Arabic Word Recognition Based on 3D Radon and Multiwavelet Neural Network

Dr. Tarik Zeyad Ismaeel
Engineering College, University of Baghdad /Baghdad
E-mail:Suad_1934@yahoo.com

Received on: 26/9/2012 & Accepted on: 4/4/2013

ABSTRACT

In this paper, an automatic speaker–independent Arabic word speech recognition system is presented using 3D Radon and Multiwavelet neural network. The approach contains combining multiwavelet theory to neural network which lead to fabricate a Multiwavenet. Position and dilation of the Multiwavenets are fixed and the weights are optimized according to learning algorithm in the network. The feature extraction for real Arabic word signals through 3D radon model is used. The proposed terminology here is training process for some words of all speakers done in Multiwavenet learning phase then test for the other sample speech signals for speakers have been used in Multiwavenet classification phase. Success theory of Multiwavenets has been generalized by extension to biorthogonal wavelets which lead to identification system development. Results show the effectiveness of the proposed system presented in this paper. The accuracy in the detection process was 86% when using utterances outside the training database and around 94% when using the whole utterances database in system test process. The proposed algorithms were implemented using MATLAB2011a.

Keywords: Speaker independent system, Multiwavelet, Multiwavenet, 3D Radon transform.

تمييز الكلمات العربية بإستعمال تحويل رادون الثلاثي الأبعاد مع تحويل المويجة المتعدد والشبكات العصبية

الخلاصة

في هذا البحث تم بناء منظومة لتمييز الكلمات العربية التي لا تشترط استخدام اصوات اشخاص محددين لكي يتم تمييز الكلمة باستخدام تحويل رادون الثلاثي وتحويل المويجة المتعدد والشبكات العصبية. النظام يشمل على بناء الشبكة العصبية للتحويل المويجي المتعدد من جمع النظامين سويةً (الشبكة العصبية وتحويل المويجة المتعدد). أوزان الشبكة العصبية تم احتسابها بالصورة الأمثل عن طريق خوارزمية

https://doi.org/10.30684/etj.31.7A14
2412-0758/University of Technology-Iraq, Baghdad, Iraq

تدريب الشبكة العصبية. تم استخدام تحويل رادون الثلاثي لاستخراج خواص الكلمات العربية المناسبة لعملية التمييز. تشتمل خطوات بناء المنظومة على مرحلتين المرحلة الأولى هي تدريب الشبكات العصبية ذات تحويل المويجة المتعدد على التعامل مع الكلمات بواسطة تغيير اوزان الشبكات العصبية في مرحلة التدريب والمرحلة الثانية هي اختبار الشبكة في تمييز الكلمات التي تصدر من اشخاص لم يتم استخدام اصواتهم في مرحلة التدريب. أثبتت المنظومة التي تم تقديمها في هذا البحث كفائتها حيث كانت نسبة النجاح عند أستعمال أصوات لم تشترك في عملية التدريب هي ٨٦% وعند الأخذ بنظر الأعتبار كافة الأصوات كانت نسبة النجاح الكلي حوالي ٩٤%. تم استخدام برنامج MATLAB2011a في بناء خوارزميات النظام المقترح.

INTRODUCTION

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. The study of these rules and their implications in human communication is the domain of linguistic. The study and classification of the sounds of speech is called phonetics. Speech can be presented in terms of its message content or information. 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 desired for him all the time. The task of a speech recognizer is to determine automatically the spoken words, regardless of the variability introduced by speaker identity, manner of speaking, and environmental conditions [2]. Speaker recognition system is usually divided into two different branches, speaker dependent and speaker independent. Most applications in which a voice is used as the key to confirm the identity of a speaker are classified as speaker verification. Speaker identification is the most difficult problem of the two [3].in this paper we propose a new technique for speaker independent speech recognition based on 3D Radon as feature extraction and multiwavelet neural network as classifier.

3D RADON TRANSFORM [4].

The general form of the radon transformation in N dimension is given by:

$$g(p,\xi) = \int g(r)\delta(p - \xi.r)dr \qquad \text{... (1)}$$

Where r and $\xi \in R^N$ but ξ has only N-1 degrees of freedom, e.g., done by normalizing $|\xi|=1$. A common approach is to express the vector ξ in hyber-spherecal coordinate. In two dimensions $\xi = (\cos\theta, \sin\theta)^T$ and the three sizes the vector can be chosen to be as follows:

$$\xi = \begin{pmatrix} \cos\phi\cos\theta \\ \sin\phi\sin\theta \\ \cos\theta \end{pmatrix}^T \quad \ldots (2)$$

The fundamental body of 3D Radon transform is in three dimensions a plane [4].

$$\xi.r = x\cos\phi\sin\theta + y\sin\phi\sin\theta + z\cos\theta \quad \ldots (3)$$

$$g(\rho,\theta,\phi) = \int g(r)\delta(\rho - x\cos\phi\sin\theta - y\sin\phi\sin\theta - z\cos\theta)dxdydz.....(4)$$

Here it's noted that the three dimension radon transform will modification the three dimension signal $g(r)$ into three dimension parameter $g(\rho,\theta,\phi)$

Algorithms of computing 3D Radon Transform

To compute the 3D Radon transform of given utterances in order to obtain the projection of these image, which reduce the dimension from three dimension utterances to two dimension projection, we can compute the 3D Radon transform according to the flowchart Figure (1).

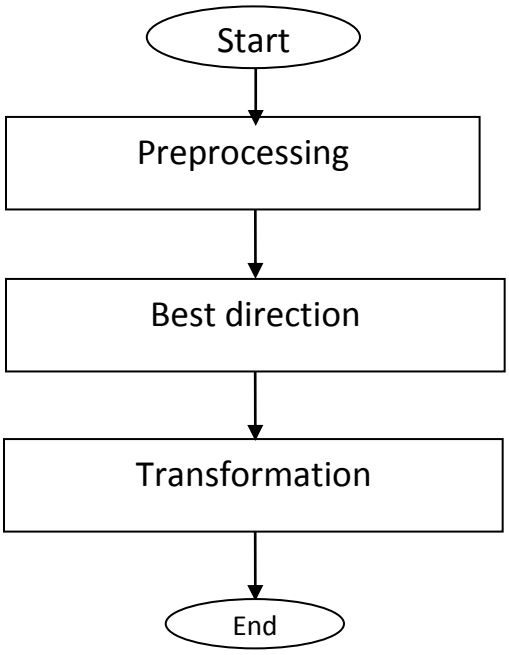


Figure (1) 3D Radon Transform flowchart diagram [5].

The preprocessing block is used to extract the required features and resize the selected volume with prime number (one of requirement of the computation of 3D Radon Transform).

One of the important facts that must be satisfied is the dimension of the given utterance to be passed to the Radon Transform space. Usually it is recommended to be prime number for such transform of any dimensions.

The Best Direction blocks are used for obtaining a new location of any input information which depends on two angles Φ and θ.

Thus, it is required to obtain a new location for arranging the input signal with respect to a given utterance. Therefore, it operates on any arbitrary data with specific size equal to the size of input data. Hence, it is results in creating a new three volumes that represents the different utterances with respect to X-prime, Y-prime, and Z-prime. This achieved by taking single slices and converting them to spherical coordinate. Thus, we can get rid of the complex computation required in Cartesian coordinate. Then rearrange the input data in accordance to the results from the applied Best Direction.

To compute the best directions, the following steps should be followed:

Step1: Define (x, y, z) arbitrary vector.
Step2: Arrange vectors in volumes in different views X, Y, and Z.
Step3: Calculate Φ and θ in Radian.
Step4: Convert X (Cart) to X (Spherical).
Step5: Convert Y (Cart) to Y (Spherical).
Step6: Round the X (Spherical) and Y (Spherical).
Step7: Rearrange the contents of data according to X (Spherical) and Y (Spherical).
Step8: Compute the Fourier slices. The resulting matrix after Fourier slices are (RR)

Last part of the 3D Radon Transform is the transformation block which is used to compute the accumulated summing of the resulting slice after the rearranging to obtain 3-D Radon Transform Projection.

BACK PROPAGATION ADAPTIVE MULTIWAVENET (BPAMWN)

Back Propagation Neural Network (BPNN) is one the most popular mapping neural network. But it has some problems such as trapping into local minima and slow convergence. Wavenets emerged as a combination of wavelet analysis and neural networks and proved powerful in dealing with shortcomings of traditional neural networks such as BPNN. Wavelets are powerful signal analysis tools. They can approximately realize the time-frequency analysis using a mother wavelet. The mother wavelet has a square window in the time-frequency space. The size of the window can be freely varied by two parameters, namely, dilation and translation. Thus, wavelets can identify the localization of unknown signals at any level. Because of these superior properties, wavelets have been successfully integrated with BPNN [6].

Multiwavelets, which consist of more than one scaling function, have better properties than traditional wavelets. It is possible to construct orthogonal (real-valued) basis for which the multiscaling functions of multiwavelets have compact support,

approximation order greater than one, and symmetric properties, which are not all simultaneously possible for traditional wavelet basis [7].

A back propagation adaptive multiwavenet (BPAWN) is proposed in this section. To design BPAWN, hidden layer sigmoidal activation function of BPNN is replaced with multiscaling function.

The architecture of the proposed multiwavenet is basically the same as the BPWN [8], except that the wavelet function of hidden layer node of BPWN is replaced with two or more scaling functions of a multiwavelet. The architecture of a single-output BPAMWN is shown in Figure (2) below.

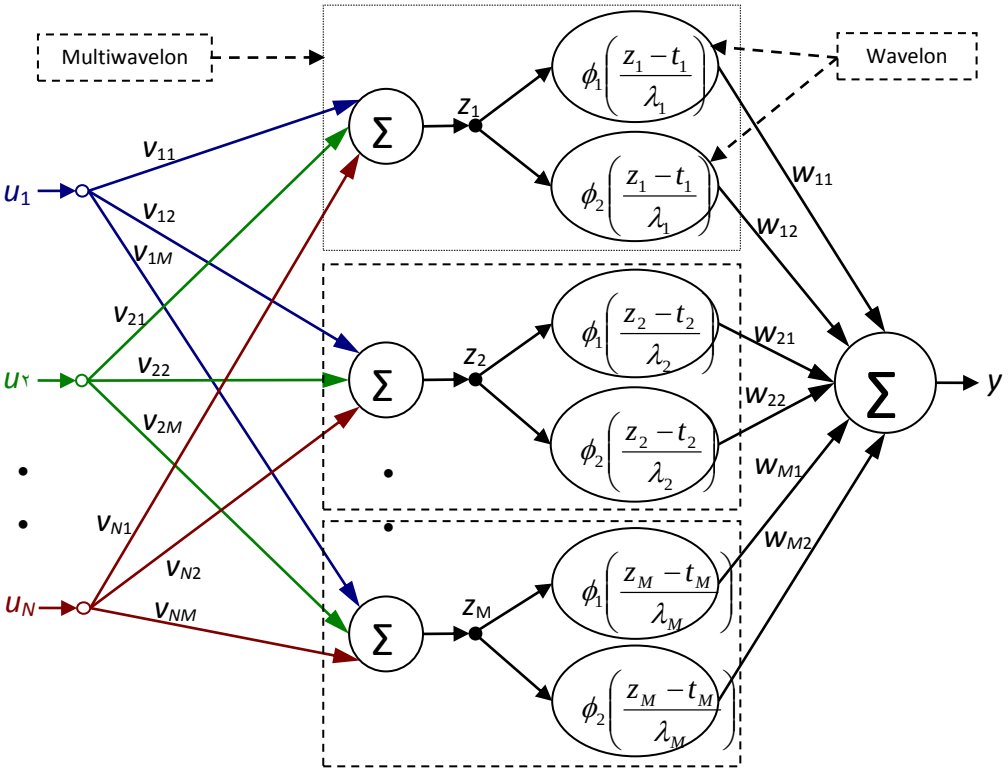


Figure (2) The proposed BPAMWN.

The output of the BPAMWN is

$$y(U_s) = \sum_{i=1}^{M}\sum_{L=1}^{r} w_{iL}\,\varphi_L\left(\frac{z_{is}-t_i}{\lambda_i}\right) \quad , \qquad s = 1 \ldots P \ldots (5)$$

$$z_{is} = \sum_{j=1}^{N} v_{ji} u_{js} \qquad \text{... (6)}$$

Where M in equation (5) is the number of multiwavelons, r is the multiplicity of the multiscaling function and each multiwavelon has r wavelons, t_i and λ_i are the translation and dilation of i^{th} multiwavelons scaling functions respectively, φ_L is the L^{th} scaling function , $U_s = \{u_{1s}, u_{2s}, \ldots, u_{Ns}\}$ is the s^{th} input vector of the total P input vectors in the training set, N is the number of elements of each input vector (input dimension), z_{is} is the inner product between the input vector U_s and the i^{th} input weight vector $V_i = \{ v_{1i}, v_{2i}, \ldots, v_{Ni} \}$ (weights between input nodes and i^{th} multiwavelon), w_{iL} is the weight between L^{th} wavelon of i^{th} multiwavelon and the output, and $y(U_s)$ is the output of the network.

Gradient descent method is used for training of the network parameters. The objective function to be minimized is:

$$C = \frac{1}{2P} \sum_{s=1}^{P} \left(y(U_s) - f(U_s) \right)^2 \qquad \ldots (7)$$

Where P is the number of training pairs, $y(U_s)$ is the output of the BPAMWN and $f(U_s)$ is the desired output.

Batch training mode is used where all training pairs $\{U_s, f(U_s)\}$, $s = 1, \ldots, P$ should be processed before parameters could be updated. Parameters are modified in the opposite direction of the gradient of C. To speed up the convergence rate, momentum term is included in parameter's update [9]. Let:

$$\tau_{is} = \frac{z_{is} - t_i}{\lambda_i},$$

$$\phi_L(\tau_{is}) = \phi_L\left(\frac{z_{is} - t_i}{\lambda_i}\right), \text{ and}$$

$$e_s = y(U_s) - f(U_s).$$

Partial derivatives are expressed as follows:

$$\frac{\partial C}{\partial w_{iL}} = \frac{1}{P} \sum_{s=1}^{P} e_s \phi_L(\tau_{is}) \qquad \text{... (8)}$$

$$\frac{\partial C}{\partial v_{ji}} = \frac{1}{P}\sum_{s=1}^{P}\sum_{L=1}^{r} e_s w_{iL} \frac{\partial \phi_L(\tau_{is})}{\partial \tau_{is}} u_{js} \lambda_i^{-1} \quad \text{... (9)}$$

$$\frac{\partial C}{\partial t_i} = -\frac{1}{P}\sum_{s=1}^{P}\sum_{L=1}^{r} e_s w_{iL} \frac{\partial \phi_L(\tau_{is})}{\partial \tau_{is}} \lambda_i^{-1} \quad \text{..... (10)}$$

$$\frac{\partial C}{\partial \lambda_i} = \frac{1}{P}\sum_{s=1}^{P}\sum_{L=1}^{r} e_s w_{iL} \frac{\partial \phi_L(\tau_{is})}{\partial \tau_{is}} \tau_{is} \lambda_i^{-1} \quad \text{... (11)}$$

Parameters can be updated as follows:
h = iteration number

$$w_{iL}^{h+1} = w_{iL}^{h} - \eta \frac{\partial C}{\partial w_{iL}} + \alpha \Delta w_{iL}^{h} \quad \text{... (12)}$$

$$v_{ji}^{h+1} = v_{ji}^{h} - \eta \frac{\partial C}{\partial v_{ji}} + \alpha \Delta v_{ji}^{h} \quad \text{... (13)}$$

$$t_i^{h+1} = t_i^{h} - \eta \frac{\partial C}{\partial t_i} + \alpha \Delta t_i^{h} \quad \text{... (14)}$$

$$\lambda_i^{h+1} = \lambda_i^{h} - \eta \frac{\partial C}{\partial \lambda_i} + \alpha \Delta \lambda_i^{h} . \quad \text{... (15)}$$

where $\Delta x^h = x^h - x^{h-1}$.

Parameter initialization has a significant impact on the convergence rate of the BPAMWN. A heuristic method for parameter initialization is proposed here.

v_{ji} is initialize to be in the range (−2, 2) randomly. w_{iL} is initialized to be zero. The input to each multiwavelon is a dilated and translated version of the inner product between the input vector U_s and the input weight vector $V_i = \{v_{1i}, v_{2i}, \ldots v_{Ni}\}$ of that multiwavelon. Therefore the initial values of dilations and translations depend on this inner product which was noted by z_{is}. Dilation of the i^{th} multiwavelon is initialized to be the difference between the maximum value of z_{is} (highest value of z_{is} when all samples are presented to the network) and the minimum value of z_{is} (lowest value of z_{is} when all samples are presented to the network) divided by two. The maximum value of z_{is} denoted by z_{imax} is calculated as the inner product of the i^{th} input weight vector V_i and U_{max}, where U_{max} is the vector of highest values of input nodes when all samples are presented to the

network (In other words if a matrix U is constructed by concatenating input vectors U_s of all samples such that it has the same number of rows as each input vector U_s and it has as many columns as there is samples, then U_{max} is the vector of the maximum values along the second dimension (the columns) of U). The minimum value of z_{is} denoted by z_{imin} is calculated in a similar manner as the inner product of the i^{th} input weight vector V_i and U_{min}, where U_{min} is the vector of lowest values of input nodes when all samples are presented to the network. Translation of the i^{th} multiwavelon is initialized to be the sum of z_{imax} and z_{imin} divided by two. The following equations express the parameter initialization more rigorously.

$$v_{ji} = rand() * 4 - 2 \text{for } j = \{1,\ldots, N\},\ i = \{1,\ldots, M\} \quad \ldots(16)$$

$$w_{iL} = 0 \text{ for } i = \{1,\ldots,M\},\ L = \{1,\ldots,r\} \quad \ldots(17)$$

$$U_{\max} = \max_{s\in\{1,\cdots,P\}} (U_s) \quad \ldots (18)$$

$$U_{\min} = \min_{s\in\{1,\cdots,P\}} (U_s) \quad \ldots (19)$$

$$z_{i\max} = V_i^T \cdot U_{\max} \quad \text{(Inner product)} \quad \ldots (20)$$

$$z_{i\min} = V_i^T \cdot U_{\min} \quad \text{(Inner product)} \quad \ldots (21)$$

$$\lambda_i = \frac{z_{i\max} - z_{i\min}}{2} \quad \ldots (22)$$

$$t_i = \frac{z_{i\max} + z_{i\min}}{2} \quad \ldots (23)$$

Where *rand*() is a function that generates random number in the range (0, 1). The activation function of output node gives two values for y either 0 or 1. For match y=1 for mismatch y=0.

PROPOSED SPEECH RECOGNITION SYSTEM

The general block diagram for the proposed model of speech recognition system is shown in Figure (3).

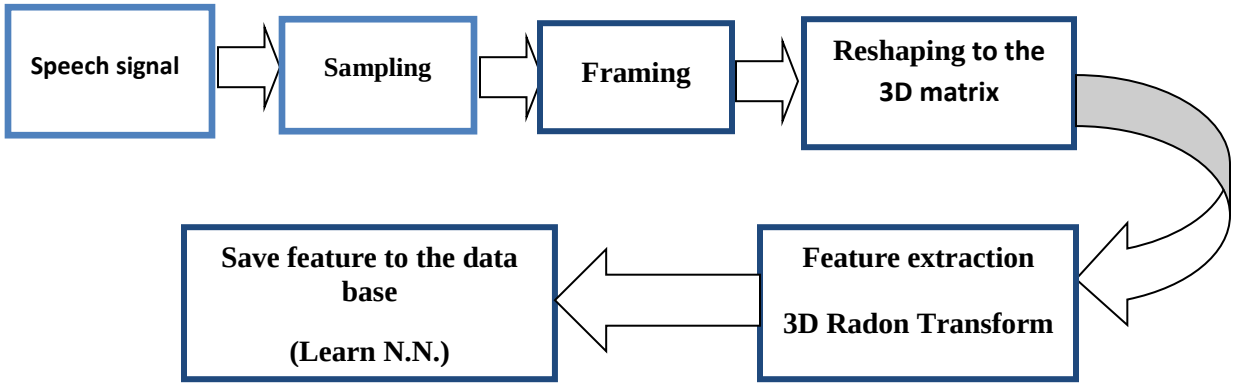
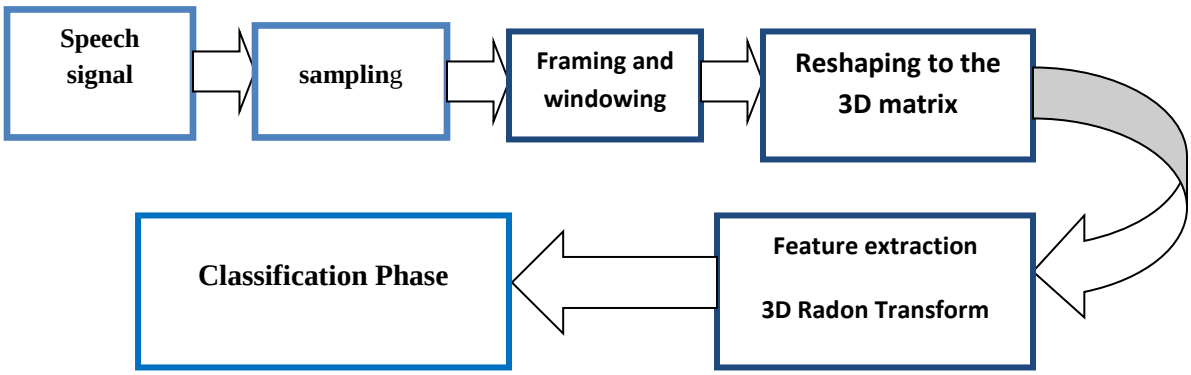


Figure (3) (a) The proposed learning phase.



Figure(3). (b). The proposed classification phase.

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 words are spoken by one speaker, and this speaker utters each word by different fifteen versions. It is clear that the lengths of the words are different; so that the versions for the same word vary in length. 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. In the second stage the speech signals are sampled to convert it from analogue to digital signal. The sampling rate has been down sampled from 44 KHz to 8 KHz. In the testing phase before the comparison between the tested word and the words in the data base the length

of the tested word must be matched with the word in the database. If it is less zero padding will be used if is higher some of the samples at the end will be eliminated.

- In third stage the continuous speech signal is blocked in frames of N samples. 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. After that 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. After windowing convert the matrix in 3D form. Then extract the feature using 3D Radon transform. After extract the feature Multiwavenet classifier is used. After the training is done, the parameters of the multiwavenet classifier, namely weights, translations and dilations are stored and are used in the identification stage when the classifier works in simulation mode. The processing in both training and identification stages is similar except for the classifier, which works in training mode in the training stage and in simulation mode in the identification stage.

EXPERIMENT RESULT

For the speaker independent proposed algorithm, 4 words should be selecting for each speaker for the training phase. Next, many different sequences of spoken words are applied for the recognition phase. These data consist of seven words [(W_1) to (W_7)] namely the, Arabic words:

W_1 يمين W_2 صباح W_3 ياسين W_4 وفي W_5 الخير W_6 مساء W_7 لندن

The results obtained for training and testing are given in (**Table 1**).

T able (1) The detailed results of matching between 12 training speakers and 8 testing speakers.

Speakers / words	Training												Testing								Evaluation
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
الخير	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	19
لندن	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	×	✓	×	17
مساء	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	19
صباح	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	20
وفي	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	20
يمين	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	20
ياسين	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×	✓	19

Samples of the words used in the experimental work spoken by the third individual are shown in Figure (4).

الخير

صباح

لندن

مساء

وفي

ياسين

يمين

6- Conclusion

In this paper, a multiwavenet system is used for the analysis of Arabic word speech signals to identify spoken word. The confirmed results show that the proposed method can make an effective interpretation for Arabic words speaker recognition system. As it can be seen from Table (1) the accuracy in the detection process was 86% when using utterances outside the training database and around 94% when using the whole utterances database in system test process. The 3D radon has been

demonstrated to be an effective tool for extracting information from the speech signals and it is robust against noise in the real word signals. Learning process is so fast and does not require any external intervention, making these methods very useful in practical applications. Generally speaking, we have shown that the use of Multiwavenet experimentally enhanced its recognition performance while also preserved its robustness to additive Gaussian noise. In the future we plan to compare also its generalization for unlearned voice and put the system in text-dependent mode. We believe that this approach will broaden a general perspective for future speaker identification system.

REFERENCES

[1]. Abbas,T. M. J. "Speech Recognition Using Features Combination", Ph.D. Thesis, Iraqi Commission for Computers & Informatics -Informatics Institute for Postgraduate Studies, August, 2005.

[2]. Amin,T. B. & I. Mahmood, "Speech Recognition Using Dynamic Time Warping", IEEE, 2nd International Conference on Advances in Space Technologies, Islamabad, Pakistan, vol. 2, pp.74-79, 29th-30th November, 2008.

[3]. Rabiner, L. R. & R. W. Schafer, "Digital Processing of Speech Signals", Englewood Cliffs, New Jersey: Prentice Hall, 1978.

[4] P. "Radon Transform Theory and implementation", Ph.D. Thesis, section digital signal processing-Technical, University of Denmark, 1996.

[5]. Toft, M. R. Shaker," 3D Finger Identification Using 3D RIDGLET Neural Transformation", Department of ComputerEngineering College of Engineering Baghdad University.

[6]. A. K." Image reconstruction using hybrid transform" M.Sc. Thesis University of Baghdad, 2010.

[7]. Ibrahim, X. Li and X. Gao, "Multiwavelet Neural Network: A Novel Model," IEEE International Conference on Systems, Man and Cybernetics, Vol. 3, pp. 2629-2632, Oct. 2003.

[8]. Rying, E. A. "A Novel Focused Local Learning Wavelet Network with Application to In Situ Monitoring during Selective Silicon Epitaxy", Ph.D. Thesis, Electrical Engineering, North Carolina State University, USA, May 2001.

[9]. Abbas,A. I. "Face Identification Using Multiwavelet-based Neural Network", M.Sc. Thesis University of Baghdad, 2010.