



Wavelet LPC with Neural Network for Spoken Arabic Digits Recognition System

K. Daqrouq^{1*}, M. Alfaouri², A. Alkhateeb¹, E. Khalaf¹ and A. Morfeq¹

¹Electrical and Computer Engineering Department, King Abdulaziz University, Saudi Arabia.

²Department of Electrical and Communications, Al-balqa Applied University Faculty of Engineering and Technology, Jordan.

Authors' contributions

This work was carried out in collaboration between all authors. Author KD designed the study, performed the statistical analysis, wrote the protocol and wrote the first draft of the manuscript and managed literature searches. Authors MA, AA, EK and AM managed the analyses of the study and literature searches. All authors read and approved the final manuscript.

Original Research Article

Received 18th July 2013
Accepted 2nd January 2014
Published 18th January 2014

ABSTRACT

The crucial problem of Arabic recognition systems is the availability of several dialects in Arabic language, particularly those with sound variations. Therefore, low recognition rate is encountered as a result of such an environment. In this research paper the authors presented dialect-independent via an enormously effectual wavelet transform (WT) based Arabic digits classier. The proposed system may be divided into two main blocks the features extraction method by combining wavelet transform with the linear prediction coding (LPC) and the classification by probabilistic neural network (PNN). The proposed classier provided a high recognition rate reaching up to 100%, in some cases, and an average rate of about 93% based on speaker-independent system. 450 Arabic spoken digit tested signals were used. The performance of the system in the noisy environment was investigated. The obtained results are very promising; however, the larger testing database may provide more credible results.

Keywords: Arabic digits; wavelet transform; speech recognition; LPC; neural network.

1. INTRODUCTION

These days interactive voice response systems are increasingly and widely used, especially involving speaker-independent recognition of given vocabularies conveyed over the telephone network or microphones investigated by [1]. The recent increase shed some light on crescendo activity in mobile communication domain inaugurating a new era of opportunities for applications of speech recognition including digits and sentences. In text to speech, or vice versa, as well as incredibly vital issues in many computer applications, where English language has achieved immense success and forms the major part of interest. On the other hand, Arabic language speech recognition has slight attraction due to its various dialects and several alphabets forms.

The major works which study the speech recognition in Arabic language deal with the morphological structure [1-3] or the phonetic features in order to recognize the distinct Arabic phonemes (pharyngeal, geminate and emphatic consonants) [4,5] and discuss their further implication in a larger vocabulary speech system. This opens an interesting field for researchers insofar as the applications in terms of implementation of recognition system dedicated and devoted to spoken isolated words or continuous speech are not extensively explored, and only a few examples have been improved and ameliorated in this research paper. A derivative scheme, named the Concurrent GRNN, implemented for accurate Arabic phonemes identification in order to automate the intensity and formants-based feature extraction, was studied in [6]. The validation tests expressed in terms of recognition rate obtained with free of noise speech signals were up to 93.37%. An isolated word speech recognition using the RNN was investigated in [7]. The achieved accuracy is 94.5% in terms of recognition rate in speaker-independent mode and 99.5% in speaker-dependent mode. Several Arabic speech recognition systems were discussed in [8].

The Fuzzy C-Means method has been added to the traditional ANN/HMM speech recognizer using RASTA-PLP features vectors. The Word Error Rate (WER) is over 14.4%. With the same approach, a method using data fusion gave a WER of 0.8% [9]. However, this method was tested only on one personal corpus and the authors indicated that the obtained improvement needed the use of three neural networks working in parallel. Another alternative hybrid method was proposed by [9], where the Support Vector Machine (SVM) and the K nearest neighbor (KNN) were substituted for the ANN in the traditional hybrid system; better recognition rate was achieved as a result, but it did not exceed 92.72% for KNN/HMM and 90.62% for SVM/HMM.

A new algorithm to recognize separate voices of some Arabic words was presented in [10], the digits from zero to ten were presented and compared. For feature extraction, transformation and hence recognition, the algorithm of minimal Eigen values of Toeplitz matrices along with other methods of speech processing and recognition were used to achieve a more accurate recognition rate of speaker-independent mode. The success rate obtained in the presented experiments was almost ideal and exceeded 98% in many cases. A hybrid method has been applied to Arabic digits recognition by [11].

From literatures papers presented by researchers, neural networks were used to identify features of Arabic language, such as the emphasis, germination and relevant vowel lengthening [7]. This was studied using ANN and other techniques [12] where many systems and configurations were considered, including time delay neural networks (TDNNs). Bearing in mind ANNs were used to recognize the 10 Malay digits [13], Saeed and Nammous (2005a) proposed a heuristic method of Arabic digit recognition, using the Probabilistic

Neural Network (PNN). The use of a neural network recognizer, with a nonparametric activation function, constitutes a promising solution to increase the performances of speech recognition systems, particularly in the case of Arabic language. [14,15] demonstrated the advantages of the GRNN speech recognizer over the MLP and the HMM in a quiet environment. But the method investigated by authors is applicable for a quiet environment. In extremely noisy environments, the recognition performance degrades considerably. Robustness to noise is essential for a professional using recognition systems particularly in mobile networks context [16,17]. Many studies have been conducted in this track [18,19]. Numerous pre-processing techniques have been developed in order to reduce or eliminate the noise effects in the speech before adding to a recognizer. Enhancement procedures like spectral subtraction [20,21] remove ambient surrounding noise. The transmission effects are reduced using equalization techniques such as cepstral normalization and adaptive filtering [22,23].

2. THE ARABIC LANGUAGE

Arabic language is one of the most widely spoken languages in the world, with an expected number of 350 millions speakers covering 22 Arabic countries. Arabic is a Semitic language, which 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 structures are built around pattern roots (CVCVCV, CVCCVC, etc.) [24].

The Arabic alphabet consists of 28 letters that can be extended to a set of 90 by additional shapes, marks, and vowels [24]. The 28 letters represent the consonants and long vowels such as *ā* and *ī* (both pronounced as /a:/), *ī* (pronounced as /i:/), and *ū* (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. Table 1 shows the complete set of Arabic diacritics. We split the Arabic diacritics into three sets: short vowels, doubled case endings, and syllabification marks. Short vowels are written as symbols either above or below the letter in text with diacritics, and dropped all together in text without diacritics. 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 the shape of a comma over a letter and kasha: it represents the /i/ sound and is an oblique dash under a letter (see Table 1) [24].

Table 1. Short Arabic vowels

Short Vowel Name (Diacritics)	Diacritics above or below Arabic letter for example B
Short / a/ -Fatha	/ba/
Short / u/- Damma	/bu/
Short / i/- Kasra	/bi/
Short / an/ TanweenAlfath	/ban/
Short / on/ TanweenAldam	/bun/
Short / en/ TanweenAlkasr	/bin/
Consonant-Sokun	/b/

It is important to realize that what we typically refer to as “Arabic” is not single linguistic variety; rather, it is a collection of different dialects. Classical Arabic is an older, literary form of the language, exemplified by the type of Arabic used in the Quran. Modern Standard

Arabic (MSA) is a version of Classical Arabic with a modernized vocabulary. MSA is a formal standard common to all Arabic-speaking countries. It is the language used in the media (newspapers, radio, TV), in official speeches, in courtrooms, and, generally speaking, in any kind of formal communication. However, it is not used for everyday, informal communication, which is typically carried out in one of the local dialects. The dialects of Arabic can roughly be divided into two groups: Western Arabic, which includes the dialects spoken in Morocco, Algeria, Tunisia, and Libya and Eastern Arabic, which can be further subdivided into Egyptian, Levantine, and Gulf Arabic. These various dialects differ considerably from each other and from Modern Standard Arabic. Differences affect all levels of the language, i.e. pronunciation, phonology, vocabulary, morphology and syntax. Table 1 lists examples of the differences between Egyptian Colloquial Arabic (ECA) and Modern Standard Arabic. ECA is that dialect which is most widely understood through-out the Arabic-speaking world due to a large number of TV programs which are produced in Egypt and exported to other Arabic countries. Native speakers from different dialect regions are for the most part capable of communicating with each other, especially if they have had some previous exposure to the other speaker's dialect. However, widely differing dialects, such as Moroccan Arabic and the Iraqi dialect, may hinder communication to the extent that speakers adopt Modern Standard Arabic as a lingua franca.

Many issues of Arabic language, such as, the phonology and the syntax, do not present difficulty for automatic speech recognition. Standard, language-independent techniques for acoustic and pronunciation modeling, such as context-dependent phones, can easily be applied to model of the acoustic-phonetic properties of Arabic. The most difficult problems in developing high-accuracy speech recognition systems to Arabic language are the predominance of non-diacritized text material, the enormous dialectal variety and the morphological complexity.

The principle problem of the dialectal variety is due to the current lack of training data for conversational Arabic; while, MSA data can readily be acquired from various media sources. Finally, morphological complexity is approved to present solemn problems for speech recognition. A high scale of affixation, derivation etc. contributes to the explosion of unlike word forms, making it difficult, if not impossible, to robustly estimate language model probabilities. Prosperous morphology also leads to elevated out-of-vocabulary rates and bigger search spaces during decoding, thus slowing down the recognition process [3].

2.1 Arabic Digits

Arabic digits from zero to nine are polysyllabic words except the first one, zero, which is a monosyllable word [2]. Table 2 shows the 10 Arabic digits along with pronunciation, signals and number of syllable [7].

Table 2. Arabic different dialects; modern, Egyptian, Jordanian and Palestinian

Gloss	MSA	EAD	JAD	PAD
'Three'	thā-lā-thāh	tā-lā-tāh	thā-lā-thēh	tā-lā-tēh
'Eight'	thā-mā-nē-yah	tā-mā-n-yah	thā-mā-n-yeh	tā-mā-n-yeh
'Two'	?ith-nān	te-nān	?ith-nen	?it-nān

Compared to other languages, Arabic digits are much more elongated. They include two to four syllables, while French, English and Mandarin digits are single or double syllables. Arabic digits can be considered as representative elements of language, because more than half of the phonemes of the Arabic language are included in the 10 digits. The fricative and

plosive consonants are more dominant and characterized by the presence of noise in the high-frequency band spectrum. In fact, these consonants are easily corrupted by noise sources. Therefore, speech recognition systems usually fails to identify them in adverse conditions [24].

2.1.1 Similarities between arabic digits

The similarity between Arabic digits, in term of pronunciation and signal morphology may lead to big recognition confusion rate [7]. In this research paper we present some of these Arabic digits similarities:

- When digit 0 is investigated against digit 1, we can observe that the second phonemes in both digits 0 and 1 are vowels /l / and /a: / respectively, and they have high similarity depending on their spectrograms. Power Spectral Density (PSD) of the two digits contains some common maximum peaks (see Fig. 1). An overlap between these phonemes may occur, hence causing a misleading match between these digits.
- The similarities between digits 0 and 2 are very little and this result is evident when spectrograms and PSD are studied. This is also confirmed by the results of digit recognition system, except noise contaminated digits.
- By investigating digits 1 and 2, we can encounter that there is a large dissimilarity between these two digits especially in the second part of their spectrograms. Digit 2 has a long vowel in the second syllable and the same syllable starts with the nasal phoneme /n/ and ends with the same nasal phoneme.
- Spectrograms of digits 1 and 3 contain big similarities. PSD of the two digits have common two core peaks at 40 and 10 in the frequency scale (see Fig. 1). The digit recognition systems, always produces immense confusions [7,24].
- Digit 1and 4 have the same penultimate phoneme, a short vowel /a/. There are moderate common peaks in PSD curves. On the other hand, there were small spectrogram similarities.
- There is little similarity between digits 2 and 3 the number and the type of syllable.
- Digits 3 and 8 have high pronunciation similarities; the sounds /h/ and /a/ are the first and second phonemes in both digits (i.e., the first syllable in both digits are exactly the same).
- There is a similarity between digits 4 and 5 in the last two phonemes. Phonemes /a/ and /h/ are the final two phonemes in both digits. Also the second phonemes in each are also the same
- There is large pronunciation dissimilarity between digits 4 and 6. Digit 6 consists mostly of unvoiced. Consonants, namely, /s/ and /t/ (twice), while digit 4 consists mostly of voiced phonemes, namely, vowels and /r/, /b/, and /? / Consonants. There is a low similarity between these digits in terms of PSD and spectrogram. There are no recognition system confusions.
- Digits 4 and 7 have a high similarity in terms of pronunciation but are different in terms of PSD and spectrogram.
- Digits 6 and 7 have the identical patterns of syllables, CVC-CVC and they have the same first phoneme, /s/, but are different in term of PSD and spectrogram.


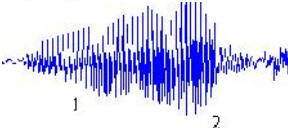
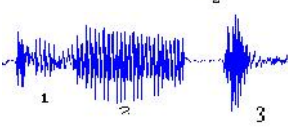

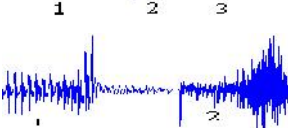
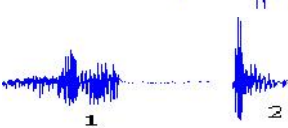
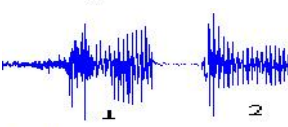
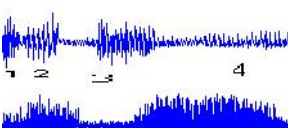
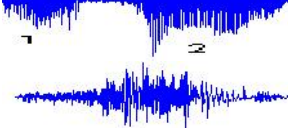

Digit	Pronunciation	Speech Signal	No. of Syllables
1	wā-hid		2
2	?ith-nān		2
3	thā-lā-thāh		3
4	'aār-bā-?āh		3
5	khm-sāh		2
6	sēt-tāh		2
7	sūb-?āh		2
8	thā-mā-nē-yah		4
9	tēs-āh		2
0	sēfr		1

Fig. 1. Arabic digits presentation

3. WAVELET TRANSFORM USING IN SPEAKER FEATURE EXTRACTION

3.1 Wavelet Packet Transform Feature Extraction Method

In order to achieve the utmost results of this work, the speech signal was decomposed into wavelet packet transform (WPT) using the common form of the equivalent low pass discrete time speech signal.

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

where X_m is the representation of the sequence of the discrete speech signal for each value, which is obtained from the data acquisition stage; $p(t)$ is the signal pulse, it represents the importance of the signal design problem whenever there is a bandwidth restriction on the channel of the signal; and T is the sampling time. The processing of the speech is achieved in conjunction with the consideration of $\phi(t - mT)$ as a scaling function of the wavelet packet, i.e., $\phi \in W_{2^N}^0$, the finite set of orthogonal subspaces as defined in [25-27] and can be constructed as:

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

Where $W_{2^N}^0 \subset L^2(R)$, $\rho N = \{(l, n)\}$ is a dyadic interval that is formed as an 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 $\{\psi_n(t)\}_{n \in N}$ so that it may be called the wavelet packet by considering ϕ as the scaling function of the processed signal. Therefore, the speech signal model in equation (1) is suited and customized as:

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

Where, the speech signal model in equation (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. Also, we may use the discrete wavelet packet transforms (DWPT) procedure defined 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 $\phi_l^n \in W_{2^l}^n$. The processes of these two can be carried out recursively by proceeding through the binary tree structure with $O(N \log N)$ computational complexity. Using equations (3), (4), and (5), with the coefficients of the linear combination are decomposed in forward order and 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 the levels having the terminals of a given tree; C_l is the set of indices of the terminals at the l th level and $f_{ln}[i]$ is the equivalent sequence generated from the combination of $h[k]$, $g[k]$ and decimation operation, which leads to form 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 equation (7) is called the constituent terminal function of ϕ_0^1 . For our research paper the tree used consists of two stages, one is three high pass nodes and the other is the three low pass nodes.

The wavelet packet is used to extract and stem additional features in order to guarantee a higher recognition rate. In this work, 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 35582 samples will reach 71166 after WPT decomposition at level two). Thus, we have to seek for a better representation of the speech features. A good survey was conducted, [28] proposed a method to calculate the entropy value of the wavelet norm in digital modulation recognition. In the biomedical field, [27] 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 [29]. Proposed a robust speech recognition scheme in a noisy environment by using wavelet-based energy as a threshold for denoising estimation. In [30], 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 [31] for speaker identification [32,28]. Proposed features extraction method for speaker recognition based on a combination of three entropies types (sure, logarithmic energy and norm). In this paper we use LPC obtained from WP tree nodes for digits feature vector constructing to be used for digits identification.

3.2 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 [33,34]. 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 Mallet's pyramid tree algorithm has been proposed (see Fig. 2) [33].

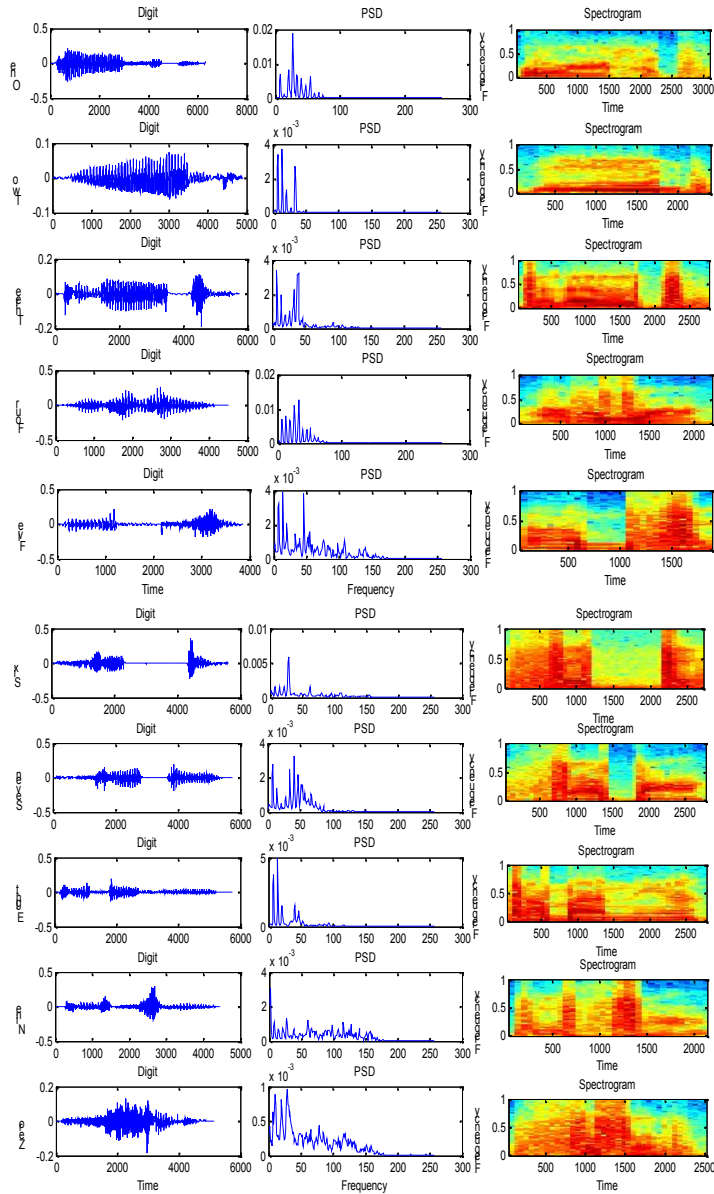


Fig. 2. Arabic digits analysis, by power spectrum density and spectrogram

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 bearing in mind that $m = \log N$ octaves. So the frequency index k varies as 1, 2, ..., m corresponds to the scales $2^1, 2^2, \dots, 2^m$. As described by the Mallet pyramid algorithm Fig. 1. The DWT coefficients of the previous stage are expressed as follows [35]:

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

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

where $W_L(p, j)$ is the p th scaling coefficient at the j th stage, $W_H(p, j)$ is the p th wavelet coefficient at the j th stage, and $h(n)$, $g(n)$ are the dilation coefficients relating to the scaling and wavelet functions, respectively. WT was proposed for recognition by [31]. In [36] and [32], 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 [35]. [37,38] 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 digits feature vector constructing to be used for text-independent digits recognition.

3.3 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. To achieve this, the signals are normalized by using the following formula [39]:

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

Where S_i is the i th element of the signal S , \bar{S} and σ are the mean and standard deviation of the vector S , respectively, and S_{Ni} is the i th element of the signal series S_N after normalization.

The LPC method is not a new technique for modeling of speech vocal tracing parameters. It was developed in the 1960s by [40] and still used in the recent papers for speech vocal tracing. The reason behind that is the representing a speaker by modeling vocal tract parameters and the data size are very suitable and well fit for speech compression throughout the digital channel [39]. In this paper, a modified LPC coefficients approach is

suggested for reducing the size of feature vectors. The proposed wavelet averaging framing LPC (AFLPC) extracts the features from Z frames of each WT speech sub signal:

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

Where Z is the number of considered frames (each frame of 20 ms duration) for the q th WT sub signal $u_q(t)$. The average of LPC coefficients calculated for Z frames of $u_q(t)$ is utilized to extract 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 signals are represented as:

$$AFPC = \{afpc_1, afpc_2, \dots, afpc_Q\} \tag{12}$$

In this paper we use AFLPC taken from WP at level two (AFWP) and taken from DWT at level 5 (AFDWT).

4. CLASSIFICATION

4.1 Proposed Probabilistic Neural Networks Algorithm

Ganchev [39-43] proposed PNN with Mel-frequency cepstral coefficients for text-independence. Although there are numerous enhanced versions of the original PNN presented by many researchers, which are either more economical or exhibit an appreciably better performance for simplicity of exposition, we adopted and invoked the original PNN for classification task (see Fig. 3). The proposed algorithm is denoted by *PNN* and depends on the following construction:

$$Net = PNN(X, P, SP),$$

Where X is a 180×24 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:

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

P is the target class vector

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

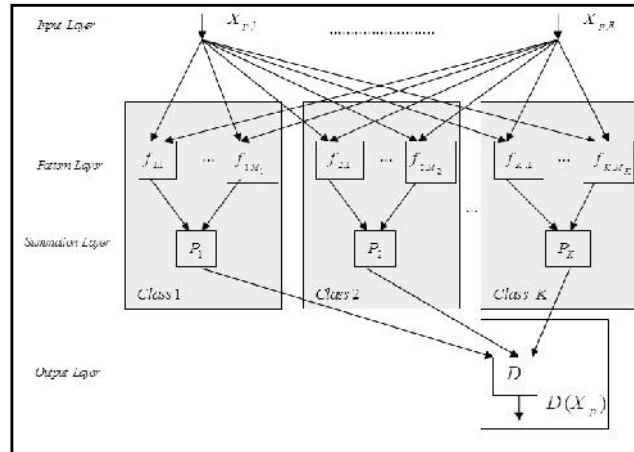


Fig. 3. Structure of the original probabilistic neural network

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.

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

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

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

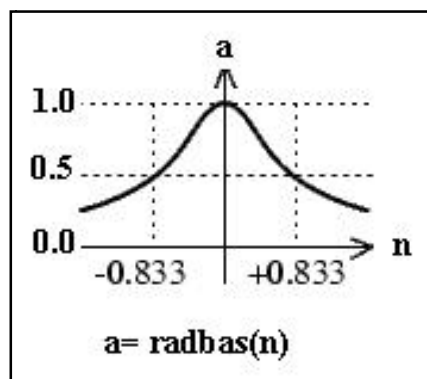


Fig. 4. Radial basis transfer function

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 is a 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^i , 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 $\pm SP$. The second layer weights are set to P .

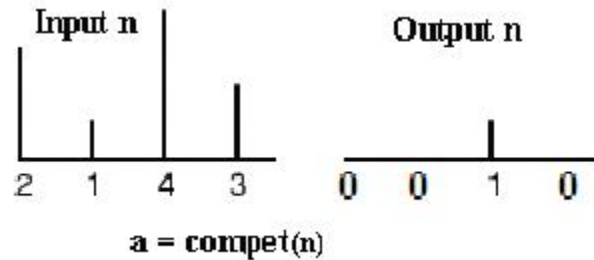


Fig. 5. Competitive transfer function

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

5. EXPERIMENTAL RESULTS AND DISCUSSION

Speech signals were recorded via PC-sound card, with sampling frequency of 8000 Hz. The Arabic digits from zero to nine were recorded, 3 times, in three Arabic dialects: Egyptian Arabic Dialect (EAD), Jordanian Arabic Dialect (JAD) and Palestinian Arabic Dialect PAD (tabulated in Tab. 2); 3 females, (aged from 20 to 30years), along with 11 males, (aged from 20 to 50 years) participated in speech digits recording. The recording was done in university office environment.

Our study of Dialect-independent classification system performance for robustness to noise is conducted by presenting some experiments depending on several considered aspects. The real noise presented by the restaurant noise will also be investigated. All experiments as well as comparison investigation will be performed in a common environment. Same testing and training signals will be utilized in case of comparison performing.

Experimental-1

Table. 3 shows the results of Recognition Rate for dialect-independent digits database of the recorded spoken Arabic digits. We tested 450 different signals for different individuals. The given experiment provides a study of classification performance via AFLDWT and AFLWP average framing wavelet transform (WP and DWT) fusion (AFLWTF). The results taken from 450 signals showed the superiority of DWT based method. The results, tabulated in Tab.4, showed the Recognition Rate for every individual spoken digit. Table 3, also showed the mixed recognition digit for certain cases. Confused similarity is seen for Arabic digit 3, where was mixed with digits 1, 7 and 8. So that, less Recognition Rate was obtained for digit 3. Large mixed recognition between digits 7 and 3 found because of the similarities between

these two digits as shown in Fig. 2. In terms of Recognition Rate in dialect-independent system according to our implementation results, Alotaibi classifier [7] has considerable results, about 80%, less than proposed classifier.

Table 3. Recognition Rate of individual Arabic digits and mixed recognized digits

Digit	0	1	2	3	4	5	6	7	8	9	
0	47										
1		49		7	2	1					
2		1	48		1						
3			2	40					3		
4	2				45			6			
5						47					
6							47				
7				1			3	44			
8	1			2		2			47		
9					2					50	
Rec. Rate [%]	94	98	96	80	90	94	94	88	94	100	Avr. 92.8 [%]

Experimental-2

In experiment-2, the system performance for noisy environment is examined using white Gaussian noise (WGN). WGN was added to the digits speech signals, the probability density function of WGN is defined as follows:

$$f_X(x) = \frac{1}{\sigma_X \sqrt{2\pi}} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$$

Where μ is mean value, σ is standard deviation and X is the random variable. The results were obtained at different SNR levels Table 4. The results proved the effect of such utilization for robustness to noise.

Table 4. Recognition rate of Arabic digits (0, 1, 2, 3 and 4) and mixed recognized digits with WGN

SNR[dB]	0	1	2	3	4
-5	0	0	0	3	3
0	0	3	0	3	3
5	0	1	0	3	3
10	0	1	2	3	4
15	0	1	2	3	4

Experimental-3

In experiment-3 a comparison between the AFDWT and AFWP methods is performed. The aspect of comparison is based on adding White Gaussian Noise to five Arabic digits signals in different magnitude as shown in Table 5. The two methods are implemented to the same signals via several SNR. The aim of this implementation is to investigate the robustness to

noise degree. We conclude according to the results contained in Table 5 that AFDWT has superior performance.

Table 5. Comparison between AFDWT and AFWP in term of real restaurant noise

SNR	2		6		7		8		9	
	AFDWT	AFWP	AFDWT	AFWP	AFDWT	AFWP	AFDWT	AFWP	AFDWT	AFWP
-4.5718	6	0	3	3	0	0	0	0	9	0
-2.2667	6	0	8	3	0	0	0	1	9	0
0.0372	6	0	3	8	0	0	0	3	9	9
2.2626	6	0	8	3	0	0	0	0	9	9
4.6439	2	0	6	6	7	0	8	2	9	9
7.1312	2	2	6	6	7	0	8	8	9	9
9.3465	2	2	6	6	7	7	8	8	9	9
11.4867	2	2	6	6	7	7	8	8	9	9
14.2551	2	2	6	6	7	7	8	8	9	9
16.0557	2	2	6	6	7	7	8	8	9	9
20.0322	2	2	6	6	7	7	8	8	9	9
23.4214	2	2	6	6	7	7	8	8	9	9

In the next experiment two conventional methods were investigated for comparison: Fast Fourier transform and feed forward back propagation (FFTNN) [34], time-frequency and feed forward back propagation (TFNN) [44], were investigated for comparison. The proposed method achieved superior performance as tabulated in Table 6.

Table 6. Comparison between several methods in terms of classification rate

Method	Classification Rate
AFDWT	92.8%
FFNN	71.91%
TNN	70.02%

6. CONCLUSION

The wavelet based features extraction method has been proposed. This approach faces the dialect-independent and speaker-independent difficulties. Probabilistic neural network is utilized in the classification part of the proposed classifier. The proposed classifier performed high recognition rate of up to 100% in some cases, with an average rate reaching up to 93%, for 450 tested digits signals. The method performance was tested in noisy environment by adding WGN as well as restaurant. The comparison between the WP and DWT was investigated. We concluded according to the results contained in Table 5 that DWT has superior performance. In terms of Recognition Rate in dialect-independent system according to our implementation results, Alotaibi classifier has considerable results- about 80% less than proposed classifier. A comparison with conventional methods was studied. The reason of this success is the utilization of the sophisticated extraction based on Wavelet Transform in conjunction with LPC. DWT has overcome the WPT in terms of recognition rate. The wavelet transform utilization in the feature extraction procedure could enhance the results significantly overcoming the conventional methods. The reason behind that is the possibility of extracting features through the different wavelet levels.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Douglas O. Interacting with computers by voice Automatic speech recognition and synthesis. *Proceeding of the IEEE*. 2003;141–159.
2. Datta S, Zabibi MA and Farook O. Exploitation of morphological in large vocabulary arabic speech recognition. *International Journal of Computer Processing of Oriental Language*. 2005;18:291–302.
3. Kirschho K. Novel approach to arabic speech recognition. final report from the jhu summer school workshop. In *Proceedings of the International Conference on ASSP*. 2002;344–347.
4. Selouani SA, Caelen J. "Recognition of arabic phonetic features using neural networks and knowledge-based system: a comparative study," *International Journal of Artificial Intelligence Tools*. 1999;73–103.
5. Debyeche M, Haton JP, Houacine A. A new vector quantization approach for discrete hmm speech recognition system. *International Scientific Journal of Computing*. 2006;72–78.
6. Shoaib M, Awais M, Masud S, Shmail S, Akhbar J. Application of concurrent generalized regression neural networks for arabic speech recognition. In *Proceedings of the IASTED International Conference on Neural Networks and Computational Intelligence*. 2004;206–210.
7. Alotaibi YA. Investigating spoken arabic digits in speech recognition setting. *Information Sciences*. 2005;173:115–139.
8. Amrouche A, Debyeche M, T-Ahmed A, Rouvaen M, Yagoub MCE. An efficient speech recognition system in adverse conditions using the nonparametric regression. *Engineering Applications of Artificial Intelligence*. 2009;85–94.
9. Bourouba H, Djemili R, Bedda M, Snani C. New hybrid system (supervised classifier/HMM) for isolated arabic speech recognition. In *Proceedings of the Second IEEE International Conference on Information and Communication Technologies*. 2006;1264–1269.
10. Saeed KM, Nammous Information Processing and Security Systems. A New Step in Arabic Speech Identification: Spoken Digit Recognition. 2005;55–66.
11. Lazli L, Sellami M. Connectionist probability estimation in hmm arabic speech recognition using fuzzy logic. *Lectures Notes in LNCS*, 2003;379–388.
12. Selouani SA, Douglas O. Hybrid architectures for complex phonetic features classification a unified approach. In *International Symposium on Signal Processing and its Applications, (Kuala Lumpur, Malaysia)*. 2001;719–722.
13. Salam M, Mohamad D and Salleh S. Neural network speaker dependent isolated malay speech recognition system: handcrafted vs. genetic algorithm. In *International Symposium on Signal Processing and its Application, (Kuala Lumpur, Malaysia)*. 2001;731–734.
14. Saeed K, Nammous M. Heuristic method of arabic speech recognition. in *Proceedings of the IEEE International Conference on Digital Signal Processing and its Applications*. 2005;528–530.
15. Amrouche A, Rouvaen JM. Arabic isolated word recognition using general regression neural network," in *Proceedings of the 46th IEEE MWSCAS*. 2003;689–692.

16. Daqrouq K and Al-Qawasmi A. "The study of wavelet filters speech enhancement method". Third Mosharaka International Conference on communications, Computers and Applications. 2009;26–28.
17. Karray L, Martin A. Towards improving speech detection robustness for speech recognition in adverse conditions. *Speech Communication*. 2003;261–276.
18. Savoji MH. A robust algorithm for accurate end pointing of speech signals. *Speech Communication*. 1989;8:45–60.
19. Junqua JC, Mak B, Reaves B. A robust algorithm for word boundary detection in the presence of noise. *IEEE Transactions Speech Audio Process*. 1994;2(3):406–412.
20. Berouti M, Schwartz R, Makhou IJ. Enhancement of speech corrupted by acoustic noise. *International Conference on Acoustics, Speech, and Signal Processing*. 1979;208–211.
21. Mokbel C, Jouvét D, Monne J. Blind equalization using adaptive filtering for improving speech recognition over telephone. In *European Conference on Speech Communication and Technology*. 1995;141–1990.
22. Hermansky H, Morgan N, Hirsch HG. Recognition of speech in additive and convolutional noise based on RASTA spectral processing. *International Conference on Acoustics, Speech, and Signal Processing*. 1993;83–86.
23. Mashinchi MR, Mashinchi MH, Selamat A. New approach for language identification based on DNA computing. *International Conference on Bioinformatics & Computational*. 2007;748–752.
24. Zitouni I, Sarikaya R. Arabic diacritic restoration approach based on maximum entropy models. *Computer Speech and Language*. 2009;23:257–276.
25. Daqrouq K and Abu-Sheikha NM. Heart rate variability analysis using wavelet transform. *Asian Journal for Information Technology*; 2005.
26. 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, *Circuits Systems Signal Processing*. 2005;24(3)287–302.
27. Behroozmand R, Almasganj F. Optimal selection of wavelet packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis. *Computers in Biology and Medicine*. 2007;37:474–485.
28. Avci E, Hanbay D, Varol A. An expert discrete wavelet adaptive network based fuzzy inference system for digital modulation recognition. *Expert System with Applications*. 2006;33:582–589.
29. Sarikaya R and Hansen JHL. High resolution speech feature parametrization for monophone-based stressed speech recognition. *IEEE Signal Processing Letters*. 2000;7(7):182–185.
30. Wu JD, Lin BF. Speaker identification using discrete wavelet packet transform technique with irregular decomposition, *Expert Systems with Applications*. 2009;36:3136–3143.
31. Avci D. An expert system for speaker identification using adaptive wavelet sure entropy, *Expert Systems with Applications*. 2009;36:6295–6300.
32. Avci E. A new optimum feature extraction and classification method for speaker recognition: GWPNN, *Expert Systems with Applications*. 2007;32:485–498
33. Mallat S. A Theory for Multiresolution Signal Decomposition: the Wavelet Representation'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1989;11:674–693.
34. HuaQuan Z, Shuang Huang D, Lei Xia X, Lyu M, Lok TM. Spectrum analysis based on windows with variable widths for online signature verification, in *18th International Conference on Pattern Recognition, ICPR*. 2006;2:2006.

35. Wu JD and Lin BF. Speaker identification based on the frame linear predictive coding, *Expert Systems with Applications*. 2009;36:8056–8063.
36. Tufekci, Gowdy, Tufekci Z, Gowdy J. Feature extraction using discrete wavelet transform for speech recognition In *Proc. South East Con.* 2000;116–123.
37. Daqrouq K. Wavelet entropy and neural network for text-independent speaker identification *Engineering Applications of Artificial Intelligence*. 2011;24:796–802.
38. Daqrouq K, Abu Sbeih I, Daoud O, Khalaf E. An investigation of speech enhancement using wavelet filtering method, *International Journal of Speech Technology*. 2010;13:(2)101-115.
39. Daqrouq K, Azzawi KYAI. Average framing linear prediction coding with wavelet transform for textindependent speaker identification system. *Comput Electr Eng*. Available: <http://dx.doi.org/10.1016/j.compeleceng.2012.04.014>, (Elsevier)
40. Ganchev T, Fakotakis N, Kokkinakis G. Comparative evaluation of various MFCC implementations on the peaker verification task, in: *Proceedings of the Specom*. 2005;1:191–194.
41. Bennani Y, Gallinari P. Neural networks for discrimination and modelization of speakers, *Speech Communication*. 1995;17:159-175.
42. Adami AG, Barone DAC. A speaker identification system using a model of artificial neural networks for an elevator application. *Information Sciences*. 2001;138:1–5.
43. Haydar A, Demirekler M and Yurtseven MK. Speaker identification through use of features selected using genetic algorithm. *Electronics Letters*. 1998;34:39–40.
44. Stridth M, Sornmo L. Shape characterization of atrial fibrillation using time time-frequency analysis. *Comput Cardiol*. 2002;29.

© 2014 Daqrouq et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:

The peer review history for this paper can be accessed here:
<http://www.sciencedomain.org/review-history.php?iid=408&id=5&aid=3383>