Arabic *Speech Recognition*

Using Two Techniques Hybrid & 3D-Multiwavelet

Dr. Talib M. Jawad Abbas

Abstract

A key issue for implementing an accurate speech recognition system is the set of acoustic features extracted from speech signal. This paper presents two techniques for comparison. The first technique converts successfully the speech signal from (1-D) into two dimensional (2-D) forms. Next, the 2-D Multiwalidlet transform is applied to each 2-D signal. The second used transformation which is 3D-Multiwavelet (DMWT). For this transform set of speakers spoke the same word which arranged as slices of 2-D signals in acoustic space. These speakers represented the word as 3-D signal. The techniques apply the neural network as a classifier and dealing with text-dependent and text-independent speech recognition. The works are tested upon a database which consist of (28) speakers and uttered 7 Arabic words for each one.

It was compared with first technique which gave the result (85.71%-100%), the second gave (71.43%-100%). It is clear that first give much better performance than the second one.

1. Introduction

One of the advantages of speech as a biometric identifier is user convenience. Firstly, it is possible to give the identity claim using the same modality as the biometric sample itself. Secondly, it has been proposed that in a fixed phrase system, the user could select him/her the pass phrase [1].

The principal difficulty for speech recognition lies in the very nature of the speech signal. The speech signal is highly non-stationary and much information is contained in the transient parts what we as humans identify as the same speech component e.g. a certain phoneme, has a large variety of different pronunciations. They vary in time a they depend a lot on the context. Therefore, a few simple operations on the speech signal are not sufficient for recognition and a complicated and sophisticated processing chain has to be designed in order to come close to an acceptable recognition rate [2].

One of the most important researches in this field was done by **Scott Axelrod and Benoit Maison**, combine Hidden Markov Models of various topologies and Nearest Neighbor classification techniques in an exponential modeling framework with a model selection algorithm to obtain significant error rate reductions on an isolated word digit recognition task [3].

Amit Kumar Mishra, presents work which starts with the assumption that the randomness of speech signals could effectively be handled by coding styles that take care of variations both in time and spectral domain. Wavelet is an ideal choice for the present case. Presently speeches of two key words from speakers (six persons) of wide range of speaking styles were taken. As in common practice the speech signals after being sampled were passed through a first order filter. Then the wavelet transform were taken using pre-defined functions in Matlab™. Those transformed code after processing was studied for distinguishing features. From the 3-D plot of the wavelet transforms, the best wavelet transforms were chosen based on visual inspection and then applied for an effective algorithm. The results were satisfactory and improved results were obtained with continuous wavelets. Computation time taken was more than that tolerable by any hard real-time system [4].

Dr. Waleed A. Mahmud Al-Jouhar and Talib M. Jawad Abbas, presents a way of data combination of technique of several features and their mapping using discrete mltiwavelet transform (DMWT).This combination was tested for isolated-word speech recognition. It is shown that this approach introduce more accurate results. This is due to the use of MWT in the combination instead of putting several logic rules. This experiment considered as a good beginning in using multiwavelet in feature combination of speech signal. It was compared with method of linear combination applying to the same data which results in **(87.75%)**. For the DMWT gave **(90.81%).** It is clear that the new method give much better performance than the conventional one [5].

This paper includes two techniques for comparison 2D-Multiwalidlet and 3D-Multiwavelet as a parameter measurement (feature extraction), and using Back-Propagation (BP) neural network as a classifier.

2.The Stages of the Speech Recognition System

The automatic speech recognition (ASR) task is divided into three stages:

- Preprocessing
- Feature Extraction
- Speech Modeling/Classification (train/test).

The performance of the overall system is dependent on each of the three factors listed above, independently and combined. Selecting a feature extraction method and classifier often depends on the available resources and the intended application of the speech recognition system. Later, explanation for these stages will be mentioned. The Structure of the Proposed Arabic Speech Recognizer simplified by block diagram in figure (1).

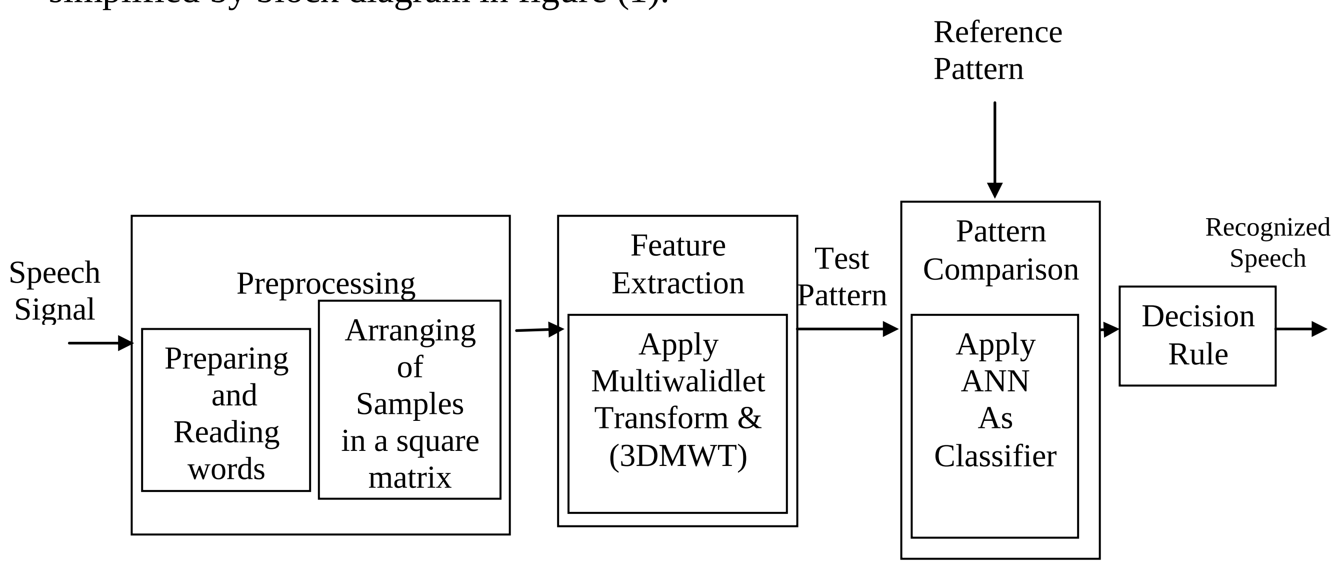

Fig (1) Structure of the Proposed Arabic Speech Recognizer

2.1The Pre-processing

This stage includes:-

1. Input of the speech signal after selecting the sampling rate (11025 Hz).

Then, the speech signal will be converted to a vector of data (samples). The length of the resulting vector is variable and depends on the type of the word.

2. Resize the variable length of the vectors in to a fixed size for all. This fixed size should be selected curvedly.

The main reason behind chose a fixed-length value is to convert the process from a vector of one dimension to a matrix form (2- dimension). This matrix should be a square form and power of 2. This is because the Multiwalidelet (or Multiwavelet in general) application assumes such requirement. For this reason, the selection of the fixed-length value must be done carefully taking into account the optimization of overall dimensions.

In this work, after many studies and test on different data, and different length, it was found that the suitable length (fixed-length) for all are (4096 samples), and this chosen come from note all length of the isolated words is very appropriate for all.

Now, the vector of any utterance must convert into new vector that length equal into 4096. The process of resizing the original length begins in computing the difference between the original length and the proposed length and from this difference will know how much addition or subtraction according to the following equation.

$$\text{Dif} = \text{original length} - \text{fixed length} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

This difference will lead us to two cases:

First: If Dif. is positive value, this mean that the original length is more than the proposed length, the following procedure must be done:

i) Divide the difference by 2 (fix (Dif. /2)), then remove half of the difference from the start of the original vector.
ii) Remove another half from the end of the original vector.
iii) The remaining vector will represent the new vector that has length equal to proposed length (power of two).

Second: If Dif. is a negative value, this means that the actual length less than the proposed length then the following procedure must be doing carefully.

i) Divided the difference by 2 (fix (abs (Dif. /2))), then add number of zero at the beginning of the new vector equal to half the difference.
ii) Put the original vector after zeros in the new vector (i.e. in this step the length of the new vector become origin vector plus zeros).
iii) Add number of zeros at the end of new vector (i.e. after the element of origin vector) equal to the half of differences.

3. Convert the vector of one dimension to square matrix form (2-dimension):-

As described previously, the length of vector is square and power of 2, the fixed length of the words is 4096, this number is square and power of 2. From this length one can obtain the matrix that will be with a dimension of 64×64.

The conversion process will be achieved:

i) Create a zero's matrix with 64×64 elements, named M.

ii) Divide the vector to 64 frames, each frame consists of 64 samples, such as :-

Frame 1 = 1... 64
Frame 2 = 65... 128
Frame 3 = 129... 192
: :
Frame 64 = 4032... 4096

iii) Here each frame represent row, so must convert these rows into column in the new matrix form, i.e. frame 1 represent column1, frame 2 represent column2,continue until reach frame 64 that represent column 64.

4. Convert from 2-dimensional array (matrix) to 3-dimensional array representation. To create a 3-dimensional signal for each word, this needs at least a combination of four speakers in matrix forms (in this work will be used eight speakers). Note that, these speakers are different but for the same words. The final matrix obtained from the preprocessing step is 64×64×8.

To apply the above procedure needs first to use the correct dimension of 64×64×8 but for simplicity it is reduced here as symbols and a lower dimension of 4×4.

The following example demonstrates this procedure for a given words. Let SP1w1, … , SP8w1 represent same words spoken by different speakers. Apply the processing on these words using the procedures described in steps 1, 2, 3 above in order to generate the matrices A, … , H respectively. Arranging these matrices as slices one by one under only one 3-D array M2.

A =

a_{11}	a_{12}	a_{13}	a_{14}
a_{21}	a_{22}	a_{23}	a_{24}
a_{31}	a_{32}	a_{33}	a_{34}
a_{41}	a_{42}	a_{43}	a_{44}

B =

b_{11}	b_{12}	b_{13}	b_{14}
b_{21}	b_{22}	b_{23}	b_{24}
b_{31}	b_{32}	b_{33}	b_{34}
b_{41}	b_{42}	b_{43}	b_{44}

C =

c_{11}	c_{12}	c_{13}	c_{14}
c_{21}	c_{22}	c_{23}	c_{24}
c_{31}	c_{32}	c_{33}	c_{34}
c_{41}	c_{42}	c_{43}	c_{44}

D =

d_{11}	d_{12}	d_{13}	d_{14}
d_{21}	d_{22}	d_{23}	d_{24}
d_{31}	d_{32}	d_{33}	d_{34}
d_{41}	d_{42}	d_{43}	d_{44}

E =

e_{11}	e_{12}	e_{13}	e_{14}
e_{21}	e_{22}	e_{23}	e_{24}
e_{31}	e_{32}	e_{33}	e_{34}
e_{41}	e_{42}	e_{43}	e_{44}

F =

f_{11}	f_{12}	f_{13}	f_{14}
f_{21}	f_{22}	f_{23}	f_{24}
f_{31}	f_{32}	f_{33}	f_{34}
f_{41}	f_{42}	f_{43}	f_{44}

G =

g_{11}	g_{12}	g_{13}	g_{14}
g_{21}	g_{22}	g_{23}	g_{24}
g_{31}	g_{32}	g_{33}	g_{34}
g_{41}	g_{42}	g_{43}	g_{44}

H =

h_{11}	h_{12}	h_{13}	h_{14}
h_{21}	h_{22}	h_{23}	h_{24}
h_{31}	h_{32}	h_{33}	h_{34}
h_{41}	h_{42}	h_{43}	h_{44}

M2 =

							m_{118}	m_{128}	m_{138}	m_{148}
						m_{117}	m_{127}	m_{1374}	m_{147}	m_{248}
					m_{116}	m_{126}	m_{136}	m_{146}	m_{247}	m_{348}
				m_{115}	m_{125}	m_{135}	m_{145}	m_{246}	m_{347}	m_{448}
			m_{114}	m_{124}	m_{134}	m_{144}	m_{245}	m_{346}	m_{447}	
		m_{113}	m_{123}	m_{133}	m_{143}	m_{244}	m_{345}	m_{446}		
	m_{112}	m_{122}	m_{132}	m_{142}	m_{243}	m_{344}	m_{445}			
m_{111}	m_{121}	m_{131}	m_{141}	m_{242}	m_{343}	m_{444}				
m_{211}	m_{221}	m_{231}	m_{241}	m_{342}	m_{443}					
m_{311}	m_{321}	m_{331}	m_{341}	m_{442}						
m_{411}	m_{421}	m_{431}	m_{441}							

Now the new matrix generated by this example is a 3-D array with a dimension of $4 \times 4 \times 8$.

3. TheFeature Measurement (Extraction)

Speech acquisition begins with a person speaking into a microphone or telephone. This act of speaking produces a sound pressure wave that forms an acoustic signal. The microphone or telephone receives the acoustic signal and converts it to an analog signal that can be understood by an electronic device. Finally, in order to store the analog signal on a computer, it must be converted to a digital signal [6].

In this work will present two techniques in feature extraction using **Hybrid Transforms** 2**D-Multiwalidlet** and **3D-Multiwavelet** hoping to fulfill high degree of accuracy in recognition.

3.1 A General Procedure For Computing 3-D DMWT

The following algorithm was used for the computation of 3-D discrete multiwavelet transform on multiwalidelet coefficients matrices using GHM four multifilter and using an over-sampled schema of preprocessing (repeated row preprocessing) [7] :-

1. ***Checking Matrix Dimensions:*** 3D–Matrix ($N \times N \times M$) should be a square matrix, $N \times N$ matrix, and must be a power of 2. So checking input matrix dimensions is the first step of the transform procedure. If the signal is not a square matrix and not a power of two, some operation must be done to resize the signal or add rows or columns of zeros to achieve this condition.
   In our work, this point is satisfied by default, because the previous step (preprocessing) is to be treated this point and generate a $64 \times 64 \times 8$ matrix.
2. ***Applying 2-D Multiwalidelet Transform***: - Here it is required to apply the two dimensional discrete multiwalidelet transform (described in 3.2) to each $N \times N$ input matrix, (i.e. applying 2-D DMWT no. of M-time, in all the M matrices ($N \times N$) in z-direction). Note that the 2-D DMWT is a critically-sampled scheme of preprocessing, thus, the result matrix is $N \times N \times M$ (i.e. $64 \times 64 \times 8$).
3. ***Applying 1-D DMWT Transform:*** In this step, one dimensional discrete multiwavelet transform (1-D DMWT) must be obtained on a vector in the z-direction, which is of length of M.
   Now, an $N \times N \times 2M$ matrix results from the $N \times N \times M$ original matrix using repeated row preprocessing.
4. ***Removing the repeated matrices:*** After doing some testing on the resulting matrices, it was found that it is necessary to remove the repeated matrices generated by using repeated row preprocessing.
   Thus, convert the resulting matrix of ($N \times N \times 2M$) to ($N \times N \times M$) by removing the even slices number (repeated slices). In our system removing [2, 4, 6,8,10,12,14] slices from $64 \times 64 \times 16$, leads to having just $64 \times 64 \times 8$ matrix. Finally, the resulting

N×N×M (64×64×8) matrix will represent the applied 3-DMWT on Multiwalidelet Coefficients.

3.2 A General Procedure For Computing Multiwalidlet Transform

This method has **hybrid** transformation effect which is not possible with other individual transforms. The following steps are followed in its computation:

1. ***Resizing:*** Here it is necessary to check matrix dimensions; matrix should be a square matrix, N×N matrix, where N must be the power of two. If the matrix is not a square matrix, a zero padding operation should be performed to the signal (adding rows or columns of zeros to get a square matrix).

2. ***Applying 2-D DMWT*****:** Here it is required to apply the two dimensions discrete multiwavelet transform (2_D DMWT), using a Critically-Sampled Scheme of preprocessing method of computation to each signal described in [8].

3. ***Subband Decomposition*****:** The resulting matrix of step two is decomposed into four subbands, these are:

- **a.** ***LL subband*****:** The upper left corner of the result coefficients matrix.
- **b.** ***LH subband*****:** The upper right corner of the result coefficients matrix.
- **c.** ***HL subband*****:** The lower left corner of the result coefficients matrix.
- **d.** ***HH subband*****:** The lower right corner of the result coefficients matrix.

4. ***Applying the Walidlet Transform:*** Each subband is analyzed via the discrete Walidlet transform described in [7].

5. ***Integration:*** Group the Walidlet coefficients of each subband to form a matrix according to their positions (LL, LH, HL, and HH). The resultant matrix is called the Multiwalidlet coefficients.

4. *Classification*

There are different types of ANN and has different architectures. Some types of ANN such as Feed Forward Neural Network (FFNN) have different number of hidden layers and different number of neurons in each layer. So it is important to select the appropriate neural network and fixing the architecture of the selected network. Despite the scarcity of the literature available on the implementation of speech recognition (SR) using NN, we have adopted a back-propagation (Bp) network model and we have found that it can achieve results as good as another model.

According to the principle that there is no general systematic method available or rules of thumb to fix these parameters so a trial and error was applied to get the best choice. It can be said in abstract that the practical and the experimental result in selecting the number of layers and neurons don't depend on certain element.

Many variations of the back propagation algorithm are proposed. In the first choice was (trainlm), but this algorithm didn't succeed here because it needs big space of memory an addition to the fact that inputs itself is big too as mentioned before. This is why execution operation on computer became very slow. Therefore another algorithm, it was succeeded and avoided the obstacle caused by earlier once (trainrp).

Performance Function was used to calculate network performance during training. In our work using two types: the mean square error performance function (MSE), it's high speed and not exacted with independent. Another function is the mean square error with regularization performance (MSERGE). It measures the network performance as the weight sum of two factors: the mean square error and mean square weight and bias values.

5. *Experimental Results of DMWT*

After presenting algorithms and preparing matrices as data and writing required programs and its implementation, the recognition result was very good. This work indicates that using hybrid 2D-Multiwalidlet transform gave better result, if a comparison is done with second technique (3D-Multiwavelet). It was compared with first technique which result (85.71%-100%), the second gave (71.43%-100%). It is clear that first give much better performance than the second one. The work is tested upon a database which consist of (28) speakers and uttered 7 Arabic words for each one.

The speech recognizer described in the paper was fully implemented in MATLAB version 7 programming language and Cool Edit pro2.1 as speech analyzer under Microsoft Window XP, and was subjected to several test inputs. The recognition performances of these experiments are shown below:

No.	Training Sequence	Test Sequence	Multiwalidelet		3D-DMWT	
			Text-De	Text-Ind	Text-De	Text-Ind
1-	S 01,S 02,S 03,S 04,S 11,S 13,S 15,S 16	S 05,S 07,S 08,S 17,S 20,S 22,S 24,S 26	100%	100%	100%	71.43%
2-	S 03,S 04,S 05,S 08,S 07,S 26,S 23,S 18	S 01,S 02,S 06,S 13,S 20,S 22,S 24,S 28	100%	85.71%	100%	85.71%
3-	S 03,S 04,S 05,S 08,S 07,S 26,S 23,S 18	S 02,S 11,S 06,S 13,S 22,S 15,S 16,S 17	100%	85.71%	100%	85.71%
4-	S 10,S 12,S 14,S 16,S 18,S 20,S 22,S 24	S 02,S 04,S 06,S 08,S 09,S 11,S 13,S 15	100%	85.71%	100%	71.43%
5-	S 20,S 22,S 24,S 26,S 21,S 23,S 25,S 27	S 12,S 14,S 16,S 18,S 11,S 13,S 15,S 17	100%	100%	100%	71.43%
6-	S 10,S 11,S 12,S 13,S 14,S 15,S 16,S 17	S 02,S 03,S 04,S 05,S 06,S 07,S 08,S 09	100%	100%	100%	100%
7-	S 10,S 11,S 12,S 13,S 14,S 15,S 16,S 17	S 01,S 03,S 04,S 05,S 20,S 07,S 08,S 27	100%	100%	100%	85.71%
8-	S 10,S 11,S 12,S 13,S 14,S 15,S 16,S 17	S 02,S 26,S 04,S 05,S 28,S 07,S 29,S 09	100%	100%	85.71%	100%
9-	S 10,S 11,S 12,S 13,S 14,S 15,S 16,S 17	S 02,S 03,S 21,S 05,S 06,S 19,S 08,S 22	100%	100%	100%	100%
10-	S 07,S 11,S 12,S 13,S 14,S 15,S 16,S 09	S 28,S 03,S 04,S 05,S 06,S 17,S 08,S 20	100%	100%	100%	71.43%
11-	S 10,S 11,S 12,S 13,S 14,S 15,S 16,S 22	S 02,S 03,S 04,S 05,S 06,S 17,S 08,S 09	100%	100%	100%	85.71%
12-	S 21,S 11,S 12,S 13,S 14,S 15,S 16,S 17	S 02,S 03,S 04,S 05,S 06,S 17,S 08,S 09	100%	100%	100%	85.71%
13-	S 10,S 11,S 24,S 13,S 18,S 15,S 16,S 17	S 02,S 03,S 04,S 05,S 06,S 17,S 08,S 09	100%	85.71%	100%	71.43%
14-	S 02,S 03,S 04,S 05,S 06,S 17,S 08,S 09	S 17,S 16,S 15,S 14,S 13,S 12,S 11,S 10	100%	100%	100%	100%
15-	S 02,S 03,S 04,S 05,S 06,S 17,S 08,S 09	S 14,S 15,S 16,S 17,S 10,S 11,S 12,S 13	100%	100%	100%	85.71%
16	S 02,S 03,S 04,S 05,S 06,S 17,S 08,S 09	S 14,S 15,S 22,S 17,S 24,S 11,S 12,S 13	100%	100%	100%	85.71%
17	S 01,S 03,S 04,S 05,S 23,S 07,S 08,S 19	S 20,S 15,S 16,S 17,S 10,S 11,S 12,S 21	100%	100%	100%	85.71%
18	S 19,S 03,S 04,S 05,S 06,S 07,S 08,S 09	S 17,S 16,S 15,S 14,S 13,S 12,S 11,S 01	100%	100%	100%	85.71%
19	S 17,S 16,S 15,S 14,S 10,S 11,S 12,S 13	S 20,S 21,S 22,S 23,S 24,S 25,S 26,S 27	100%	100%	100%	71.43%
20	S 24,S 03,S 04,S 05,S 06,S 07,S 08,S 20	S 10,S 11,S 12,S 13,S 14,S 15,S 16,S 21	100%	100%	100%	85.71%

Table of Results to 2D-Multiwalidlet & 3D-Multiwavelet

Note: Speaker=S

6. Conclusions

1. The selection of 64*64 resulted in more cancellation cases than addition, this is very important since the addition some time causes problems while the cancellation was performed on the start or end of the words which usually give indication of noise.
2. At the beginning four slices was selected to apply 3D-Multiwavelet and then six and eight, but the best result was the selection of last one.

3. The results show that the best neural network type that can be used for ASR is the feed forward neural network trained with Rp algorithm among the other types. These coincide with the ability of this type of network and training algorithm in solving classification problems.
4. The **hybrid** technique which used was Multiwalidlet. This technique is a combination of 2-dimentional Discrete Multiwavelet Transform (DMWT) and Walidlet Transform, which gave best recognition.

7. References

[1] Harsh Gupta, Ville Hautamäki, Tomi Kinnunen and Pasi Fränti "Field Evaluation of Text-Dependent Speaker Recognition in an Access Control Application", Speech and Image Processing Unit, Department of Computer Science, University of Joensuu, Joensuu, Finland
{gupta, villeh, tkinnu, franti}@cs.joensuu.fi

[2] Dharmendra P Kanejiya, "Speech Recognition Using Hidden Markov Model With Neural Network Probability, Estimators". IIT Delhi , 2000.

[3] Scott Axelrod and Benoit Maison, "Combination of Hidden Markov Models With Dynamic Time Warping, For Speech Recognition". IBM T.J. Watson Research Center, P.O. Box 218, Yorktwon Heights, NY 10598, USA, 2004.

[4] Amit Kumar Mishra ,"Speech Processing Using Wavelets For Speaker Identification", Defense R & D Organization, Chandipur, Balasore, Orissa India-756 025, 2004, Email: akmishra@ieee.org

[5] Dr. Waleed A. Mahmud Al-Jouhar and Talib M. Jawad Abbas, "Feature Combination & Mapping Using Multiwavelet Transform", Journal of Al-Rafidain University College For Sciences, No. 19 8th Year 1427-2006, E-mail: Jouranl@alrafidain.net.

[6] M. M. El Choubassi, H. E. El Khoury, C. E. Jabra Alagha, J. A. Skaf and M. A. Al-Alaoui, "Arabic Speech Recognition Using Recurrent Neural Networks" . Electrical and Computer Engineering Department Faculty of Engineering and Architecture – American University of Beirut 1107 2020, P.O. Box: 11-0236, LEBANON, 2002. adnan@aub.edu.lb.

[7] Ghaida Wajeh Ahmed, "Speaker Recognition Using Hybrid Transform", Msc. Thesis, Iraqi Commission for Computers & Informatics- Informatics Institute for Postgraduate Studies, Baghdad 2006.

[8] R. Alaa AL-Deen, "Aproposed Walidelet Transform with It's Application in Image Enhancement", Msc. Thesis, Iraqi Commission for Computers & Informatics-Informatics Institute for Postgraduate Studies, Baghdad 2005.

تمييز الكلام العربى
باستخدام أسلوبين هجين ومتعدد المويجة ثلاثى الأبعاد

أن العنصر الأساسي في بناء وتنفيذ نظام تمييز كلام هي في اختيار المجموعة المناسبة لاستنباط الخواص. في هذا البحث تم اعتماد أسلوبين، الأسلوب الأول يعمل على تحويل مصفوفة أحادية الأبعاد الى مصفوفة ثنائية الأبعاد وتطبيق (Multiwalidlet) على كل أشارة ثنائية الأبعاد لغرض استنباط الخواص. الأسلوب الثاني استخدام محول متعدد المويجة ثلاثي الأبعاد(3D-Multiwavelet) ،ولهذا الغرض مجموعة من المتكلمين تنطق نفس الكلمة رتبت على شكل شرائح في مصفوفة ثلاثية. الأسلوبان طبق الشبكة العصبية كمصنف وتعامل مع تمييز الكلام ذو النص المعتمد والغير معتمد. تم فحص العمل على قاعدة بيانات تتكون من (٢٨) متكلم نطقوا (٧) كلمات عربية لكل منهم. كان معدل النسبية المئوية لدقة الأسلوب الأول هو (٨٥.٧١%-١٠٠%) والثاني هو (٧١.٤٣%-١٠٠%) و يظهر أن الأول كان الأكفأ من الثاني في تمييز الكلام.