SPEECH RECOGNITION OF ARABIC WORDS USING ARTIFICIAL NEURAL NETWORKS

Dr. Sadiq Jassim Abou-Loukh
University of Baghdad - College of Engineering - Electrical Eng. Dept.
doctor_sadiq@yahoo.com

ABSTRUCT:
The speech recognition system has been widely used by many researchers using different methods to fulfill a fast and accurate system. Speech signal recognition is a typical classification problem, which generally includes two main parts: feature extraction and classification. In this paper, a new approach to achieve speech recognition task is proposed by using transformation techniques for feature extraction methods; namely, slantlet transform (SLT), discrete wavelet transforms (DWT) type Daubechies Db1 and Db4. Furthermore, a modified artificial neural network (ANN) with dynamic time warping (DTW) algorithm is developed to train a speech recognition system to be used for classification and recognition purposes. Twenty three Arabic words were recorded fifteen different times in a studio by one speaker to form a database. The performance of the proposed system using this database has been evaluated by computer simulation using MATLAB package. The result shows recognition accuracy of 65%, 70% and 80% using DWT (Db1), DWT (Db4) and SLT respectively.

KEYWORDS: Speech Recognition, Discrete Wavelet Transform, Slantlet Transform Dynamic Time Warping, Artificial Neural Network

تمييز الكلمات العربية باستخدام الشبكة العصبية الاصطناعية

د. صادق جاسم ابو اللوخ
جامعة بغداد – كلية الهندسة – قسم الهندسة الكهربائية

الخلاصة:
استعمل نظام تمييز الكلام بصورة واسعة بوساطة عدد من الباحثين باستخدام طرائق مختلفة لتحقيق نظام تمييز سريع ودقيق. ان تمييز اشارة الكلام تعد مشكلة تصنيف نوعية وهي تضم بصورة عامة جزئين اساسيين: استخلاص الميزات والتصنيف. تضمن هذا العمل اقتراح ثلاثة طرق لاستخلاص الخصائص وهي تحويل المويجي المتقطع (DWT) بنوعيه Db4 and Db1 وتحويل المويل (SLT). تم تطوير نظام يعتمد على استخدام الشبكات العصبية الاصطناعية مع طريقة ميلان الزمن الديناميكي لغرض التمييز. ثلاثة وعشرون كلمة عربية بخمسة عشر ازمان مختلفة مسجلة في الاستوديو بوساطة متكلم واحد لتشكيل قاعدة بيانات. اداء النظام المقترح تم عن طريق تمثيل قاعدة البيانات باستخدام حقيبة الـ MATLAB . بينت النتائج ان دقة التمييز هي (٦٥%، ٧٠% و٨٠%) باستخدام (DWT Db1, DWT Db4 and SLT) على التوالي.

1. INTRODUCTION

Speech recognition is currently used in many real time applications, such as cellular telephones, computers and security systems. Speech signals are composed of a sequence of sounds, these sounds and the transitions between them serve as a symbolic representation of information. The arrangement of these sounds (symbols) is governed by the rules of language. An alternative way of characterizing speech is in terms of the signal carrying the message information, i.e., the acoustic waveform [1].

Speech is one of the most important tools for communication between human and his environment, therefore manufacturing of automatic system recognition (ASR) is a popular and challenging area in developing human computer interaction. The task of a speech recognizer is to determine automatically the spoken words, regardless of the variability introduced by the speaker identity, manner of speaking, and environmental conditions [2]. The design of the speech recognition system requires careful attention to the following issues; speech preprocessing, feature extraction techniques, speech recognition, and performance evaluation.

As will be illustrated in the following subsections, the preprocessing stage includes three different processes, namely; sampling, framing, and windowing. While in feature extraction phase, the speech signal is converted into a stream of feature vectors coefficient which contain only that information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of information irrelevant for correct classification. Currently, most speech recognition systems are based on dynamic time warping (DTW), hidden Markov model (HMM) and artificial neural network (ANN) [3,4].

In this paper a slantlet transform (SLT) based approach and discrete wavelet transform (DWT) type Daubechies (Db1) and (Db4) were used to extract the features from the speech signal. In speech recognition dynamic time warping (DTW) is often used to determine if two waveforms represent the same spoken phrase. A modified ANN based on the DTW algorithm is developed to train the system and to be used as a classifier for identifying the spoken word.

The rest of the paper is organized as follows. In the next two sections, we introduce two types of transforms implemented in this work. An overview of the proposed speech recognition system applied to the Arabic words database is presented in section 4. Section 5 presents the performance of the system on Arabic words and attains the experimental recognition results. Finally, section 6 gives a summary and conclusions of the work.

2. DISCRETE WAVELET TRANSFORM

The discrete wavelet transform (DWT) has proved to be a useful tool for the analysis of non-stationary signals like speech. The DWT employs two sets of functions, called scaling and wavelet functions, which are associated with low-pass and high-pass filters, respectively. It is defined by the following equation [5].

$$g(t) = \sum_{k} c_{j0}(k) 2^{j0/2} \varphi(2^{j0} t - k) + \sum_{k} \sum_{j=j0}^{\infty} d_{j}(k) 2^{j/2} \psi(2^{j} t - k) \qquad (1)$$

Where: $\varphi(t)$'s is the scaling functions , $\psi(t)$'s is the wavelet functions, k is the time translation index and j is the scale parameter.

The DWT is a linear transformation that operates on a data vector whose length is an integer power of two, transforming into a numerically different vector of the same length. It is a tool that separates data into different frequency components, and then studies each component with a resolution matched to its scale. DWT is computed with a cascade of low and high pass filtering followed by a factor 2 down sampling.

The DWT analyzes the signal at different frequency bands with different resolutions by decomposing the signal into a coarse approximation and detail information. The decomposition of the signal into different frequency bands is simply obtained by successive high-pass and low-pass filtering of the time domain signal. The original signal x [n] is first passed through a half band high-pass filter g [n] and a low-pass filter h [n] as shown in Fig.1 [5].

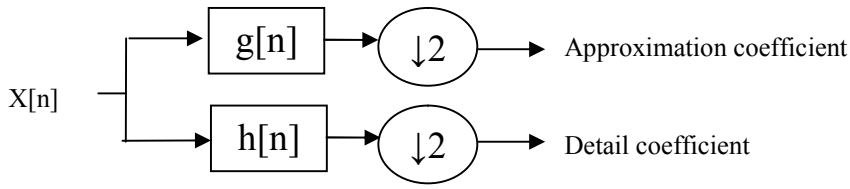

Fig.1: Approximation and detail coefficients using DWT

After each filtering, half of the samples/data can be segregated (or eliminated) as per the Nyquist's rule. Since the high-pass filtered signal has now the highest frequency of π/2 radians instead of π, the signal can therefore be down sampled by 2. This constitutes one level of decomposition. Output from the high-pass filter is downloaded as the level 1 detail D1, and the output from the low- pass filter becomes the level 1 approximation, A1. Starting afresh with the A1, the process can be successively repeated as per requirements. This filtering and eliminating can be repeated on the scaling coefficients to give the two scale structure. Repeating this on the scaling coefficients is called iterating the filter bank. Iterating the filter bank again gives the three scale structure as shown in Fig.2 [5].

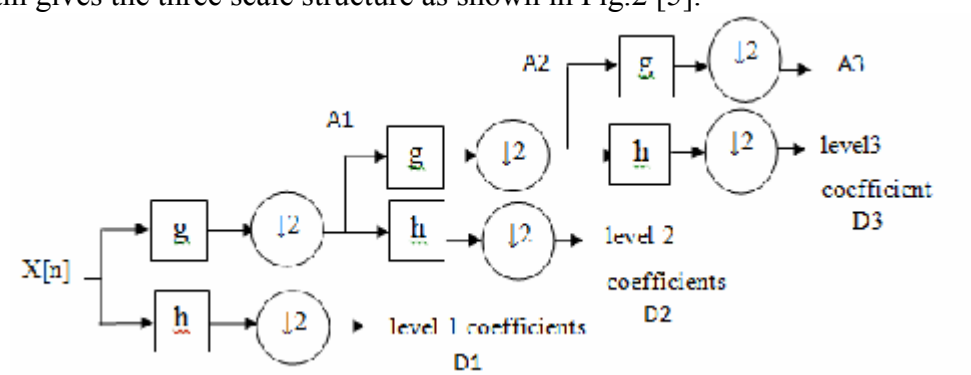

Fig. 2: Three scale filter bank analysis

3. SLANTLET TRANSFORM

The slantlet transform (SLT) is based on an improved version of the usual DWT, where the support of the discrete-time basis functions is reduced [6]. The SLT is an orthogonal DWT with two zero moments and with improved time localization, the basis of the slantlet is based on a filter bank structure where different filters are used for each scale. Consider a usual two-scale iterated DWT filter bank shown in Fig. 3 (a) and its equivalent form in Fig. 3 (b). The slantlet filter bank is based on the structure of the equivalent form shown in Fig. 3 (b), but it is occupied by different filters that are not products. With this extra degree of freedom obtained by giving up the product form, filters of shorter length are designed to satisfy the orthogonality and
zero moment conditions [6]. For two-channel case the Daubechies filter is the shortest filter which makes the filter bank orthogonal and has K zero moments. For K=2 zero moments the iterated filters of Fig. 3 (b) are of lengths 10 and 4 but the slantlet filter bank with K=2 zero moments shown in Fig. 3 (c) has filter lengths 8 and 4. Thus the two-scale slantlet filter bank has a filter length which is two samples less than that of a two-scale iterated Daubechies filter bank. This difference grows with the increased number of stages. The slantlet filters are piecewise linear. Even though there is no tree structure for slantlet it can be efficiently implemented like an iterated DWT filter bank. Therefore, computational complexities of the slantlet are of the same order as that of the DWT, but slantlet transform gives better performance.

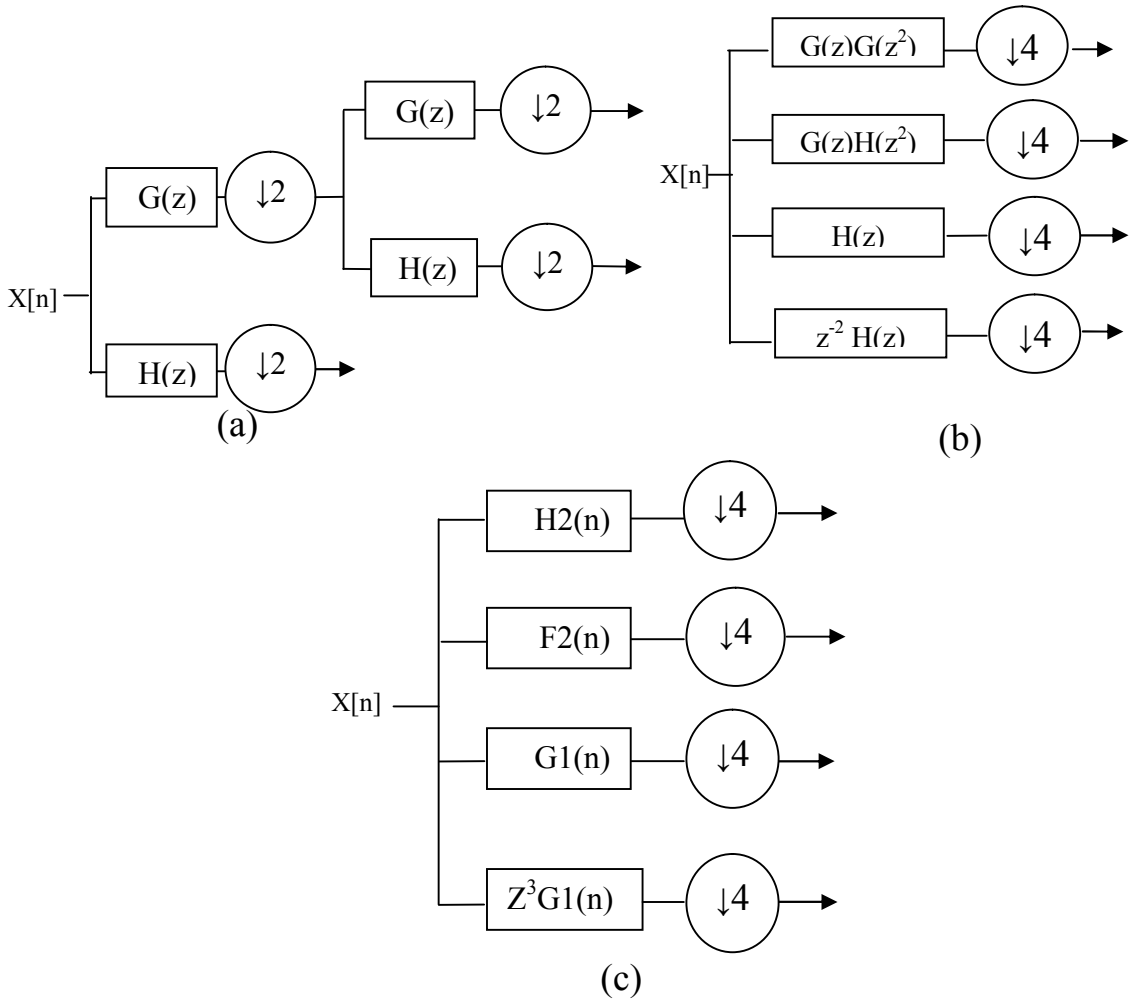

Fig. 3: (a) Two-scale iterated filter bank DWT
(b) Equivalent form using the DWT
(c) Two-scale filter bank using SLT

4. PROPOSED SPEECH RECOGNITION MODEL

The block diagram for the proposed model of speech recognition system is shown in Fig. 4. The architecture of the proposed speech recognition process contains six main stages:

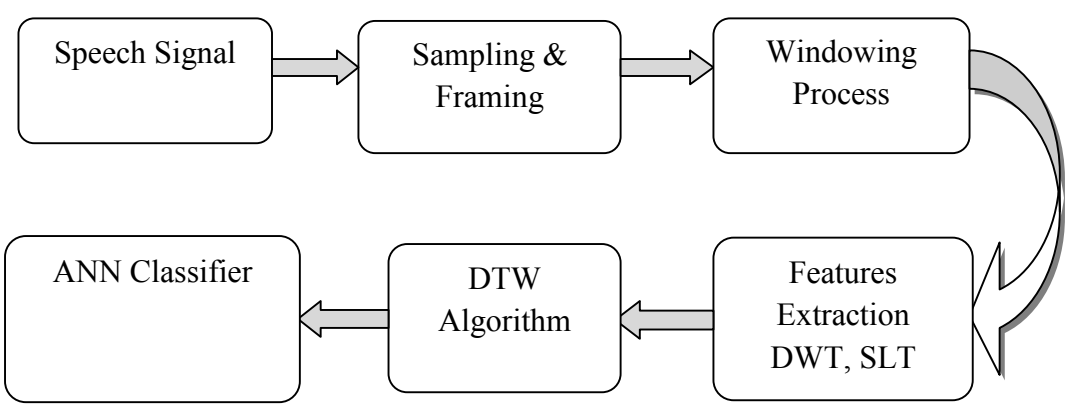

Fig. 4: Proposed model of the speech recognition system

4.1 Speech Signal

The speech signal was recorded in a nearly noise free environment by microphone in a studio, this database consists of twenty three Arabic words. These Arabic words are spoken by one speaker, and this speaker utters each word by different fifteen versions. It is clear that the length of the words is different, also each version of the same word varies in length too. The total number of words in the database is 345 utterances for one speaker and this database was used for training and evaluation of the proposed algorithm. The twenty three Arabic words are shown in Table 1.

4.2 Sampling and Framing

The speech signals are sampled to convert it from analogue to digital signal. Since we deal with speech signal, which is non-stationary signal (vary with time), the framing process is essential to deal with frames not with whole signal. After this stage the speech signal has many frames and the number of frames depends on the number of samples for each word. The number of samples for each frame is 256 samples.

4.3 Hamming Windowing

Each frame of the word was multiplied by the Hamming window; the advantage of this multiplication is to minimize the signal discontinuities at the beginning and the end of each frame.

Table 1: The words and their corresponding number of versions

Word number	Word name	The total number Of versions used	Base dataset	Test dataset
1	اشارة	15	10	5
2	لندن	15	10	5
3	افتح	15	10	5
4	الخير	15	10	5
5	تصميم	15	10	5
6	ثقيل	15	10	5
7	خاص	15	10	5
8	دوران	15	10	5
9	رازق	15	10	5
10	رحمن	15	10	5
11	زيارة	15	10	5
12	صباح	15	10	5
13	صديق	15	10	5
14	عمودي	15	10	5
15	كامل	15	10	5
16	محمد	15	10	5
17	معلومات	15	10	5
18	نظام	15	10	5
19	وفاء	15	10	5
20	ياسين	15	10	5
21	يمين	15	10	5
22	مساء	15	10	5
23	زهرة	15	10	5

4.4 Feature Extraction

The goal of feature extraction is to represent any speech signal by a finite number of measures (or features of the signal). This is because the integrity of the information in the acoustic signal is too much to process and not all of the information is relevant for specific tasks. A typical feature extraction algorithm tends to build a computational model through some linear or nonlinear transform of the data so that the extracted features are as representative as possible. Two methods, namely, DWT and SLT were used for feature extraction as explained in the following sections.

4.4.1 DWT Coefficients Extraction

The DWT was applied to each frame. In this work 1 to 8 DWT level were applied. The 3-level wavelet decomposition structure is shown in Fig. 2, where the result is four different subsets, three subsets for the details (the wavelet function coefficients) and the subset four is the approximation subset (the scaling function coefficients).

Most of the energy of the speech signal lies in the lower frequency bands, the other sub-bands contain most detail information of the signal and they are discarded, since the frequency band covered by these levels contains much noise and is less necessary for representing the approximate shape of the speech signal. Hence take the approximation coefficients and discard the detail coefficients. In this work the DWT Daubechies (Db1) and (Db4) were used.

4.4.2 SLT Coefficients Extraction

The slantlet filter bank used to extract the features of the speech signal is 3-scale (L=3) filter bank. The structure of this filter bank is shown in Fig.5, where 6 different filters are used. These filters are constructed using the derivations explained previously. The low pass filter h3 (n) output is the approximation of the signal and the other outputs are the details so it can be efficient to take only the coefficients of the low pass filter as features of the signal and discard the remaining coefficients without losing much information about the signal. Since the slantlet transform gives better time localization the results of the classification will be better using this transformation method.

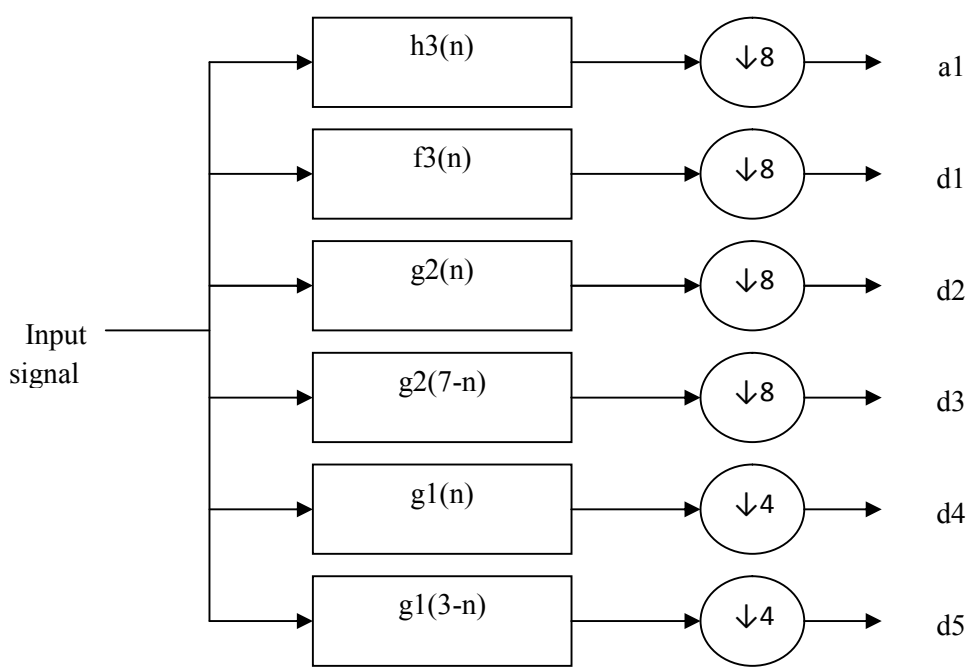

Fig. 5: Three scale slantlet filter bank

Where:

- a1: is 32 samples which are the outputs of the low pass filter h3 after down sampling of 8. Only this vector will be used as the features of the speech signal.
- d1: is 32 samples which are the outputs of the band pass filter f3 after down sampling of 8.
- d2: is 32 samples which are the outputs of the band pass filter g2 after down sampling of 8.
- d3: is 32 samples which are the outputs of the band pass filter (shifted time reverse of g2) after down sampling of 8.
- d4: is 64 samples which are the outputs of the band pass filter g1 after down sampling of 4.
- d5: is 64 samples which are the outputs of the band pass filter (shifted time reverse of g1) after down sampling of 4.

4.5 DYNAMIC TIME WARPING ALGORITHM

Dynamic time warping (DTW) is a technique that finds the optimal alignment between two time series. In speech recognizes the need for DTW arises because repetitions of the same words might have different duration intervals. The DTW algorithm which is based on dynamic programming finds on optimal match between two sequences of feature vectors by allowing for stretching and compression of sections of sequence [7]. Therefore, after feature extraction phase, each word version represented by one feature vector, these feature vectors are different in length. The lengths of all versions of the same word are not identical and this problem was solved by using the DTW algorithm which equalizes the versions length of the same word.

The procedure of the DTW algorithm for equalizing all the versions is explained by the following points [8]:

a.Take fifteen versions of the first word (اشارة), these versions is (v1, v2, v3, v4, v5, v6, v7, v8, v9, v10, v11, v12, v13, v14, v15).

b. Choose the reference version by the following procedure:

1. Find the number of frames for each version

F1=number of frames of v1=224
F2=number of frames of v2=192
F3=number of frames of v3=208
F4=number of frames of v4=232
F5=number of frames of v5=256
F6=number of frames of v6=184
F7=number of frames of v7=224
F8=number of frames of v8=192
F9=number of frames of v9=232
F10=number of frames of v10=232
F11=number of frames of v11=224
F12=number of frames of v12=192
F13=number of frames of v13=240
F14=number of frames of v14=192
F15=number of frames of v15=200

2. Take the average value of numbers of these frames

$$AV = \frac{(F1 + F2 + F3 + ... + F15)}{15} = \frac{(224 + 192 + 208 + ... + 200)}{15} = 214.9333$$

AV must be an integer, therefore AV=214, and if AV is increased by 1, the average value becomes A=AV+1=215.

3. Compare the value of A with the number of frames for each version, and take the version which has number of frames equal or nearest to the value of A. See here F3 is nearest to A; therefore the third version is the reference version.

c. The reference version (v3) has 208 frames and each frame consists of 256 samples, and the other fourteen versions must have the same length of reference version, i.e., all fifteen versions must have 208 frames and each frame has 256 samples, therefore each version is equalized with reference version, thus equalizing is done by applying the DTW algorithm.

4.6 ARTIFIAL NEURAL NETWORK

In the present work, the artificial neural network (ANN) is used for the recognition purposes. The ANN derives their power due to their massively parallel structure, and the ability to learn from experience. They can be used for fairly accurate classification of input data into categories, provided they are previously trained to do so [9]. The knowledge gained from the learning experience is stored in the form of connection weights, which are used to make decisions on fresh input.

The ANN used as a classifier uses the features obtained from the feature extraction process for training and testing. The ANN structure used in the proposed work is shown in Fig. 6, and its specifications are illustrated in Table 2 [8]. The activation function for the input layer and the hidden layer neurons is the Tan-sigmoid function and the activation function for the output layer neuron is the linear function.

The neural network after identifying its parameters [identifying the number of layers and the activation functions] was trained using the back propagation algorithm (BPA). The BPA is a supervised learning algorithm, in which a mean square error function is defined, and the learning process aims to reduce the overall system error to a minimum. After the training process of the neural network, the testing process was done to test the performance of the neural network in classifying the input patterns. The maximum number of iterations for the neural network used in this work is 1000.

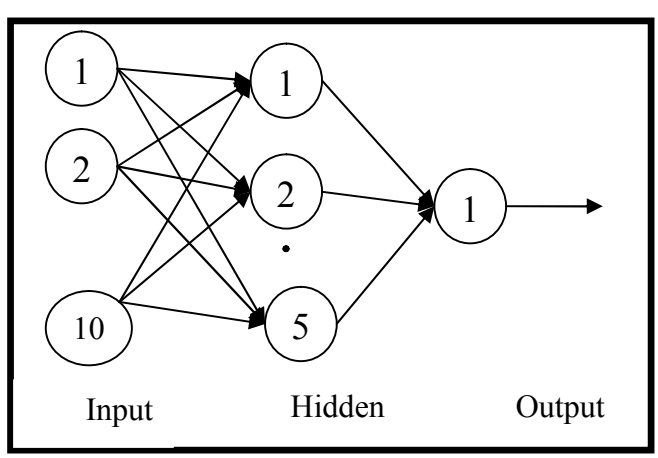

Fig.6: Structure of ANN

Table 2: Neural network specifications

No.	Item	Value
1	Training pattern vector number	230
2	Layer number	3
3	No. of neurons in the input layer	10
4	No. of neurons in the hidden layer	5
5	No. of neurons in the output layer	1
6	Goal MSE	0
7	Maximum number of iterations	1000

5. RECOGNITION RESULTS

In this work, the proposed system has been applied on twenty three Arabic words. These words and their corresponding number of versions are shown in Table 1.The number of versions of each word has been divided into two parts:

a. First part of these versions are used for the base of the ANN, called "base versions", ten versions have been taken for each word.

b. The other part is used for the test of the ANN called "test versions", five versions have been taken for each word, the test versions are tested by the ANN and their resultant error is used to give the measure of the generalization ability of this network.

In this work three different transforms were used to extract the feature from the speech signal namely; DWT Db1, DWT Db4 and SLT. The need for DTW algorithm arises because the repetition of the same word might have different duration intervals. The ANN based on the DTW algorithm was used for classification and recognition purposes. To compare the performance of these transforms, the recognition rate or accuracy of each one has been computed by the following equation:

$$\text{Accuracy} = \frac{\text{Total number of correct recognition}}{\text{Total number of testing versions}} * 100\% \qquad (2)$$

Since speech recognition consists of several stages, the most significant ones are the feature extraction and recognition stages. Therefore, two feature extraction techniques are evaluated. The first proposed model based on DWT type Db1 and Db4 was used, while the SLT was used in the second model as explained previously. After that all feature vectors for each version is applied to the DTW to equalize the length of these versions with reference versions, finally these equalizing versions enter into a neural network. The ANN structure used is shown in figure (6) and its specifications are illustrated in Table (2). The proposed models were examined using computer simulation with MATLAB package. The accuracy of recognition for the different models used is calculated as below.

The results of recognition using DWT (Db1) are as follows:

A total number of testing error is equal to 35

The percentage of error (e) is calculated by

$$e = \frac{35}{100} * 100 = 35\%$$

Therefore, the recognition rate =100 – 35= 65%

The results of recognition using DWT (Db4) are as follows:
A total number of testing error is equal to 30
The percentage of error (e) is calculated by

$$e = \frac{30}{100} * 100 = 30\%$$

Therefore, the recognition rate =100-30=70%

The results of the recognition using SLT are as follows:
A total number of testing error is equal to 20
The percentage of error e is calculated by

$$e = \frac{20}{100} * 100 = 20\%$$

Therefore, the recognition rate =100 – 20= 80%

Figure (7) Show the comparison between the accuracy of recognition of the proposed systems using SLT and DWT type (Db1) and (Db4).

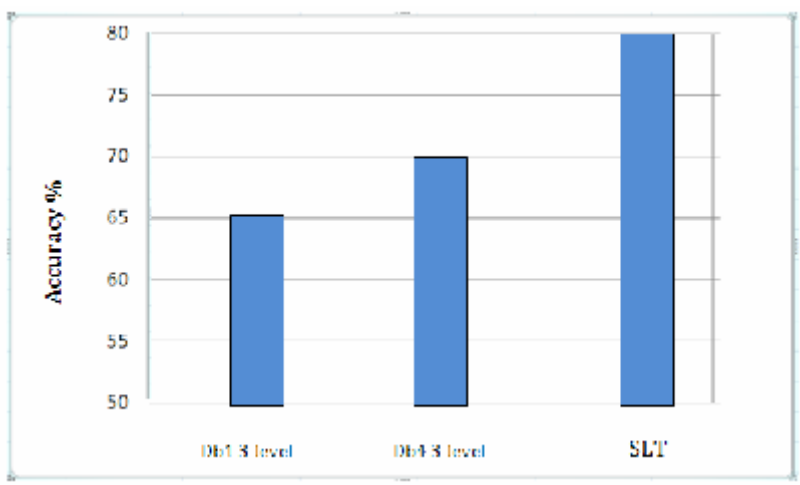

Fig. 7: Comparison of the recognition accuracy of the proposed systems

6. CONCLUSIONS

Currently, development of speech recognition is widely used in the industrial software market. In this paper, a new method is developed to recognize the Arabic words based on ANN with DTW. Three different types of transformation methods were used to extract features from the speech signal. The recognition system used these features as the input to the ANN classifier.

The results show that the DWT type Db1 and Db4 have nearly the same accuracy of Arabic word recognition. The number of scales in SLT depends on the number of samples in input signals and is governed by the relation ($L = \log_2 N$), N is the number of samples in the input signal and L is number of scales in SLT. In DWT, the number of decomposition level is not governed by above relation, but it can be chosen, and choosing the right decomposition level will affect recognition results. Since the digital filter is achieved by shifting and delaying operations and the slantlet transform has short filters with length approach to (2/3) of the length of the iterated wavelet filter bank, therefore the SLT is faster than DWT. The SLT is an orthogonal transform and provides improved time localization than DWT, therefore it

will improve the recognition results. The comparison of the accuracy of the SLT system with other systems used in this work gives a conclusion that the SLT gives improved accuracy for the recognition of Arabic words.

7. REFERENCES

[1] Throat, R.A. and Jadhav, R.A. **,** " Speech Recognition System", International Conference on Advanced in Computing, Communication and Control, pp.607-609, Mumbai, 2009.

[3] Bourouba, H., Bedda, M., and Djemeli, R., "Isolated Words Recognition System Based on Hybrid Approach DTW/ GHMM", Informatica, Vol. 30, pp. 373-384, 2006.

[2] Anusuya, M.A. and Katti, S.K., "Speech Recognition by Machine: A Review", International Journal of Computer Science and Information Security (IJCSIS), Vol. 6, No. 3, pp.181-205, 2009.

[4] Pour, M.M. and Farokhi, F., "An Advanced Method for Speech Recognition", World Academy of Science, Engineering and Technology, 49, pp.995-1000, 2009.

[5] Burrus, C.S., Gopinath, R.A., and Guo, H., "Introduction to Wavelets and Wavelet Transform", Prentice Hall, 1998.

[6] Selesnick, I.W., "The Slantlet Transform", IEEE Transaction on Signal Processing, Vol. 47, No. 5, pp. 1304-1313, MAY, 1999.

[7] Furtunà, T.F., "Dynamic Programming Algorithms in Speech Recognition", Revista Informatica Economică, Vol. 46, No.2, pp. 94-99, 2008.

[8] Gata, S.M.,"A Multi Transform based Dynamic Time Warping Isolated Words Speech Recognition System", M.Sc. Thesis, University of Baghdad, Electrical Engineering Department, 2010.

[9] Paul, D. and Parekh, R., "Automated Speech Recognition of Isolated Words using Neural Networks", International Journal of Engineering Science and Technology (IJEST) , Vol. 3, No. 6, pp. 4993-5000, 2011.