GMM for the whole database reached the highest values with an average of 96.88% due to the big number of training speakers. The small number of training speakers for PNN caused its best average recognition rate to be 97.07%.

Another experiment was conducted to assess the performance of the system in noisy environments. Table 7 summarizes the results of the speaker identification corresponding to white Gaussian noise and real noise (restaurant noise, which seems like babbling) with the signal-to-noise ratio (SNR) of 0 and 5 dB references. SNR was calculated as follows:

$$\mathrm{SNR} = 20\log 10\frac{\sum s}{\sum(s - s_n)},$$

where s is free of a noise speech signal and s_n is a noisy speech signal. Three approaches used in the experimental investigation for comparison: WPLPCF, DWTLPCF and Eigen vector with AFLPC (EWPLPCF). The recognition rate of WPLPCF reached the lowest value. The best recognition rate selections obtained were 58.56 (with 0 dB) and 70.52 (with 5 dB) for DWTLPCF. The reason for DWT's success over WP is that the feature vector is obtained from level 5 (depth 5), where the sub signals are

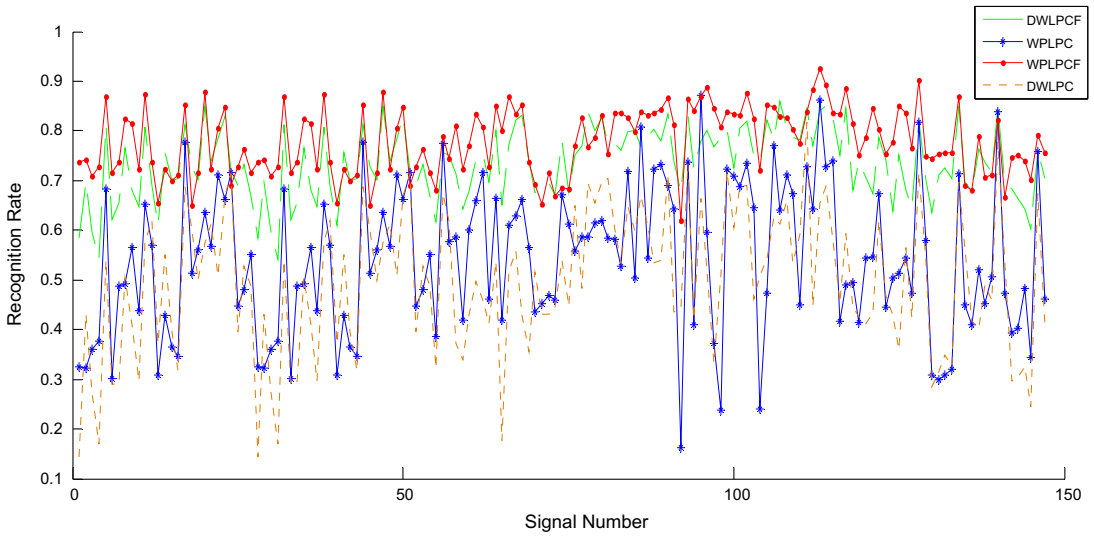

Fig. 7. Correlation coefficient method taken for 150 different signals of 15 speakers.

Table 4
Optimal WP level used for entropies methods.

Identification method	Recognition rate (%)			
	Level 3	Level 5	Level 7	Level 9
Shannon and WP	69.44	82.64	86.11	85.42
Sure and WP	40.74	41.67	42.59	44.44
Log energy and WP	66.67	67.59	67.59	69.44
Energy index and WP	71.29	88.88	89.81	83.33

Table 5
Comparison between different feature extraction methods.

Identification method	No. of signals	Recognition rate (%)
WPID	1128	92.81
GWPNN	1128	85.64
MDWTLPC	1128	92.39
EWPLPC	1128	94.19
EDWTLPC	1128	91.84
Shannon and WP	1128	89.11
Sure and WP	1128	47.44
MFGMM	1128	90.42
Log energy and WP	1128	72.44
WPLPCF	1128	97.36

Table 6
Comparison between GMM and PNN with AFLPC.

Identification method	Classifier	Number of training speakers	Number of training patterns for each speaker	Recognition rate (%)
AFLPC	GMM	10	6	65.57
AFLPC	GMM	47	15	96.88
AFLPC	PNN	10	6	97.07
AFLPC	PNN	47	15	69.91

filtered in lower depth than in WPLPCF at level 2. It is shown that the use of the Eigen vector in conjunction with WPLPCF can improve the robustness of an identification system.

Table 7
Comparison between DWT and WP with AWGN.

Identification method	Recognition rate (%) AWGN		Recognition rate (%) Restaurant noise	
	0 dB	5 dB	0 dB	5 dB
DWTLPCF	58.56	70.52	45.21	65.89
WPLPCF	26.15	37.26	22.01	30.42
EWPLPCF	58.23	66.37	43.33	62.45

5. Conclusion

This work proposed a speaker identification system based AFLPC. The benefit of AFLPC is its capability to reduce the huge speech data into a few values, and the computing speed is also accomplished. In the beginning of feature extraction, WT is applied with LPC coefficients by analyzing the vocal tract parameters of a speaker. Then AFLPC coefficients are extracted from LPC obtained from wavelet coefficients and used as a representative speaker feature vector. For classification, PNN is applied. The speaker identification performance of this method was demonstrated on a total of 47 individual speakers (31 male speakers and 16 female speakers). Experimental results showed both DWT and WP linked with AFLPC are suitable for the feature extraction method. However, WP resulted in better performance in terms of recognition rate (97.36%). As a comparison with other published methods, WPLPCF produced a higher recognition rate. The experimental results revealed the proposed AFLPC technique with DWT can accomplish better results for a speaker identification system in an AWGN environment; 58.56% for 0 dB and 70.52% for 5 dB. The experimental results of the two classification approaches GMM and PNN within AFLPC were obtained. The recognition rates of GMM for the whole database reached the highest values with average 96.88% in case of big number of training speakers, while the best average recognition rate selection obtained was 97.07% for PNN in case of small number of training speakers.

References

[1] Antonini M, Barlaud M, Mathieu P, Daubechies I. Image coding using wavelet transform. IEEE Trans Image Proc 1992;1(2):205–20.
[2] Quiroga RQ. Quantitative analysis of EEG signals: time–frequency methods and chaos theory. Lü beck: Institute of Physiology, Medical University; 1998.
[3] Lung S-Y, Chen C-C. Further reduced form of Karhunen–Loeve transform for text independent speaker recognition. Electron Lett 1998;34:1380–2.
[4] Mallat S. A wavelet tour of signal processing. San Diego, CA: Academic Press; 1998.
[5] Hosseinzadeh D, Krishnan S. Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs. In: Proceedings of the international conference on signal processing; 2007. p. 365–8.
[6] Afify M, Siohan O. Comments on vocal tract length normalization equals linear ransformation in cepstral space. IEEE Trans Audio Speech Lang Process 2007;15:1731–2.
[7] Rioul OM, Vetterli V. Wavelets and signal processing. IEEE Trans Signal Process Mag 1991:14–38.
[8] Long CJ, Datta S. Wavelet based feature extraction for phoneme recognition. Proceeding of the fourth international conference of spoken language processing, Philadelphia, USA, vol. 1. p. 264–7.
[9] Lung S-Y. Improved wavelet feature extraction using kernel analysis for text independent speaker recognition. Digital Signal Process 2010;20:1400–7.
[10] Chen C-C, Chen C-Ta, Tsai C-M. Hard-limited Karhunen–Loeve transform for text independent speaker recognition. Electron Lett 1997;33:2014–5.
[11] Vetterli M, Kovacevic J. Wavelets and sub band coding. Englewood Cliffs, NJ: Prentice Hall; 1995.
[12] Mallat S, Zhong S. Characterization of signals from multiscale edges. IEEE Trans Pattern Anal Mach Intell 1992;14:710–32.
[13] Mallat S. Zero-crossings of a wavelet transform. IEEE Trans Inform Theory 1991;37:1019–33.
[14] Mallat S. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 1989;11:674–93.
[15] Specht DF. A general regression neural network. IEEE Trans Neural Netw 1991;19(4):1560–8.
[16] Tufekci Z, Gowdy J. Feature extraction using discrete wavelet transform for speech recognition. In: Proceedings of the SOUTHEASTCON. p. 116–23.
[17] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans ASSP 1980;28:357–66.
[18] Mitchell RA. Hybrid statistical recognition algorithm for aircraft identification. Dayton, OH: University of Dayton Press; 1997.
[19] Lung S-Y. Feature extracted from wavelet eigenfunction estimation for text-independent speaker recognition. Pattern Recognit 2004;37:1543–4.
[20] Chen C-T, Lung S-Y, Yang C-F, Lee M-C. Speaker recognition based on 80/20 genetic algorithm. In: IASTED international conference on signal processing, pattern recognition, and application, Greece. p. 547–9.
[21] Nathan KS, Silverman HF. Time-varying feature selection and classification of unvoiced stop consonants. IEEE Trans Speech Audio Process 1994;2:395–405.
[22] Sarikaya R, Pellom B, Hansen J. Wavelet packet transform features with application to speaker identification. NORSIG; 1998.
[23] Wu J-D, Lin B-F. Speaker identification based on the frame linear predictive coding. Expert Syst Appl 2009;36:8056–63.
[24] Wu J-D, Lin B-F. Speaker identification using discrete wavelet packet transform technique with irregular decomposition. Expert Syst Appl 2009;36:3136–43.
[25] Avci D. An expert system for speaker identification using adaptive wavelet sure entropy. Expert Syst Appl 2009;36:6295–300.
[26] Avci E. A new optimum feature extraction and classification method for speaker recognition: GWPNN. Expert Syst Appl 2007;32:485–98.
[27] Avci E, Hanbay D, Varol A. An expert discrete wavelet adaptive network based fuzzy inference system for digital modulation recognition. Expert Syst Appl 2006;33:582–9.
[28] Daqrouq K. Wavelet entropy and neural network for text-independent speaker identification. Eng Appl Artif Intell 2011;24:796–802.
[29] Daqrouq K, Abu Sbeih I, Daoud O, Khalaf E. An investigation of speech enhancement using wavelet filtering method. Int J Speech Technol 2010;13(2):101–15.
[30] Xiang B, Berger T. Efficient text-independent speaker verification with structural Gaussian mixture modes and neural network. IEEE Trans Speech Audio Process 2003;11:447–56.
[31] Gowdy J, Tufekci Z. Mel-scaled discrete wavelet coefficients for speech recognition. Proceedings of the ICASSP, vol. 1. p. 1351–4.

[32] Xia YY, Xie YM, Zhu RG. An engineering geology evaluation method based on an artificial neural network and its application. Eng Geol 1997;47:149–56.
[33] Specht DF. Probabilistic neural networks. Neural Netw 1990;3(1):109–18.
[34] Specht DF. Enhancements to probabilistic neural networks. In: Proceedings of the IEEE international joint conference on neural networks, Baltimore, MD, June 7–11.
[35] Ganchev T, Tasoulis D, Vrahatis M, Fakotakis D. Generalized locally recurrent probabilistic neural networks with application to text-independent speaker verification. Neurocomputing 2007;70:1424–38.
[36] Ganchev T, Fakotakis N, Kokkinakis G. Comparative evaluation of various MFCC implementations on the peaker verification task. Proceedings of the SPECOM-2005, vol. 1. p. 191–4.
[37] Adami AG, Barone DAC. A speaker identification system using a model of artificial neural networks for an elevator application. Inform Sci 2001;138:1–5.
[38] Haydar A, Demirekler M, Yurtseven MK. Speaker identification through use of features selected using genetic algorithm. Electron Lett 1998;34:39–40.
[39] Huang X, Acero A, Hon H-W. Spoken language processing: a guide to theory, algorithm, and system development. New Jersey: Prentice-Hall; 2001.
[40] Huang Y, Wueng S, Ou C, Cheng C, Su K. Nutritional status of functionally dependent and nonfunctionally dependent elderly in Taiwan. J Am College Nutr 2001;20:135–42.
[41] Hermansky H. Perceptual linear prediction (PLP) analysis for speech. J Acoust Soc Amer 1990;87:1738–52.
[42] Rabiner L, Juang B-H. Fundamentals of speech recognition. Englewood Cliffs, New Jersey: Prentice-Hall; 1993.
[43] Kwong S, Chau CW, Man KF, Tang KS. Optimisation of HMM topology and its model parameters by genetic algorithms. Pattern Recognit 2001;34:509–22.
[44] Yang ZR, Chen S. Robust maximum likelihood training of heteroscedastic probabilistic neural networks. Neural Netw 1998;11(4):739–48.
[45] Engin A. A new optimum feature extraction and classification method for speaker recognition: GWPNN. Expert Syst Appl 2007;32:485–98.
[46] Uchida S, Ronee MA, Sakoe H. Using eigen-deformations in handwritten character recognition. Proceedings of the 16th ICPR, vol. 1. p. 572–5.
[47] Daubechies I. Orthonormal bases of compactly supported wavelets. Commun Pure Appl Math 1988;41:909–96.
[48] Lei Z, Jiandong L, Jing L, Guanghui Z. A novel wavelet packet division multiplexing based on maximum likelihood algorithm and optimum pilot symbol assisted modulation for Rayleigh fading channels. Circ Syst Signal Process 2005;24(3):287–302.
[49] Behroozmand R, Almasganj F. Optimal selection of waveletpacket-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis. Comput Biol Med 2007;37:474–85.
[50] Sarikaya R, Hansen JHL. High resolution speech feature parametrization for monophone-based stressed speech recognition. IEEE Signal Process Lett 2000;7(7):182–5.
[51] Souani C, Mohamed Abid, Kholdoun Torki, Rached Tourki. VLSI design of 1-D DWT architecture with parallel filters. Integration 2000;29(2):181–207.
[52] Delac K, Grgic M, Grgic S. Face recognition in JPEG and JPEG2000 compressed domain. Image Vis Comput 2009;27:1108–20.
[53] Atal BS. The history of linear prediction. IEEE Signal Process Mag 2006;23:154–61.
[54] Tang KS, Man KF, Kwong S, He Q. Genetic algorithm and their applications. IEEE Signal Process Mag 1996;13(6):22–37.
[55] Chau CW, Kwong S, Diu CK, Fahrner WR. Optimisation of HMM by a genetic algorithm. In: Proceedings of the ICASSP 3. p. 1727–30.
[56] Bennani Y, Gallinari P. Neural networks for discrimination and modelization of speakers. Speech Commun 1995;17:159–75.
[57] Bayes T. An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond 1763;53:370–418.
[58] Kinnunen T, Haizhou L. An overview of text-independent speaker recognition: from features to super vectors. Speech Commun 2010;52:12–40.
[59] Ghaffari A, Golbayani H, Ghasemi M. A new mathematical based QRS detector using continuous wavelet transform. Comput Electr Eng 2008;34(2):81–91.
[60] Lu W, Sun W, Lu H. Robust watermarking based on DWT and nonnegative matrix factorization. Comput Electr Eng 2009;35(1):183–8.
[61] Banković Z, Stepanović D, Bojanić S, Nieto-Taladriz O. Improving network security using genetic algorithm approach. Comput Electr Eng 2007;33(5–6):438–51.
[62] Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans Speech Audio Process 1995;3(1):72–83.

Khaled Daqrouq received the B.S. and M.S. degrees in biomedical engineering from the Wroclaw University of Technology in Poland, in 1995, as one certificate, and the Ph.D. degree in electronics engineering from the Wroclaw University of Technology, in Poland, in 2001. He is currently associate professor at the King Abdulaziz University, in SA. His research interests are ECG signal processing, wavelet transform applications in speech recognition and the general area of speech and audio signal processing and improving auditory prostheses in noisy environments.

Khalooq Y. Al Azzawi received his B.Sc. in Electrical Engineering from University of Mosul in Iraq in 1970, a Post Graduate Diploma in Communications Engineering from Manchester University (UMIST) in England in 1976, and M.Sc. degree in Digital Communications & Electronics from Loughborough University of Technology in England in 1977. He is currently associate professor in Electromechanical Department in Baghdad University of Technology. His research interests are FDNR Networks in filters, and Signal processing, specially in wavelet transform applications in speech recognition.