Applied Soft Computing 27 (2015) 231–239

Contents lists available at ScienceDirect

Applied Soft Computing

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

Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers

Khaled Daqrouq [a,*], Tarek A. Tutunji [b]

[a] Electrical & Computer Engineering Department, King Abdulaziz University, Jeddah, Saudi Arabia
[b] Mechantronics Engineering Department, Philadelphia University, Jordan

ARTICLE INFO

Article history:
Received 15 December 2011
Received in revised form 3 November 2014
Accepted 18 November 2014
Available online 26 November 2014

Keywords:
Speaker verification and identification
Wavelet packet
Neural networks
Formants

ABSTRACT

This paper proposes a new method for speaker feature extraction based on Formants, Wavelet Entropy and Neural Networks denoted as FWENN. In the first stage, five formants and seven Shannon entropy wavelet packet are extracted from the speakers' signals as the speaker feature vector. In the second stage, these 12 feature extraction coefficients are used as inputs to feed-forward neural networks. Probabilistic neural network is also proposed for comparison. In contrast to conventional speaker recognition methods that extract features from sentences (or words), the proposed method extracts the features from vowels. Advantages of using vowels include the ability to recognize speakers when only partially-recorded words are available. This may be useful for deaf-mute persons or when the recordings are damaged. Experimental results show that the proposed method succeeds in the speaker verification and identification tasks with high classification rate. This is accomplished with minimum amount of information, using only 12 coefficient features (i.e. vector length) and only one vowel signal, which is the major contribution of this work. The results are further compared to well-known classical algorithms for speaker recognition and are found to be superior.

1. Introduction

Speech processing applications include speech recognition and speaker identification. Speaker identification system is a technology with a potentially large market due to broad applications that range from automation using operator-assisted service to speech-to-text aiding system [1,2].

In general, a speaker identification system can be implemented by observing the voiced/unvoiced components or by analyzing the speech energy distribution. Such systems can be divided into two main steps: feature extraction and speaker classification [3]. Several digital signal processing methods have been used by researchers: Linear Predictive Coding (LPC) technique [1], Mel Frequency Cepstral Coefficients (MFCC) [4], Discrete Wavelet Transform (DWT) [5] and Wavelet Packet Transform (WPT) [3].

Due to its success in analyzing non-stationary signals, DWT has become a powerful alternative to the Fourier methods in many speech/speaker identification applications. The main advantage of wavelets is its optimal time–frequency resolution in all frequency ranges. This is a result of varying the window size for different frequencies: wide for slow frequencies and narrow for fast frequencies [1,6,7].

Previous studies showed that using Wavelet Packet (WP) entropy as features in recognition tasks is effective. In [27], a method to calculate the wavelet norm entropy value in digital modulation recognition was proposed. In [36] a combination of genetic algorithm and WPT was presented and the energy features were determined from a group of WP coefficients. The work was applied in bio-medic where the results were used for pathological classification and evaluation. Energy indexes of WP were proposed for speaker identification [3] and sure entropy was considered terminal node signal waveforms obtained from DWT [1] and applied to speaker identification. Others, [28], used features

Abbreviations: ACF, Autocorrelation Function; ANN, Artificial Neural Network; AR, Auto-Regressive; DFT, Discrete Fourier Transform; DWT, Discrete Wavelet Transform; DWT-NN, Discrete Wavelet Transform with Neural Networks; FSVM, Feature-extraction Support Vector Machine; FWENN, Formants Wavelet Entropy with Neural Networks; GWP-NN, Genetic Wavelet Packet with Neural Networks; IMFCC, Inverted Mel Frequency Cepstral Coefficient; LPC, Linear Predictive Coding; LPC-NN, Linear Predictive Coding with Neural Networks; LPC-WP, Linear Predictive Coding with Wavelet Packet; MFCC, Mel Frequency Cepstral Coefficient; MFCC-NN, Mel Frequency Cepstral Coefficient with Neural Network; PSD, Power Spectrum Density; SWPF, Sure entropy with WP Formants; SVM, Support Vector Machine; WP, Wavelet Packet; WPID, Wavelet Packet Index Distribution; WPT, Wavelet Packet Transform.

* Corresponding author at: P.O. Box 80204, Jeddah 21589, Saudi Arabia. Tel.: +966 5 66 980400; fax: +966 5 695 2686.
E-mail address: haleddaq@yahoo.com (K. Daqrouq).

http://dx.doi.org/10.1016/j.asoc.2014.11.016
1568-4946/

extraction methods based on a combination of three entropy types (sure, logarithmic energy and norm).

Formant can be described as a function of the supralaryngeal vocal tract. The air in the oral and nasal cavities vibrates at a range of frequencies as a response to the vibratory movement of the vocal folds and air passing through the glottis. These resonant frequencies are affected by the size/shape of the vocal tract and by the tongue and lip positions [9]. Vocal tract resonances are often studied in terms of vowel formant frequencies. Because the male vocal tract is about 15% longer than the female vocal tract, men's speech signals have lower formant frequencies than women's [10]. During voiced speech, resonant frequencies of the vocal track are recognized as formants with valuable features for both automatic speech recognition and speech synthesis [11].

The use of formants for 1-D and 2-D continuous motion control created a new vocal interface that allowed people, especially individuals with motor impairments, to interact with computer-based devices [8,49].

Researchers used different methods for formant tracking that included: Linear Predictive Coding spectral analysis [12], hidden Markov model based methods [13,60], nonlinear predictors [14], and Kalman filtering framework [15].

Artificial Neural Network (ANN) models have been effectively used for speaker classification [38,53,57,59]. Researchers used radial basis function networks [17,18,54] and developed ANN-based techniques using a cascade neural network [16] for speaker verification. Others compared ANN with second order statistical techniques for speaker verification [19]. Committee neural networks to improve the reliability of ANN based classification systems were developed [21,22] and the use of these committee networks for text-dependent speaker verification was addressed [20]. Support Vector Machines (SVM), a special case of Tikhonov regularization that belongs to the general linear classifier family, have been used for speaker recognition [46–47].

Researchers used a combination of MFCC and parametric feature-sets' algorithms to improve the accuracy of speaker recognition systems in adverse environments. Some studies, such as [39], emphasized their work on text-dependent speaker identification, which deals with feature extraction by means of LPC coefficients. A Gaussian shaped filter was used for calculating MFCC and IMFCC instead of typical triangular shaped bins. A system using four transform techniques was suggested in [40,41]. The feature vectors were the row mean of the transforms for different groupings. Experiments were performed on Discrete Fourier Transform (DFT), Discrete Cosine Transform, Discrete Sine Transform and Walsh Transform. All these methods showed an accuracy of more than 80% for the different groupings considered. However, the results showed that Discrete Sine Transform had the best performance. In [42] IMFCC, which covers high frequency, was used to improve speaker recognition rate.

Researchers investigated fundamental and formant frequencies for a speaker recognition task [50]. It was concluded from a detailed comparison that the long-term formant distributions contributed to the rejection of the suspect. Grigoras continued this study to calculate likelihood ratios based on the density estimation of formant frequencies on distinct vowel phonemes ([a], [e], [i], [o]) [51]. Rose [52] suggested the comparison of vowel phonemes by likelihood ratio computation, and recognition of human speech phonemes by fuzzy method was proposed in [58].

In our study, vowels are used for speaker recognition. And because the formants are recommended in case of vowels [51,52], they are studied here in detail. In order to enhance the recognition results, WP entropy is utilized. The reason behind WP entropy is to extract additional features over different band passes of frequency by Shannon entropy.

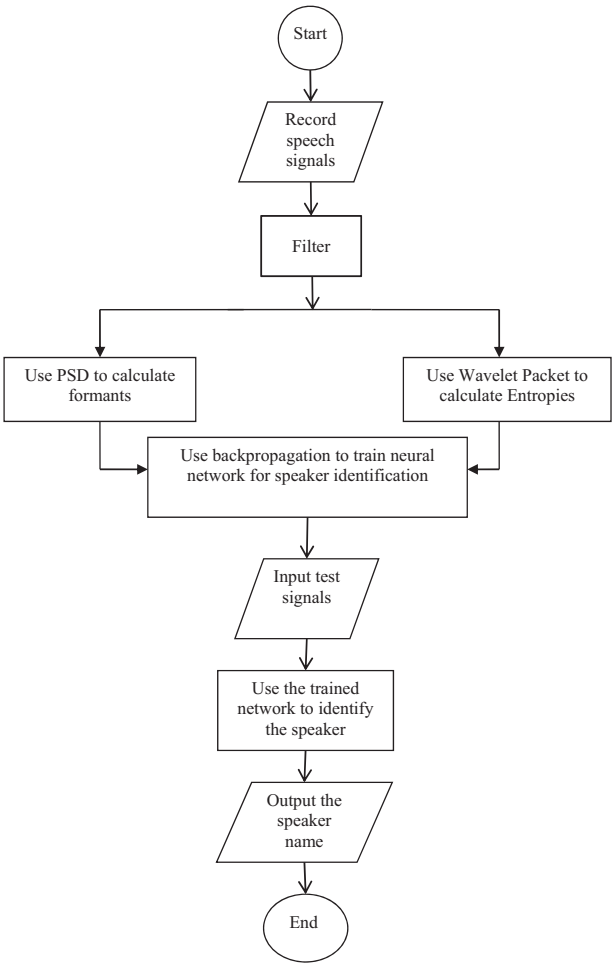

Fig. 1. Flow chart procedure for FWENN method.

This paper presents a new method for speaker identification that uses formants and wavelet packet entropy within a feed-forward neural network. The objective was to develop the method using partially-recorded speech signals only. The major contribution of this research is the development of an accurate speaker identification method that uses simple computations with minimum amount of information. The developed method is capable of dealing with vowels as the only input from the speech signal. This method might be used for forensic and criminal investigation as well as for deaf-mute speakers' recognition.

The paper is organized as follows: Section 1 describes the general structure of proposed method. Section 2 discusses feature's extraction using Power Spectrum Density and Wavelets while Section 3 describes the neural networks used in this work. Section 5 provides the experimental results and Section 6 concludes the paper.

2. FWENN method

The method developed and proposed in this work, Formant and Wave Entropy within Neural Networks (FWENN), is explained in this section. The proposed method is based on several steps (as shown as a flow chart in Fig. 1) and can be divided into four stages: recording and filtering the speech signals, extracting features, classification, and speaker retrieval.

The emphasis here will be on the second and third stages: extracting features and classification [37]. Extracting features of the speech signals were performed using two techniques: formants

using Power Spectrum Density (PSD) and entropies using WP. These concepts will be explained in Section 3. Classification was done using neural networks and will be explained in Section 4.

3. Features' extraction

Periodic excitation is seen in the spectrum of certain sounds, especially vowels. The speech organs form certain shapes to produce the vowel sound and therefore regions of resonance and anti-resonance are formed in the vocal tract. Location of these resonances in the frequency spectrum depends on the form and shape of the vocal tract. Since the physical structure of the speech organs is a characteristic of each speaker, differences among speakers can also be found in the position of their formant frequencies. These resonances affect the overall spectrum shape and are referred to as formants. A few of these formant frequencies can be sampled at an appropriate rate and used for speaker recognition. These features are normally used in combination with other features. Unlike our previous works that investigated DWT for feature extraction [24,25], in this work, the formants are used with the WP entropy to identify the speakers.

This paper proposes to use formants and WP parameters as inputs to ANN for speaker classification. Therefore, it is necessary to introduce the two concepts: feature extraction by formants and WP algorithms. The discussion of these concepts will be limited to their use in speaker identification.

3.1. Features' extraction by formants using PSD

Formants are the spectral peaks of the sound spectrum of vowels or the acoustic resonances of the human vocal tract. In a wide band spectrogram they show up as black bars. In a small band spectrogram the fundamental frequency and the harmonics are visible as well. The sound production can be modeled as a time varying linear system (having these resonances), that is excited by a sequence of impulses. Using a linear system of order + n and taking the inverse, the model spectrum of the linear system can be created from the speech signal. Using a small order can result in noisy resonances while using a large order can introduce artificial harmonics: For small n, the resonances are smeared, while for large n, some of the peaks are actually not formants but harmonics. An order n corresponding to about 1 ms seems to be a good choice resulting in five formants.

The first five vocal resonant frequencies, i.e. formants (F1, F2, F3, F4, F5), during voiced-speech are distinguishable for each person and therefore are proposed as the speaker features. For voiced-speech, the glottis signal is periodic with a fundamental frequency (i.e. pitch, F0). Variations of the pitch during the duration of the utterance provide the contour, which can be used as a feature for speech recognition. The speech utterance is normalized and the contour is determined. The vector that contains the average values of pitch of all segments is thereafter used as a feature for speaker recognition. The pitch might be sufficient for the speaker identification, but is usually assisted by the formants for the best speaker recognition [26].

The filtered speech signal can be used as an input to a power spectrum algorithm in order to identify the first five formants. These formants can be used as the unique features for the speaker. The PSD algorithm can be used to identify the formants in two steps: First, the PSD is estimated using the Yule–Walker Auto-Regressive (AR) method. Then, the local maxima are identified.

Power spectrum can be found by taking the Fourier Transform of the Autocorrelation Function (ACF)

$$P_x(e^{j\omega}) = \sum_{n=-\infty}^{\infty} r_x(n)e^{-j\omega n} \tag{1}$$

where $r_x(n)$ is the ACF of the signal $x(n)$. The autocorrelation values are estimated from finite data record, $x(n)$ for $0 \le n \le N-1$, and is defined as

$$\hat{r}_x(k) = \frac{1}{N}\sum_{n=0}^{N-1-k} x(n+k)x^*(n) \quad k = 0, 1, \ldots, p \tag{2}$$

Eq. (1) is one estimate of the PSD, but has some disadvantages that include excessive variance estimation. A better estimate is the Yule–Walker method.

The Yule–Walker method estimates the PSD of the input using the Yule–Walker AR method. The concept is to minimize the forward prediction error by fitting an AR model to the windowed input data. This method is also called autocorrelation method [30]. The PSD is estimated using the following

$$P_x(e^{j\omega}) = \frac{|b(0)|^2}{\left|1 + \sum_{k=1}^{p} a(k)e^{-j\omega k}\right|^2} \tag{3}$$

The parameters $a(k)$ and $b(0)$ can found from the autocorrelation estimates and will be described next.

The AR model is described in the equation

$$\hat{r}_x(n) = -\sum_{k=1}^{p} a(k)\hat{r}_x(n-k) \tag{4}$$

Which can be expanded into matrix form as

$$\begin{bmatrix} \hat{r}_x(0) & \hat{r}_x(1) & \cdots & \hat{r}_x(p-1) \\ \hat{r}_x(1) & \hat{r}_x(0) & \cdots & \hat{r}_x(p-2) \\ \cdots & \cdots & \cdots & \cdots \\ \hat{r}_x(p-1) & \hat{r}_x(p-2) & \cdots & \hat{r}_x(n0) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ \cdots \\ a(p) \end{bmatrix} = - \begin{bmatrix} \hat{r}_x(1) \\ \hat{r}_x(2) \\ \cdots \\ \hat{r}_x(p) \end{bmatrix} \tag{5}$$

This formulation is defined as the Yule–Walker equations and the Levinson–Durbin recursion are used to solve the equations in order to obtain the AR parameters: $a(1), \ldots, a(p)$. On the other hand, the parameter $b(0)$ is calculated using

$$|\hat{b}(0)| = \hat{r}_x(0) + \sum_{k=1}^{p} a(k)\hat{r}_x(k) \tag{6}$$

The above parameters, $b(0)$ and $a(k)$, can now be substituted to estimate the PSD.

In the context of this paper, Arabic vowels were used for speaker identification. However, the proposed method can be applied to other languages. Figs. 2 and 3 illustrate the applied results for the first five formants (spectrum peaks) of Arabic vowels for speakers' models discrimination. Fig. 2 illustrates the formants of Arabic vowels for two speakers. For each speaker, four speech signals with Arabic vowel ي sounds /e/ were recorded (see Appendix A). Note that the spectrum features for each speaker are similar while there are clear differences in magnitudes and indexes between the two speakers. Fig. 3 illustrates the formants of two Arabic vowels: ا sounds /a/ and Arabic vowel ي (i.e. vowel-independent). Here, the overlap between the two speaker spectrum features is much larger because different vowels are used (i.e. vowel-independent case). This is a limitation due to the use of different vowels for each speaker.

Frequency information, specifically the indexes of local maxima, contains the distinguishable speaker features. Table 1 shows the

Table 1
Comparison among formants (indexes) and magnitudes.

Speaker	Signal number (vowel E)	F1		F2		F3		F4		F5	
		Index	Mag.	Index	Mag.	Index	Mag.	Index	Mag.	Index	Mag.
1	1	7	0.5934	39	4.2136	80	0.2432	127	2.9991	N/A	N/A
	2	7	2.4974	40	4.1294	80	0.134	127	3.0025	N/A	N/A
	3	7	0.3612	40	4.7808	81	0.3719	N/A	N/A	N/A	N/A
	4	7	2.2286	41	2.5217	78	0.1622	127	2.9925	N/A	N/A
	5	7	0.9950	23	0.0268	40	1.8998	78	0.0183	127	2.9958
2	1	1	0.5550	20	1.0868	50	3.053	82	2.6169	113	0.2566
	2	1	0.1981	19	0.7639	49	0.2269	83	2.0576	113	0.8389
	3	1	0.8332	21	1.1704	48	3.5790	81	0.4615	110	2.0189
	4	1	1.1253	22	1.8774	46	1.6640	73	0.0524	111	0.8447
	5	1	1.0500	21	0.1506	48	1.3245	82	0.4500	110	04166
3	1	7	0.9968	34	0.5306	80	0.4455	97	0.0326	127	3.009
	2	7	2.0944	18	0.1030	31	0.0798	80	1.4715	N/A	N/A
	3	7	1.5067	78	0.5869	98	0.0331	N/A	N/A	N/A	N/A
	4	7	2.0874	79	1.3096	100	0.0064	127	3.0044	N/A	N/A
	5	7	1.1680	78	0.5193	127	3.008	N/A	N/A	N/A	N/A
4	1	2	0.0760	41	0.2450	61	0.1483	N/A	N/A	N/A	N/A
	2	4	0.4334	42	0.7514	63	0.3955	127	2.9697	N/A	N/A
	3	4	0.0664	16	0.0442	62	0.2867	82	0.2373	114	0.0073
	4	8	1.2975	43	0.0967	63	0.0791	127	2.9458	N/A	N/A
	5	19	0.5030	54	0.0738	92	0.0004	120	0.0039	N/A	N/A

N/A: Non available.

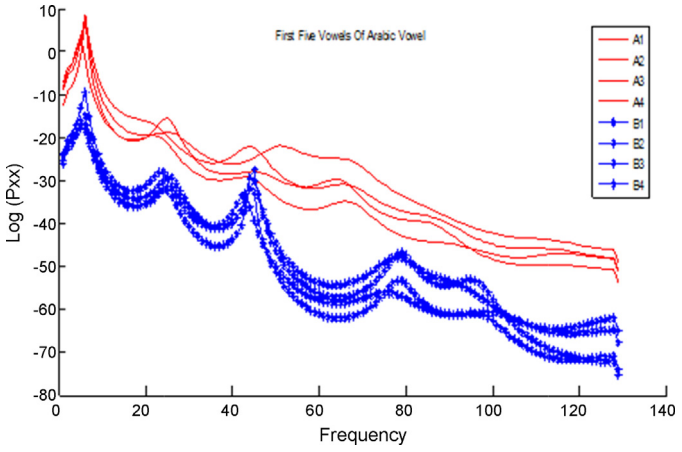

Fig. 2. PSD result showing the five formants of Arabic vowels for two speakers: A and B (vowel-dependent). Speakers A and B are both male. In A1, A2, A3, A4 four signals of one speaker were used, and in B1, B2, B3, B4 four signals of another speaker were used.

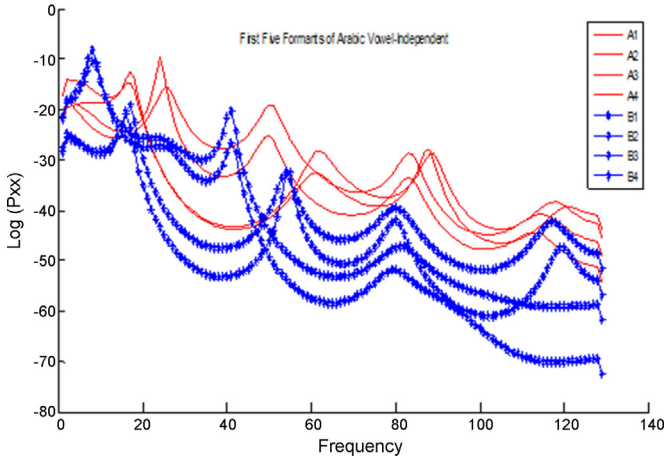

Fig. 3. PSD result showing the five formants of Arabic vowels for two speakers: A and B (vowel-independent). Speakers A and B are both male. In A1, A2, A3, A4 four signals of one speaker were used, and in B1, B2, B3, B4 four signals of another speaker were used.

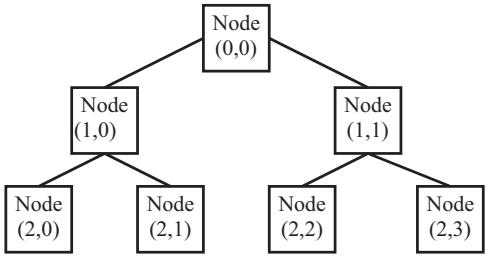

Fig. 4. Wavelet packet tree at depth two.

formants calculations for five speakers. Note that the indexes vary among different speakers, but are consistent for each speaker.

3.2. Features' extraction by entropies using WP

A general case of the wavelet decomposition is the WP method. The mother wavelet function is defined by

$$\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right) \tag{7}$$

where a and b are the scale and shift parameters, respectively. By varying a and b, the mother wavelet is scaled and translated. The wavelet transform is obtained by the inner product of the data function $x(t)$ and the mother wavelet $\psi(t)$

$$W_{\psi_X}(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t) * \psi\left(\frac{t-b}{a}\right) dt \tag{8}$$

The WP uses a recursive binary tree, as shown in Fig. 4, for the recursive decomposition of the data. A pair of low-pass and high-pass filters, denoted as $h[n]$ and $g[n]$, respectively, are used to generate two sequences, with the purpose of capturing different frequency sub-band features of the original signal. The two wavelet orthogonal bases generated from a previous node are defined as:

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} h[n]\psi_j^p(k-2^j n) \tag{9}$$

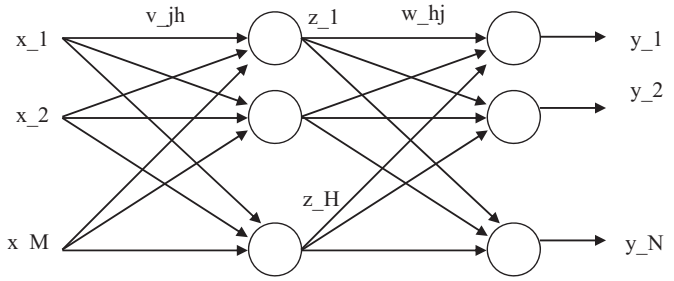

Fig. 5. Feed-forward multi-layer architecture.

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} g[n]\psi_j^p(k-2^j n) \tag{10}$$

where $\psi[n]$ is the wavelet function, while j and p are the number of decomposition levels and the number of nodes in the previous level, respectively [3,44]. In this study, WPT is applied at the feature extraction stage, but the large amount of data might cause some difficulties. Therefore, a better representation for the speech features is needed and is explained next.

For a given orthogonal wavelet function, a library of WP bases is generated. Each of these bases proposes a particular way of coding signals, maintaining global energy and reconstructing exact features. The WP is used to extract extra features to guarantee higher recognition rate. In this work, WPT is used at the feature extraction stage, but this data is not suitable for classifier due to a great amount of data length. Therefore, there is a need to find a better demonstration for the speech features.

The WP features' extraction method can be summarized as follows:

- Decompose the vowel signal at WP depth of level two with Daubechies type and calculate the Shannon entropy for each sub-signal. The WP extracts additional features to the Shannon entropy and therefore enhances the recognition rate.
- Calculate Shannon entropy for all seven nodes of wavelet packet using the equation

$$E(x) = -\sum_i x_i^2 \log(x_i^2) \tag{11}$$

where x is the signal under consideration and x_i are the signal coefficients that form the orthonormal basis. These seven entropies (in addition to the five formants) will be used to identify different speakers.

4. Classification

Two different classification approaches were used in the experimental investigation: feed-forward neural networks and probabilistic neural networks.

4.1. Feed-forward neural networks

Feed-forward networks are usually composed of multi-layer nodes (Fig. 5). The direction of the data goes only in one-way (i.e. forward). Consider a three-layer neural network which receives inputs $x_1, x_2, \ldots, x_m$, processes it to the hidden layer and then to the output layer to give the outputs $y_1, y_2, \ldots, y_n$.

The connecting arrows between the nodes have weights (the network variables). These weights are: v_{jh} (connects node input i with hidden node h) and w_{hj} (connects node input h with hidden node j). The outputs of the hidden and output layers at each pattern, k, are given by

$$z_h(k) = f\left(\sum_{i=1}^{M} v_{ih}x_i(k)\right) \quad h = 1, \ldots, H \tag{12}$$

$$y_j(k) = g\left(\sum_{h=1}^{H} w_{hj}z_h(k)\right) \quad j = 1, \ldots, N \tag{13}$$

where f and g are the activation function. The most commonly used activation functions are the sigmoidal and hyper-tangent.

Let the desired outputs be $d_1, d_2, \ldots, d_n$. The learning objective is to determine the weight values that minimize the difference between the desired and network outputs for all patterns. Let the error criterion be defined as follows:

$$SSE = \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{N}(e_j(k))^2 = \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{N}(y_j(k) - d_j(k))^2 \tag{14}$$

where k refers to the pattern number and j refers to the output node number. The weights are updated recursively

$$w_{hj}(iter+1) = w_{hj}(iter) + s(iter) \tag{15}$$

$$v_{ih}(iter+1) = v_{ih}(iter) + s(iter) \tag{16}$$

Here $s(iter)$ is the search direction at the specified iteration. An efficient update is the use of Levenberg–Marquardt method to find the search direction. This is provided next for the output weights w_{jh} [30]

$$E(k) = \sum_{j=1}^{N}(y_j(k) - d_j(k))^2 \tag{17}$$

$$J = \frac{\partial E(k)}{\partial w_{ih}} \tag{18}$$

$$s(iter) = -(J^T J + \lambda I)^{-1}(J^T E(k)) \tag{19}$$

where J is the Jacobian matrix and I is the identity matrix. The parameter λ is updated for each search step: increased if the algorithm is divergent and decreased if the algorithm is convergent. There are two main advantages of using this parameter: enforce descending function values in the optimization sequence and increase the numerical stability of the algorithm.

The derivative of the error w.r.t., the hidden weights, v_{ih}, involves more steps: The target values at the hidden layer are not available and a chain rule is used to approximate the hidden error. This algorithm is called *backpropagation* [29].

A major application for neural networks is pattern classification. In this paper, the formant and wavelet information presented in the two previous sections were used as input/output data for the neural network for classification. The total number of inputs used was 12 (five formants and seven entropies).

4.2. Probabilistic neural network

Probabilistic neural networks are implementations of statistical algorithms and are generally used as classifiers. These networks are unsupervised feed-forward networks with four layers: input, pattern, summation, and output. A probabilistic function, such as the Gaussian, is used for each pattern node. The network weights are updated according to the input patterns. The patterns are then classified using the nearest-neighborhood function according to the Gaussian classifiers. The mean and variance for each node function can also be updated during training to minimize the distance

between the patterns and their closest classifiers. More theoretical details of such networks can be found in [31].

5. Results and discussion

The experimental setup was as follows: Speech signals were recorded via PC-sound card with a spectral frequency of 4000 Hz and sampling frequency of 8000 Hz. Eighty people participated in the recordings with each person recording a minimum of 20 times. Each recording was a partially spoken word emphasizing a particular Arabic vowel (E, A and O) (see Appendix A). The age of the speakers varied from 30 to 45 years and included 48 males and 32 females. The recording process was provided in normal university office conditions (i.e. recording in an office with the door closed and with no obvious noisy surroundings).

Even though the methods proposed for speaker and speech recognition, diacritics or diacritics restoration have been maturing over time, they are still inadequate in terms of accuracy [47,48]. In the approach we present in this paper, we propose research study of the speaker recognition by means of speech vowels signal. Therefore, the presented study may be considered as an investigation work aiming to build a system that classifies the speakers by only a short part of his speech signal. This speech part signal is the separated spoken vowel. This helps greatly in case only uncompleted word speech signals are available. We solved the problem by using the conventional recognition method (feature extraction and then classification). This assists to find out a distinguished speaker recognition system by only the vowels speech signals. This approach is based on a combination between the formants and seven Shannon entropy wavelet packets as a feature extraction method and neural network for classification.

The extracted features (five formants and seven entropies) were calculated for each person from filtered speech signals. Filters using the multistage wavelet enhancement method were used [23]. The input matrix, X, contained n columns (that represented the number of speakers). Each column had 12 entrees: five formants and seven entropies.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \dots & \dots & \dots & \dots \\ x_{12,1} & x_{12,2} & \dots & x_{12,n} \end{bmatrix} \tag{20}$$

One way to configure the desired output matrix is to use the number of columns that is equal to the number of speakers. Binary-decoded columns were used where the binary value of each column represented the speaker's order. As an example, for six speakers, the first column would be '1000', the second would be '0100', the third would be '1100', ..., and the sixth column would be '0110'. The matrix used is shown next

$$D = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{21}$$

However, in order to improve the performance of the network, several patterns were recorded for each person (not just one column). Therefore, the final input and output matrices were of the form

$$X = [\, X_1 \quad X_2 \quad \dots \quad X_r \,] \tag{22}$$

$$D = [\, D_1 \quad D_2 \quad \dots \quad D_r \,] \tag{23}$$

where r is the number of recordings for each person.

The neural network parameters used were determined empirically for the best performance results and are provided in Table 2.

Table 2
Parameters used for the neural network.

Functions	Description
Network type	Feed-forward back propagation
No. of layers	Three layers: input, hidden, and output
No. of neurons in layers	12 – inputs, 30 – hidden, and 4 – outputs
Training function	Levenberg–Marquardt
Performance function (mse)	10^{-5}
Differentiable transfer function	Sigmoidal
No. of epochs	200
Maximum validation failures	5
Minimum performance gradient	10^{-10}
Initial mu	10^{-2}
mu increase factor	10
mu decrease factor	0.1
Maximum mu	10^{10}

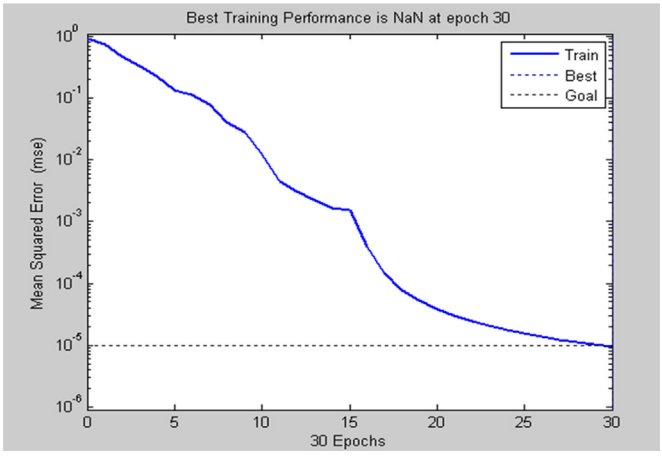

Fig. 6. Network training convergence.

The recorded signals were matched with the speaker's identity and were used as input/output pairs as explained in Section 4. A typical network convergence run is shown in Fig. 6.

The conducted experiments tackled two main recognition tasks: the verification task and the identification task. All the conducted experiments used a database with 80 speakers. For each speaker, the data was divided between training and testing data.

For verification [55], the system was tested for its accuracy and effectiveness on the 80 speakers (i.e. classes). At each run, 50% of the signals were used from *chosen class* (i.e. *correct speaker*) while the other 50% of the signals were used from *another class* (i.e. *imposter*). The verification recognition rates' results were determined with regard to the False Positive Error (FPE), which refers to the total number of the accepted imposter testing signals divided by the total number of the testing signals, and the False Negative Error (FNE), which refers to the number of rejected genuine testing signals divided by the total number of the testing signals. The recognition rate was calculated as 100% minus the total of FPE and FNE. We would like to note here that the proposed method had a 100% recognition rate on the training data.

The results are compared to two established methods in literature: MFCC [32] with ANN referred to as MFCC-NN, and LPC with ANN referred to as LPC-NN [33]. The results are tabulated in Table 3. Note that FWENN had better recognition rates: An increase of about 12% over MFCC-NN and 27% over LPC-NN. Those are significant results showing that the proposed method is more suitable for identifying speakers when 'partially-spoken vowels' are used as inputs.

In the next experiments, identification task was performed where the first half of the signals in each class were used for training and the second half for testing. The proposed method (i.e. FWENN)

Table 3
Recognition rate for verification of experiment results.

Recognition rate [%]	FNE [%]	FPE [%]	Method
89.16	5.60	5.23	FWENN
77.32	14.23	8.11	MFCC-NN
61.88	15	23.12	LPC-NN

Table 4
Average recognition rate results.

Identification method	Number of speakers	Recognition rate [%] (Vowel-dependent)	Recognition rate [%] (Vowel-independent)
FWENN	80	90.09	82.50
MFCC-NN	80	79.66	74.03
LPC-NN	80	66.63	59.45
DWT-NN	80	81.44	79.32
GWP-NN	80	85.47	80.07

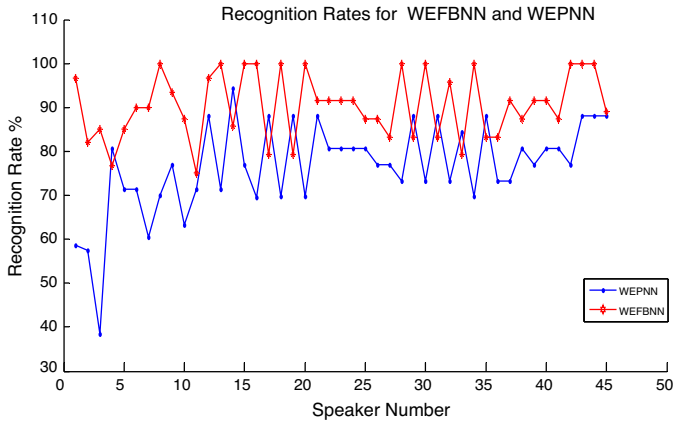

Fig. 7. The experimental results of two different classification approaches (regular feed-forward and probabilistic) used in the experimental investigation for comparison of vowel-dependent system.

was further compared to two more modern methods that use wavelet transform: Genetic Wavelet Packet with ANN denoted by GWP-NN [34] and Discrete Wavelet Transform (at level 5) with the proposed feature extraction method and ANN denoted by DWT-NN. The results tabulated in Table 4 can be used to compare recognition rates among five methods. Results indicated that proposed method is superior with highest recognition rates of 90% and 82.5% for vowel-dependent and vowel-independent respectively.

In WP case, it was found that the recognition rates improved upon increasing the number of features (by increasing WP level). However, the improvement implies a trade-off between the recognition rate and extraction time (see Table 7).

To further test the FWENN method, the feed-forward neural network was replaced with a probabilistic neural network. Fig. 7 shows the experimental results of the two classification approaches within FWENN. The recognition rates for the database of the probabilistic network reached the lowest values with average 76.56%. The best average recognition rate selection obtained was of 90.09% for FWENN.

A comparative study of the proposed feature extraction method with other WP-based feature extraction methods was performed. The Eigen vector with LPC [43] in conjunction with WP (LPC-WP), Wavelet Packet energy Index Distribution method (WPID) [3], and Sure entropy in conjunction with WP Formants at level seven (SWPF) [44] were employed for comparison. For all these methods feed-forward neural network classifier was utilized. Finally, our proposed feature extraction method with Support Vector Machines (SVM) classifier (FSVM) [47]. The results were conducted

Table 5
A comparison among different WP-based methods.

Recognition rate [%]	Identification method
90.09	FWENN
84.34	LPC-WP
83.10	WPID
80.76	SHWPF
87.21	FSVM

Table 6
Comparison between FWENN and DWT-NN in noisy environments.

Identification method	Recognition rate [%]	
	0 dB	5 dB
FWENN	48.44	68.45
DWT-NN	46.10	50.23

Table 7
The computational complexity in terms of feature extraction vector length and simulating time by Shannon entropy via WP.

WP level	Level 2	Level 5	Level 7
Simulation time (s)	0.58	1.31	3.40
Vector length	7	63	255
Recognition rate	90.09%	91.04	91.69%

for the whole recorded database. The best recognition rate selection obtained was 90.09% for proposed method (Table 5).

Another experiment was conducted to assess the performance of the system in the noisy environments for vowel-dependent system. Table 6 summarizes the results of speaker identification corresponding to white Gaussian noise with SNR of 0 dB and 5 dB references. Two approaches were used in the experimental investigation for comparison: FWENN and DWT-NN. The best recognition rate selection obtained was 48.44 (with 0 dB) and 68.45 (with 5 dB) for DWT-NN. The reason of DWT being successful over WP is that the feature vector is obtained from level 5, where the sub-signals were filtered in lower depth than in WP at level 2. Our proposed method could easily overcome this problem by increasing the number of WP tree levels from 2 to 5 or 7. However, the improvement implies a trade-off between the recognition rate and increasing the features' dimensionality.

Very few researchers studied the use of formants for speaker recognition [56]. This work concentrated on the use of formants of spoken vowels in order to recognize different speakers. Additional features extracted by Shannon entropy enhanced the recognition rate by a further 7%. However, the improvement implied a trade-off between the recognition rate and extraction time.

The results of different Shannon entropy feature vector lengths by several WP levels are shown in Table 7. Notice that the recognition rate of the proposed method was enhanced from 90.09% to 91.69% when the number of WP nodes increased from 7 to 255 (from level 2 to 7). The fact that we can get very good performance with a short vector (i.e. only 12 coefficients) is a major contribution of this work.

6. Conclusion

In this paper, a new method for speaker recognition (verification and identification) was described. The method used feature extractions (formants and Shannon entropy) as inputs to a neural network for classification.

Power spectrum using Yule–Walker equations was used for identifying the formants while wavelet packet was used for calculating the entropies. Furthermore, two different network

architectures were investigated: feed-forward neural networks and probabilistic neural networks.

The significance of this work is that the recorded signals used for recognition were vowels. Advantages of using vowels include the ability to recognize speakers when only partially-recorded words are available. This may be useful for people with speaking disability, such as deaf-mute persons.

The proposed method, FWENN, was compared to several established methods in literature (MFCC-NN, LPC-NN, DWT-NN, and GWP-NN) using vowel-dependent and vowel-independent recordings from a combination of 80 speakers under different noise levels. Experimental results showed that the proposed method had a high recognition rate for verification and identification. In comparison to other published methods, results indicated that the proposed method is superior with a highest recognition.

Acknowledgements

This project's paper was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, under grant no. 12-135-35-RG. The authors, therefore, acknowledge with thanks DSR technical and financial support.

Appendix A. Arabic vowels

Arabic language is one of the most important and broadly spoken language in the world. An expected number of 350 million people distributed all over the world (mainly covering 22 Arabic countries) speak Arabic. Arabic is a Semitic language that is characterized by the existence of particular consonants like pharyngeal, glottal and emphatic consonants. Furthermore, it presents some phonetics and morpho-syntactic particularities. The morpho-syntactic structure is built, around pattern roots (CVCVCV, CVCCVC, etc.), as shown in [35].

The Arabic alphabet consists of 28 letters that can be extended to a set of 90 by additional shapes, marks, and vowels. The 28 letters represent the consonants and long vowels such as ى and اٰ (both pronounced as/a:/), ي (pronounced as/i:/), and و (pronounced as/u:/). The short vowels and certain other phonetic information such as consonant doubling (shadda) are not represented by letters, but by diacritics. A diacritic is a short stroke placed above or below the consonant. We find three short vowels: fatha: it represents the /a/ sound and is an oblique dash over a letter, damma: it represents the /u/ sound and has shape of a comma over a letter and kasra: it represents the /i/ sound and is an oblique dash under a letter. The long and short vowels are presented in Tables A.1 and A.2.

Table A.1
Long Arabic vowels.

Long vowel name	Connected with letter 'ب' (sounds B)	Pronunciation
Alef	فب	/baa/
Waw	+F	/buu/
Yaa□	يب	/bii/

Table A.2
Short Arabic vowels.

Short vowel name (diacritics)	Diacritics above or below letter 'ب' (sounds B)	Pronunciation
Fatha	بَ	/ba/
Damma	بF	/bu/
Kasra	بِ	/bi/
Tanween Alfath	بًا	/ban/
Tanween Aldam	بFF	/bun/
Tanween Alkasr	بٍ	/bin/
Sokun	بْ	/b/

References

[1] D. Avci, An expert system for speaker identification using adaptive wavelet sure entropy, Expert Syst. Appl. 36 (2009) 6295–6300.
[2] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digit. Signal Process. 10 (1–3) (2000) 19–41.
[3] J.-D. Wu, B.-F. Lin, Speaker identification using discrete wavelet packet transform technique with irregular decomposition, Expert Syst. Appl. 36 (2009) 3136–3143.
[4] R. Sarikaya, B.L. Pellom, J.H.L. Hansen, Wavelet packet transform features with application to speaker identification, in: Proceedings of the IEEE Nordic Signal Processing Symposium, 1998, pp. 81–84.
[5] E.S. Fonseca, R.C. Guido, P.R. Scalassara, C.D. Maciel, J.C. Pereira, Wavelet time–frequency analysis and least squares GWPNN support vector machines for the identification of voice disorders, Comput. Biol. Med. 37 (2007) 571–578.
[6] R.R. Coifman, M.L. Wickerhauser, Entropy based algorithms for best basis selection, IEEE Trans. Inf. Theory 32 (1992) 712–718.
[7] E. Visser, M. Otsuka, T. Lee, A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments, Speech Commun. 41 (1992) 393–407.
[8] J. Malkin, X. Li, J. Bilmes, A Graphical Model for Formant Tracking, SSLI Lab, Department of Electrical Engineering, University of Washington, Seattle, 2005.
[9] M.P. Gelfer, V.A. Mikos, The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels, J. Voice 19 (4) (2007) 544–554.
[10] J. Bachorowski, M. Owren, Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech, J. Acoust. Soc. Am. 106 (2) (1999) 1054–1063.
[11] X. Huang, A. Acero, H.-W. Hon, Spoken Language Processing, Prentice Hall PTR, 2001.
[12] S. Kadambe, G.F. Boudreaux-Bartels, Application of the wavelet transform for pitch detection of speech signals, IEEE Trans. Inf. Theory 32 (March) (1992) 712–718.
[13] Acero, Formant analysis and synthesis using hidden Markov models, in: Proc. Eur. Conf. Speech Communication Technology, 1999.
[14] L. Deng, A. Bazzi, Acero, Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint, in: Proc. Eur. Conf. Speech Communication Technology, 2003.
[15] L. Deng, L. Lee, H. Attias, A. Acero, A structured speech model with continuous hidden dynamics and prediction residual training for tracking vocal track resonances, in: IEEE ICASSP, 2004.
[16] C.A. Norton, S.A. Zahorian, Speaker verification based on speaker position in a multidimensional speaker identification space, in: Intelligent Engineering Systems Trough Artificial Neural Networks, 5ASME Press, New York, 1995, pp. 739–744.
[17] M. Zaki, A. Ghalwash, A. Elkouny, Speaker recognition system using a cascade neural network, Int. J. Neural Syst. 7 (1996) 203–212.
[18] M.W. Mak, S.Y. Kung, Estimation of elliptical basis function parameters by EM algorithm with application to speaker recognition, IEEE Trans. Neural Netw. 11 (2000) 961–969.
[19] M. Homayounpour, G. Chollet, Neural nets approach to speaker verification, in: Proceedings of ICASSP95 (Proceedings of the International Conference of Acoustics, Speech and Signal Processing, '95), 1995, pp. 335–356.
[20] N.P. Reddy, O.A. Buch, Speaker verification using committee neural network, Comput. Methods Programs Biomed. 72 (2003) 109–115;
Artificial Neural Networks, vol. 5, ASME Press, New York, 2000, pp. 739–744.
[21] N.P. Reddy, D. Prabhu, S. Palreddy, V. Gupta, S. Suryanarayanan, E.P. Canilang, Redundant neural networks for reliable diagnosis: applications to dysphagia diagnosis, in: C. Dagli, A. Akay, C. Philips, B. Fernadez, J. Ghosh (Eds.), Intelligent Engineering Systems Through Artificial Neural Networks, vol. 5, ASME Press, New York, 1995, pp. 739–744.
[22] A. Das, N.P. Reddy, J. Narayanan, Hybrid fuzzy logic committee neural networks for recognition of swallow acceleration signals, Comput. Methods Programs Biomed. 64 (2001) 87–99.
[23] K. Daqrouq, I. Abu Sbeih, O. Daoud, E. Khalaf, An investigation of speech enhancement using wavelet filtering method, Int. J. Speech Technol. 13 (2) (2010) 101–115.
[24] K. Daqrouq, E. Khalaf, A. Al-Qawasmi, T. Abu-Hilal, Wavelet formants speaker identification based system via neural network, Int. J. Recent Trends Eng. 2 (5) (2009).
[25] W. Al-Sawalmeh, K. Daqrouq, O. Daoud, A. Al-Qawasmi, Speaker identification system-based Mel frequency and wavelet transform using neural network classifier, Eur. J. Sci. Res. 41 (4) (2010) 515–525.
[26] A. Cherif, Bouafif, T. Dabbabi, Pitch detection and formants analysis of Arabic speech processing, Appl. Acoust. 62 (2001) 1129–1140.
[27] E. Avci, D. Hanbay, A. Varol, An expert discrete wavelet adaptive network based fuzzy inference system for digital modulation recognition, Expert Syst. Appl. 33 (2006) 582–589.
[28] E. Avci, A new optimum feature extraction and classification method for speaker recognition: GWPNN, Expert Syst. Appl. 32 (2007) 485–498.
[29] S. Haykin, Neural Networks: A Comprehensive Foundation, Macmillan College Publishing Company, NY, 1994, ISBN 10:0023527617/ISBN 13:9780023527616, From Cheshire Book Centre (CHESHIRE, ENG, United Kingdom).
[30] R. Isermann, M. Munchhof, Identification of Dynamic Systems: An Introduction with Applications, Springer-Verlag, Berlin, Heidelberg, 2011.

[31] T. Ganchev, D. Tasoulis, M. Vrahatis, D. Fakotakis, Generalized locally recurrent probabilistic neural networks with application to text-independent speaker verification, Neurocomputing 70 (2007) 1424–1438.
[32] T. Ganchev, N. Fakotakis, G. Kokkinakis, Comparative evaluation of various MFCC implementations on the speaker verification task, in: Proceedings of the SPECOM-2005, vol. 1, 2005, pp. 191–194.
[33] Y. Bennani, P. Gallinari, Neural networks for discrimination and modelization of speakers, Speech Commun. 17 (1995) 159–175.
[34] A. Engin, A new optimum feature extraction and classification method for speaker recognition: GWPNN, Expert Syst. Appl. 32 (2007) 485–498.
[35] I. Zitouni, R. Sarikaya, Arabic diacritic restoration approach based on maximum entropy models, Comput. Speech Lang. 23 (2009) 257–276.
[36] R. Behroozmand, F. Almasganj, Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis, Comput. Biol. Med. 37 (2007) 474–485.
[37] D. Mashao, M. Skosan, Combining classifier decisions for robust speaker identification, Pattern Recognit. 39 (January (1)) (2006).
[38] R.V. Pawar, P.P. Kajave, S.N. Mali, Speaker identification using neural networks, Proc. World Acad. Sci. Eng. Technol. 7 (August) (2005).
[39] S. Chakroborty, G. Saha, Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter, Int. J. Signal Process. 5 (1) (2009).
[40] H.B. Kekre, V. Kulkarni, Comparative analysis of speaker identification using row mean of DFT, DCT, DST and Walsh transforms, Int. J. Comput. Sci. Inf. Secur. 9 (1) (2011).
[41] H. Kekre, V. Kulkarni, Speaker identification using row mean of DCT and Walsh Hadamard transform, Int. J. Comput. Sci. Eng. (2011) 6–12.
[42] S. Singh, E.G. Rajan, Vector quantization approach for speaker recognition using MFCC and inverted MFCC, Int. J. Comput. Appl. 17 (March (1)) (2011) 1–7.
[43] S. Uchida, M.A. Ronee, H. Sakoe, Using eigen-deformations in handwritten character recognition, in: Proceedings of the 16th ICPR, vol. 1, 2002, pp. 572–575.
[44] K. Daqrouq, Wavelet entropy and neural network for text-independent speaker identification, Eng. Appl. Artif. Intell. 24 (2011) 796–802.
[46] S. Rahati, Quchani, K. Rahbar, Discrete word speech recognition using hybrid self-adaptive HMM/SVM classifier, J. Tech. Eng. 1 (2) (2007) 79–90.
[47] P. Rama Koteswara Rao, Pitch and pitch strength based effective speaker recognition: a technique by blending of MPCA and SVM, Am. J. Sci. Res. 59 (2012) 11–22.
[48] Y. Alotaibi, A. Hussain, Speech recognition system and formant based analysis of spoken Arabic vowels, in: Proceedings of the First International Conference, FGIT, Jeju Island, Korea, December 10–12, 2009.
[49] Y. Alotaibi, A. Hussain, Formant based analysis of spoken Arabic vowels, in: Proceedings BioID_MultiComm, Madrid, Spain, 2009.
[50] F. Nolan, C. Grigoras, A case for formant analysis in forensic speaker identification, Speech Lang. Law 12 (2) (2005) 143–173.
[51] C. Grigoras, Forensic voice analysis based on long term formant distributions, in: 4th European Academy of Forensic Science Conference, June 2006.
[52] P. Rose, Forensic Speaker Identification, Taylor & Francis, London, 2002.
[53] J. Nirmal, M. Zaveri, S. Patnaik, P. Kachare, Voice conversion using General Regression Neural Network, Appl. Soft Comput. 24 (November) (2014) 1–12.
[54] X. Hong, S. Chen, A. Qatawneh, K. Daqrouq, M. Sheikh, A. Morfeq, A radial basis function network classifier to maximize leave-one-out mutual information, Appl. Soft Comput. 23 (October) (2014) 9–18.
[55] S. Sarkar, K. Sreenivasa Rao, Stochastic feature compensation methods for speaker verification in noisy environments, Appl. Soft Comput. 19 (June) (2014) 198–214.
[56] V. Asadpour, M.M. Homayounpour, F. Towhidkha, Audio–visual speaker identification using dynamic facial movements and utterance phonetic content, Appl. Soft Comput. 11 (March (2)) (2011) 2083–2093.
[57] R.H. Laskar, D. Chakrabarty, F.A. Talukdara, K. Sreenivasa Raoc, K. Banerjee, Comparing ANN and GMM in a voice conversion framework, Appl. Soft Comput. 12 (November (11)) (2012) 3332–3342.
[58] R. Halavati, S.B. Shouraki, S.H. Zadeh, Recognition of human speech phonemes using a novel fuzzy approach, Appl. Soft Comput. 7 (June (3)) (2007) 828–839.
[59] M. Sarma, K.K. Sarma, An ANN based approach to recognize initial phonemes of spoken words of Assamese language, Appl. Soft Comput. 13 (May (5)) (2013) 2281–2291.
[60] S. Jothilakshmi, Automatic system to detect the type of voice pathology, Appl. Soft Comput. 21 (August) (2014) 244–249.