Isolated Word Speech Recognition Using Mixed Transform

Assist. Prof. Dr. Sadiq Jassim Abou-Loukh and **Eng. Shahad Mujeeb Abdul-Razzaq**

University of Baghdad, College of Engineering, Electrical Engineering Department
Email: doctor_sadiq@yahoo.com

ABSTRACT :

Methods of speech recognition have been the subject of several studies over the past decade. Speech recognition has been one of the most exciting areas of the signal processing. Mixed transform is a useful tool for speech signal processing; it is developed for its abilities of improvement in feature extraction. Speech recognition includes three important stages, preprocessing, feature extraction, and classification. Recognition accuracy is so affected by the features extraction stage; therefore different models of mixed transform for feature extraction were proposed. The properties of the recorded isolated word will be 1-D, which achieve the conversion of each 1-D word into a 2-D form. The second step of the word recognizer requires, the application of 2-D FFT, Radon transform, the 1-D IFFT, and 1-D discrete wavelet transforms were used in the first proposed model, while discrete multicircularlet transform was used in the second proposed model. The final stage of the proposed models includes the use of the dynamic time warping algorithm for recognition tasks. The performance of the proposed systems was evaluated using forty different isolated Arabic words that are recorded fifteen times in a studio for speaker dependant. The result shows recognition accuracy of (91% and 89%) using discrete wavelet transform type Daubechies (Db1) and (Db4) respectively, and the accuracy score between (87%-93%) was achieved using discrete multicircularlet transform for 9 sub bands.

KEYWORDS: Mixed Transform, Radon Transform, Discrete Wavelet Transform, Discrete Multicircularlet Transform, Dynamic Time Warping.

تمييز الكلمات المفصولة باستخدام التحويلات الخليطة

أ.م.د. صادق جاسم ابو اللوخ **م.م. شهد مجيب عبد الرزاق**

قسم الهندسة الكهربائية / كلية الهندسة / جامعة بغداد

الخلاصة:

طرائق تمييز الكلام كان موضوع كثير من الدراسات خلال العقد الماضي. الكلام هو الطريقة الطبيعية للتواصل بين البشر ويعتبر تمييز الكلام واحد من المجالات المهمة في معالجة الإشارة. التحويلات الخليطة هي أداة مفيدة في معالجة إشارة الكلام، وقد تم تطويرها من اجل تحسين تمثيل الإشارة المستخلصة. يتضمن تمييز الكلام ثلاث أجزاء أساسية: معالجة مسبقة للإشارة، استخلاص الميزات، والتصنيف. تتاثر دقة تمييزالكلام بمرحلة استخلاص الميزات لذلك فقد تم اقتراح نماذج مختلفة من التحويلات الخليطة. ان خصائص الكلمات المسجلة ستكون احادية الابعاد (1-D) مما سيمكننا تحويلها الى صيغة ثنائية الابعاد (2-D). المرحلة الثانية في التصنيف تتطلب تطبيق التحويلات الخليطة، تحويل فورير ثنائي الابعاد يطبق على الإشارة ثنائية الأبعاد ثم تحويل رادون ثم تحويل فورير المعكوس احادي البعد. بعد ذلك تم استخدام تحويل المويجي المتقطع في النموذج الأول، بينما تم استخدام التحويل الدائري المتعدد في النموذج الثاني. المرحلة النهائية تتضمن استخدام تحويل الزمن الديناميكي لغرض التمييز بين الكلمات. أربعون كلمة عربية مسجلة بخمسة عشر زمن مختلف في الاستوديو بواسطة متكلم واحد استخدمت كقاعدة بيانات في هذا العمل. أداء كل الطرق المستخدمة تم تحليلها وتقييمها بواسطة الحاسوب باستخدام لغة MATLAB (2010a) . إن دقة تمييز الكلام في النموذج الأول تساوي (91% and 89%) عندما استعمل التحويل المويجي المتقطع نوع Db4 وDb1 على التوالي بينما كانت الدقة في النموذج الثاني بين (87%-93%) عندما استخدمت تسعة أحزمة مختلفة من التحويل الدائري المتعدد.

الكلمات الرئيسية: التحويلات الخليطة، تحويل رادون، التحويل المويجي المتقطع ، تحويل الزمن الديناميكي، التحويل الدائري المتعدد

1. INTRODUCTION

Speech recognition is the process to recognize speech utters by a speaker. For human beings, it is a natural and simple task. However, it is an extremely complex and difficult job to make a computer respond to even simple spoken command. Speech recognition becomes a challenging task to create an intelligent recognizer that emulates a human being's ability in speech perception under all environments (Rabiner, 1993).

Many research works have been done the recognition of Arabic words. These studies include the use of neural networks and dynamic time warping. Mutasher, 2010, presented different models for speech recognition based on artificial neural network (ANN) and dynamic time wrapping (DTW) algorithm, and in each model, two transformation methods namely, discrete wavelet transform and slantlet transform are used to extract features from speech signal. Qassim, 2006, developed speech feature extraction using a hybrid technique based on discrete wavelet transform which is applied to each Arabic phoneme for single words. He proposed a technique for training and recognition tasks includes the use of feed-forward back propagation neural network.

Speech recognition consists of several stages, that the most significant of them are the feature extraction and recognition stages. Therefore, several feature extraction techniques are evaluated based on discrete wavelet transform (DWT), radon transform, and discrete multicircularlet transform (DMCT).

In this paper, isolated Arabic word recognition system is proposed based on different mixed transform techniques and DTW algorithm as a decision network.

2. ISOLATED WORDS DATABASE

A database is created for Arabic language using single speaker. Each word is repeated 15 times. We have used forty different isolated words for creating the database. The samples stored in the database are recorded by using a high quality studio-recording microphone at a sampling rate of 8 KHz. Recognition has been made on these 40 isolated spoken words under the same configuration. Our database consists of a total of 600 utterances of the spoken words.

The spoken words are preprocessed, numbered, and stored in appropriate classes in the database. The general block diagram of the speech recognition system is shown in **Fig.1**. Basic speech recognition system includes three stages: the preprocessing stage, the feature extraction stage, and the classification stage.

a. Sampling: the speech signals are sampled to convert it from analog to digital.

b. Framing and windowing: at this stage the speech signal is blocked in frames of N sample. Since we deal with speech signal, which is non-stationary signal, the framing process is essential to deal with frame not with the original signal. After this stage the speech signal has many frames and the number of frames depends on the number of samples for each word, so all length of utters must be resized into a length which is agreed with the proposed length. The main reason behind choosing proposed length, is to get high performance of the algorithms, to get best feature extraction coefficient and also to convert the process from 1D into the 2D process matrix. In this proposed system, after many studies and tests on different data (words), and different lengths, it is found that the suitable (proposed) length of all words is

256, and this choice comes from noting all length of words and finding that this length is very appropriate for all. Windowing includes multiplying each frame of the word by the hamming window; the advantage of the multiplication is to minimize the signal discontinuities at the beginning and the end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and the end of each frame. It is clear that the hamming window does not go to zero at the extremes (Mutasher, 2010). In this paper, a mixed transform was used to extract the features from the speech signal, while the DTW algorithm was used to recognize the input speech.

3. MIXED TRANSFORM

Transform techniques have proved invaluable in signal analysis and in coding for efficient transmission and storage of signal data. The goal of a transform is to represent as much of the signal information in few transform coefficients as possible. However, a particular transform is only efficient for representing signals which are of the same class as the basis functions of the transform. However, a signal may be represented efficiently only if the basis functions of the selected transform are similar in structure to the signal. Since signals such as speech and images are highly dynamic, consisting of regions with various combinations of narrow and broadband components , a single transform with fixed basis functions is rarely optimal for representing such signals (Albert, 1999).

In this work, the feature extraction of the speech signal will be done using two different mixed transforms, as explained below.

3.1 Radon Transform

The Radon transform was utilized for DSP purposes, in a novel procedure known as Finite Radon Transform (FRAT), which converts a matrix of data into a set of independent projections, and it can be reversed to retrieve the original data from the set of projections. This objective is carried out by applying 2D FFT on the matrix, to obtain the frequency-domain version of the data, then reordering the matrix elements through an optimum ordering procedure. Optimum ordering algorithm, which states that, the output matrix always contains principal directions, which are: the first column, the main diagonal, the first row, and the reverse diagonal, respectively. For example 4*4 matrix the optimum ordering will be shown below (Abdulwahid, 2010):

$$\begin{bmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 6 & 5 & 8 & 10 & 7 \\ 3 & 11 & 9 & 0 & 0 & 0 \\ 4 & 16 & 13 & 14 & 12 & 15 \end{bmatrix}$$

3.2 Discrete Wavelet Transform

The DWT can be defined as the process of decomposing a signal or function into an expansion in terms of a basis function, usually called mother wavelet, from which two types of filters, low-pass and high-pass, can be generated. These filters can be arranged in a tree structure, called a filter bank, whose outputs will be separate signals, which stand for the signal content in separate bands of frequency.

Fig. 2 illustrates a two-band filter bank, which consists of a high-pass (details) filter $h_1(-n)$, and a low-pass (approximation) filter $h_0(-n)$. The outputs d_j and c_j are found by the following convolution sums (Trivedi, 2011),

$$c_j(k) = \sum_m h_0(m-2k)c_{j+1}(m) \qquad (1)$$

$$d_j(k) = \sum_m h_1(m-2k)c_{j+1}(m) \qquad (2)$$

Notice how the convolved filter sequence jumps by two elements instead of one, due to the down sampler (Decimator) step, which takes a signal $x(n)$ as an input and produces an output of $y(n) = x(2n)$.

Further decomposition on the approximation coefficients c_j, resulting in the new outputs d_{j-1} and c_{j-1} (Trivedi, 2011).

Most of the energy of the speech signal lies in the lower frequency bands. The other sub-bands contain more detailed information of the signal and they are discarded, since the frequency band covered by these levels contains much noise and less necessary for representing the approximate shape of the speech signal. Hence take the approximation a3 and discard (d1, d2, d3). **Fig.3** shows 3-levels DWT.

3.3 Multiwavelet Transform

As in the scalar wavelet case, the theory of multiwavelets is based on the idea of multiresolution analysis , analyzing the signal at different scales or resolutions. The difference is that multiwavelets have several scaling functions. The standard multiresolution has one scaling function $\phi(t)$ (Ibraheem, 2010). The multiwavelet two-scale equation resemble those for scalar wavelets:

$$\Phi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} H_k \Phi(2t-k) \qquad (3)$$

$$\Psi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} G_k \Psi(2t-k) \qquad (4)$$

One famous multiwavelet filter is the GHM filter proposed by Geronimo, Hardin, and Massopust (Geronimo, 1994). Their system contains the two scaling functions and $\Phi_2(t)$ and the two wavelets $\Psi_1(t)$ and $\Psi_2(t)$ (Strela, 1998). According to equations 3 and 4, the GHM two scaling and wavelet functions satisfy the following two-scale dilation equations:

$$\begin{bmatrix} \psi_1(t) \\ \psi_2(t) \end{bmatrix} = \sqrt{2} \sum_k G_k \begin{bmatrix} \psi_1(2t-k) \\ \psi_2(2t-k) \end{bmatrix} \qquad (5)$$

$$\begin{bmatrix} \phi_1(t) \\ \phi_2(t) \end{bmatrix} = \sqrt{2} \sum_k H_k \begin{bmatrix} \phi_1(2t-k) \\ \phi_2(2t-k) \end{bmatrix} \qquad (6)$$

Where H_k for GHM system are four scaling matrices H_0, H_1, H_2, and H_3.

$$H_0 = \begin{bmatrix} \frac{3}{5\sqrt{2}} & \frac{4}{5} \\ -\frac{1}{20} & -\frac{3}{10\sqrt{2}} \end{bmatrix}, H_1 = \begin{bmatrix} \frac{3}{5\sqrt{2}} & 0 \\ \frac{9}{20} & \frac{1}{\sqrt{2}} \end{bmatrix},$$
$$H_2 = \begin{bmatrix} 0 & 0 \\ \frac{9}{20} & -\frac{3}{10\sqrt{2}} \end{bmatrix}, H_3 = \begin{bmatrix} 0 & 0 \\ -\frac{1}{20} & 0 \end{bmatrix} \qquad (7)$$

Also, G_k for GHM system are four wavelet matrices G_0, G_1, G_2, and G_3.

$$G_0 = \begin{bmatrix} -\frac{1}{20} & -\frac{3}{10\sqrt{2}} \\ \frac{1}{10\sqrt{2}} & \frac{3}{10} \end{bmatrix}, G_1 = \begin{bmatrix} \frac{9}{20} & -\frac{1}{\sqrt{2}} \\ -\frac{9}{10\sqrt{2}} & 0 \end{bmatrix}, \qquad (8)$$
$$G_2 = \begin{bmatrix} \frac{9}{20} & -\frac{3}{10\sqrt{2}} \\ \frac{9}{10\sqrt{2}} & -\frac{3}{10} \end{bmatrix}, G_3 = \begin{bmatrix} -\frac{1}{20} & 0 \\ -\frac{1}{10\sqrt{2}} & 0 \end{bmatrix}$$

3.4 Discrete MultiCircularlet Transform

The GHM characteristics offers a combination of orthogonality, symmetry, and compact support, which cannot be achieved by any scalar wavelet basis except for the Haar basis, its basis functions are exploited to generate more efficient basis functions for compression purposes where the number of zeros is as large as possible.

The GHM basis functions are exploited to generate more efficient basis functions for compression purposes. This can be achieved through the following steps (Alubady, 2009).

1- Taking the 2-D convolution for each basis function matrix with itself to generate new basis functions. This is done as follows:-

a. Write the polynomial representation of the matrix, e.g. for matrix $G_1(n_{1,}n_2)\,\&\,G_2(n_{1,}n_2)$:

$$G_1 = p\downarrow \overset{q \longrightarrow}{\begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}} \tag{9}$$

$$G_2 = p\downarrow \overset{q \longrightarrow}{\begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix}} \tag{10}$$

The polynomial representation for eq. (9) Will Be

$$G_1(p,q) = a_1 + a_3 p + a_2 q + a_4 pq \tag{11}$$

The polynomial representation for eq. (10) Will Be

$$G_2(p,q) = b_1 + b_3 p + b_2 q + b_4 pq \tag{12}$$

b. Use table lookup method to find the convolution result as follows:

	a_1	a_3p	a_2q	a_4pq
b_1	a_1b_1	a_3b_1p	a_2b_1p	a_4b_1pq
b_3p	a_1b_3p	$a_3b_3p^2$	a_2b_3pq	$a_4b_3p^2q$
b_2q	a_1b_2q	a_3b_2pq	$a_2b_2q^2$	$a_4b_2pq^2$
b_4pq	a_1b_4pq	$a_3b_4p^2q$	$a_2b_4pq^2$	$a_4b_4p^2q^2$

The output polynomial representation is equal to:

$$a_1b_1$$
$$a_3b_1p + a_1b_3p$$
$$a_2b_1q + a_3b_3p^2 + a_1b_2q$$
$$a_4b_1q + a_2b_3pq + a_3b_2pq + a_1b_4pq$$
$$a_4b_3p^2q + a_2b_2q^2 + a_3b_4p^2q$$
$$a_4b_2pq^2 + a_2b_4pq^2$$
$$a_4b_4p^2q^2$$

c. Arrange the result in matrix form to which gives:

$$\begin{matrix} a_1b_1 & a_2b_1 + a_1b_2 & a_2b_2 \\ a_3b_1 + a_1b_3 & a_2b_3{+}a_3b_2{+}a_1b_4{+}a_4b_1 & a_4b_2 + a_2b_4 \\ a_3b_3 & a_4b_3 + a_3b_4 & a_4b_4 \end{matrix}$$

d. Fold the 3rd column on the 1st column & next folding the 3rd row on the 1st row.

i- Folding the 3rd column results in

$$\begin{pmatrix} a_1b_1 + a_2b_2 & a_2b_1 + a_1b_2 \\ a_1b_3 + a_3b_1 + a_2b_4 + a_4b_2 & a_1b_4 + a_2b_3 + a_3b_2 + a_4b_1 \\ a_3b_3 + a_4b_4 & a_3b_4 + a_4b_3 \\ & \end{pmatrix}$$

ii- Folding the 3rd row results in

$$\begin{pmatrix} a_1b_1 + a_2b_2 + a_3b_3 + a_4b_4 & a_1b_2 + a_2b_1 + a_3b_4 + a_4b_3 \\ a_1b_3 + a_2b_4 + a_3b_1 + a_4b_2 & a_1b_4 + a_2b_3 + a_3b_2 + a_4b_1 \end{pmatrix}$$

The individual coefficients values of these matrices are generated using the following procedure (Alubady, 2009):

1- Apply the 2-D convolution between the G’s & H’s . This can be achieved as follows:

a) compute $A_{i1} = H_i \otimes H_i$

b)compute $B_{i1} = G_i \otimes G_i$

where i= 0,1,2,3

2. Now compute the 2-D Convolution between the resultant of step 1 & the G's & H's. This can be done through the following way:

a) compute $A_{i2} = A_{i1} \otimes H_i$

b)compute $B_{i2} = B_{i1} \otimes G_i$

where i = 0,1,2,3

3. The process was repeated several times. It was found that the optimal results was at the third step.

The proposed matrix coefficients A's and B's was obtained by performing the following computations:

a) compute $A_i = A_{i2} \otimes H_i$

b)compute $B_i = B_{i2} \otimes G_i$

where i = 0,1,2,3

The proposed new multifilter bases functions which denoted by A's and B's are stated as

A_0, A_1, A_2, A_3 & B_0, B_1, B_2, B_3

The proposed basis functions are:

The A's 2x2 matrices are:

$$A_0 = \begin{bmatrix} 1.4561 & 1.4131 \\ -1.0265 & -0.9857 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1.6896 & 1.5814 \\ 1.4376 & 1.5450 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0.0977 & -0.0945 \\ 0 & 0 \end{bmatrix}, \quad A_3 = 1.0e-005 * \begin{bmatrix} 0.6250 & 0 \\ 0 & 0 \end{bmatrix} \tag{13}$$

The B's 2x2 matrices are:

$$B_0 = \begin{bmatrix} 0.0460 & 0.0343 \\ -0.0459 & -0.0342 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 2.7696 & -2.4406 \\ -2.4142 & 2.7225 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1.6610 & -1.6066 \\ 1.6581 & -1.6038 \end{bmatrix}, \quad B_3 = \begin{bmatrix} -0.0302 & -0.0302 \\ 0.0052 & 0.0052 \end{bmatrix} \tag{14}$$

Due to the good characteristics of the transformed 1-D and 2-D signal by the basis functions obtained from third convolution, it will be adopted as the new transform named "Multicircularlet Transform"(Alubady,2009).

The AB multifilter bank coefficients are 2 by 2 matrices, and during the convolution step they must multiply vectors (instead of scalars). This means that multifilter banks needs 2 input rows. This transformation is called preprocessing. The most obvious way to get two input rows from a given signal is to repeat the signal. Two rows go into the multifilter bank. This procedure is called "Repeated Row" which introduces over sampling of the data by a factor of 2 .

For computing DMWT, the transformation matrix can be written as follows:

$$W = \begin{bmatrix} A_0 & A_1 & A_2 & A_3 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & A_0 & A_1 & A_2 & A_3 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ A_2 & A_3 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & A_0 & A_1 \\ B_0 & B_1 & B_2 & B_3 & \vdots & \vdots & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & B_0 & B_1 & B_2 & B_3 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & B_0 & B_1 & B_2 & B_3 \\ B_2 & B_3 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & B_0 & B_1 \end{bmatrix}$$

where A_i , B_i are the impulse responses of the AB multifilter bank, equations 13 and 14.

The organization and statistics of multicircularlet subbands differ from the scalar wavelet case. During a single level of decomposition using a scalar wavelet transform, the 2-D signal data is replaced with four blocks, while in multifilter bank used here, two channels are corresponds for each bank, so there will be two sets of scaling coefficients and two sets of wavelet coefficients.

4. DYNAMIC TIME WARPING

The Dynamic Time Warping (DTW) is the mathematical technique which is used to find the cumulative distance along the optimum path without having to calculate the cumulative distance along all possible paths (Chapaneri,2012). In **Fig.4** consider a point i, j somewhere in the middle of both words. If this point is on the optimum path, then the constraints of the path necessitate that the immediately preceding point on the path is i-1, j or i-1, j-1 or i, j-1. These three points are associated with a horizontal, diagonal or vertical path step respectively. Let D(i, j) be the cumulative distance along the optimum path from the beginning of the word to point i, j, thus:

$$D(i,j)=\sum_{x,y=1,1}^{i,j} d(x,y) \qquad (15)$$

As there are only the three possibilities for the point before i,j it follows that:

$$D(i,j) = d(i,j) + \min [D(i-1, j) , D(i-1, j-1), D(i,j-1)] \qquad (16)$$

The best way to get to point i, j is thus to get to one of the immediately preceding points by the best way, and then take the appropriate step to i, j. The value of D(1, 1) must be equal to d(1, 1) as this point is the beginning of all possible paths. To reach points along the bottom and the left-hand side of **Fig.4** there is only one possible direction (horizontal or vertical, respectively). Therefore, starting with the value of D(l, 1), values of D(i, 1) or values of D(1, j) can be calculated in turn for increasing values of i or j. Let us assume that we calculate the vertical column, D(1, j*),* using a reduced form of eq. (15) that does not have to consider values of D(i-l, j) or D(i-1, j-1). (As the scheme is symmetrical we could equally well have chosen the horizontal direction instead). When the first column values for D(1, j) are known, eq.(16) can be applied successively to calculate D(i, j) for columns 2 to n. The value obtained for D(n,m*)* is the score for the best way of matching the two words(Holmes, 2001).

Suppose we have two time series *Q* and *C*, of length *n* and *m* respectively, where

Q = q1,q2,…,qi,…,qn

C = c1,c2,…,cj,…,cm

To align two sequences using DTW we construct a *n*-by-*m* matrix where the (it, get) element of the matrix contains the distance d (Qi, cj) between the two points *Qi* and *cj* (Typically the Euclidean distance is used, so d (Qi, cj) = (Qi - cj) 2). Each matrix element (I, j) corresponds to the alignment between the points Qi and cj.

A warping path *W* is a contiguous set of matrix elements that defines a mapping between series. The kth element of *W* is defined as $w_k = (i, j)_k$ so we have:

W = (*w*1,*w*2,….*wj* ,…,*wk*)

The best warp (the shortest) path could be found by exhaustively searching all possible paths and selecting that path having the minimum measure**.** This, however, is not a practical solution, because the number of possible paths is large and exponentially related to the size of the lattice.

If a point (Qi, cj) lies on the optimal path, then the sub path from (q1, c1) to (Qi, cj) is also locally optimized. This means that the best path from (q1, c1) to (qn, cm) can be recursively found by locally optimizing paths one grid unit with time-beginning at (q1**,** c1) and ending at (qn, cm). This is done in practice by running the recurrence relation in eq. (16) which defines the cumulative distance D (I, j) for each point , i.e. assigning a partial sum to each grid (Li Dong, 2006).

The global warp cost of the two sequences is defined as shown below:

$$GC = \frac{1}{P}\sum_{I=1}^{P} w_i \qquad (17)$$

Where w_i are those elements that belong to warping path, and p is the number of them.

There are three conditions imposed on the DTW algorithm that ensure them a quick convergence (Furtuna, 2008):

1. Monotony – the path never returns, that means that both indices I and j used for crossing through sequences never decrease.
2. Continuity – the path advances gradually, step by step; indices I and j increase by maximum 1 unit on a step.
3. Boundary – the path starts in left-down corner and ends in a right-up corner in the distance matrix.

5. PROPOSED SPEECH RECOGNITION SYSTEM

The general block diagram for the proposed speech recognition system is shown in **Fig. 5**

5.1 Speech Signal

The proposed models have been applied on forty Arabic words; these words were recorded by a microphone in the studio by one speaker and stored as files, the format of these files is wave format. The type of digital speech signal is (PCM), and the sampling rate is 8 KHz. These words are:

(اشارة، لندن، افتح، الخير، تصميم، نبيل، مربع، رازق،
رحمن، شارع، صباح، صديق، عمودي، كامل، محمد،
معلومات، نظام، وفاء، ياسين، ورود، بغداد، خشب،
ثلاثون، ابراهيم، تردد، واحد، عباس، جبل، مصعد، قادر،
زائد، قلم، خاص، محمول، سيارات، هاتف، كتاب، دفتر،
صالح، مستقبل)

5.2 Sampling

The speech signals are sampled to convert it from analog to digital. The sampling rate has been down sampled from 44 KHz to 8 KHz.

5.3 Framing

At this 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 frame not with the 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 265 samples.

5.4 Hamming Windowing

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

5.5 Feature Extraction

Since speech recognition is so affected by the feature extraction stage, therefore two different mixed transforms are used in this work.

5.5.1 The First Mixed Transform

The first mixed transform consist of applying 2-D FFT for each frame after resizing it into 16× 16 matrix, then find the best sequence of direction for each frame after that applying the 1-D IFFT then applying the DWT to the resultant matrix for each row of the frame.

Most of the energy of the speech signal lies in the lower frequency bands. The other sub-bands contain more detailed information of the signal and they are discarded, since the frequency band covered by these levels contains much noise and less necessary for representing the approximate shape of the speech signal . Hence take the approximation a3 and discard (d1, d2, d3). In the proposed work the DWT that used are Daubechies (Db1) and (Db4) type.

5.5.2 The Second Mixed Transform

In this proposed model the same preprocess and classification stages was used as in the first model, in feature extraction stage DMCT was used instead of DWT. The DMCT was applied to each frame of each version that results in 16 sub bands for each frame, each sub band treated independently to examine the recognition rate when each sub band used separately. Nine sub bands (sb1, sb2, sb3, sb5, sb6, sb7, sb9, sb10, sb11) which represent the approximation sub bands will be later represent the feature vector for the uterus and it will be ready to be used by the classifier, as it can be shown in **Fig. 6** where the shaded bands represent the detailed bands with a lowest recognition rate that can be ignored. **Fig. 7** shows the recognition steps using the second mixed transform.

5.6 Dynamic Time Warping Algorithm

After taking the mixed transform for all versions and represent each version by one feature vector, these feature vectors are different even in the same word, so these data are suitable to enter DTW classifier. The DTW algorithm allows a non-linear warping alignment of one signal to another by minimizing the distance between the two. This warping between two signals can be used to determine the similarity between them and thus it is very useful feature for recognition.

The algorithm of DTW for classification of words is explained as follows:

a. Take fifteen version for each word and divide these versions, ten versions of each word (10*40=400 version) for basing, and five versions (5*40=200 version) for testing.

b. Calculate the distance value between each version from test with each version from base for all words.

The calculation of distance is done by finding the warping path between two versions and then calculate the global warp cost of the two versions defined by eq. (17)

6. RESULTS

In this work, the proposed systems have been applied on forty Arabic words. The number of versions of each word have been divided into two parts:

a. One part of these versions used for basing called “basing versions’’, ten versions have been taken for each word.

b. The other part used for testing called ‘’testing versions’’ five versions have been taken for each word. The test versions are tested on the DTW and their resultant error is used to give the measure of the generalization ability of this algorithm.

The proposed speech recognition system is done by different models two mixed transforms are used to recognize the speech signal. The first one consists of cascaded mixed transforms including (2D FFT, Radon transform, 1D IFFT) and followed by the DWT. The second mixed transform use the same first three stages as the first one, but differs in the last transform where DMCT is used. To compare the performance of each one of them, the accuracy or recognition rate has been computed as follows:

$$Accuracy = \frac{Total\ number\ of\ correct\ recognition}{Total\ number\ of\ testing\ version}\ X100\%$$

Table 1 shows the comparison between the accuracy of different recognition system, while **Fig.8** shows the percentage of accuracy for each recognition system.

7. CONCLUSIONS

In this work, two mixed transforms are used to recognize the speech signal. The first one consists of cascaded mixed transforms including (2D FFT, Radon transform, 1D IFFT) and followed by the DWT that gave a recognition rate of 91%. The second mixed transform use the first as the first one, but differs in the last transform where DMCT is used, that gave a recognition rate of 93%. The mixed transform succeeded as a technique for combining features of speech recognition, since the coefficients pass through a cascaded transforms which enhance its low frequency component. The proposed mixed transform is a combination of multicircularlet transform with a 2D-FFT, Radon transform and 1D-IFFT to achieve better coefficient decomposition. The DMCT offers a good distribution of the signal in the frequency - spatial domain. It was shown that, this will result in a better decomposition of the coefficients, so it can be used as the extraction stage in the proposed speech recognition. DTW algorithm was used as a classifier to the features that are extracted by the mixed transform; this algorithm is a powerful tool to find the minimum distance between features of the same word, and there is no need to equalize their lengths. With all these advantages of using the proposed algorithm, the disadvantage of using the proposed mixed transform is the execution time because the complex computation of the proposed mixed transforms that take more time of executing a single transform alone.

8. REFERENCES

Abdulwahid ,H. " **Design And Simulation of A Multidimensional Radon-Based OFDM System** ", M.Sc. Thesis, Nahrain University, Communications Engineering Department, June 2010.

Albert P. B. and Mikhae1 B. W. **"A Survey of Mixed Transform Techniques for Speech and Image Coding",** IEEE Xplor, pp.106-109,1999.

Alubady,I. **" A Proposed Multicircularlet Mixed Transform and Its Application for Image Compression"**, M.Sc. Thesis, University of Baghdad, Electrical Engineering Department, 2009.

Chapaneri, S.V. **" Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping",** International J. of Computer Applications, Vol. 40, No.3, pp.6-12, February 2012.

Furtuna, T.F., **"Dynamic Programming Algorithms in Speech Recognition",** Revista Informatica Economică, Vol.46, No.2, pp. 94-99, Bucharest, 2008.

Geronimo, J., Hardin, D. & Massopust, P., **"Fractal Function and Wavelet Expansion Based on Several Functions"**, J. Approx. Theory, Vol. 78, PP. 373-401, 1994.

Holmes, J. and Holmes, W., **"Speech Synthesis and Recognition"**, Second Edition, London and New York, 2001.

Ibraheem, A. K**., "Image Reconstruction Using Hybrid Transform"**, M.Sc. Thesis, University of Baghdad, Electrical Engineering Department, 2010.

Li Dong, X., Kui Gu, C. & Ou Wang, **"A Local Segmented Dynamic Time Warping Distance Measure Algorithm for Time Series Data Mining",** Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, pp.1247-1252, August 2006.

Mutasher, S., **“A Multi Transform Based Dynamic Time Warping Isolated Word Speech Recognition System’’**, M.Sc. Thesis, University of Baghdad, Electrical Engineering Department, April,2010.

Qassim, A**., “Arabic Phonemes Recognition Using Hybrid Technique",** M.Sc. Thesis, University of Technology, Electrical and Electronic Engineering Department, January 2006.

Rabiner, L. and Juang, B. H., **"Fundemantals of Speech Recognition",** Prentice –Hell , New Jercy, 1993.

Strela, V. and Walden, A.T. **" Orthogonal and Biorthogonal Multiwavelets for Signal Denoising and Image Compression"** Proc. SPIE, 3391, pp.96-107, 1998.

Trivedi,N., Kumar,V., and Singh,S., **"Speech Recognition by Wavelet Analysis",** International J. of Computer Applications, Vol.15, No.8, pp.27-32, February 2011.

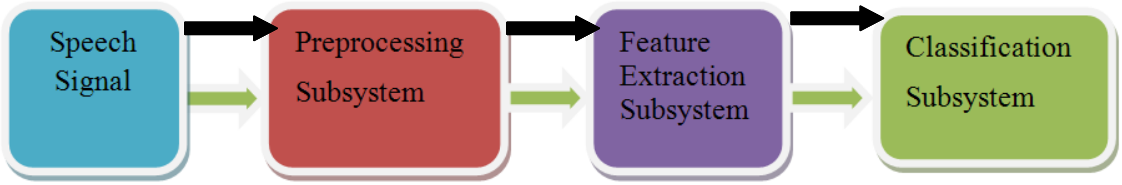


Fig.1: General block diagram of the speech recognition system

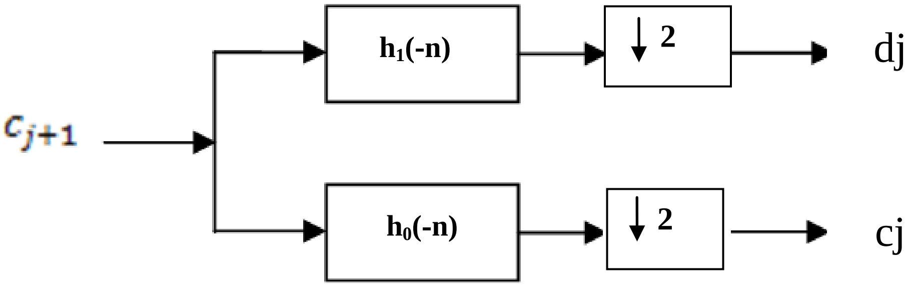


Fig.2:- Two-band filter bank **.**

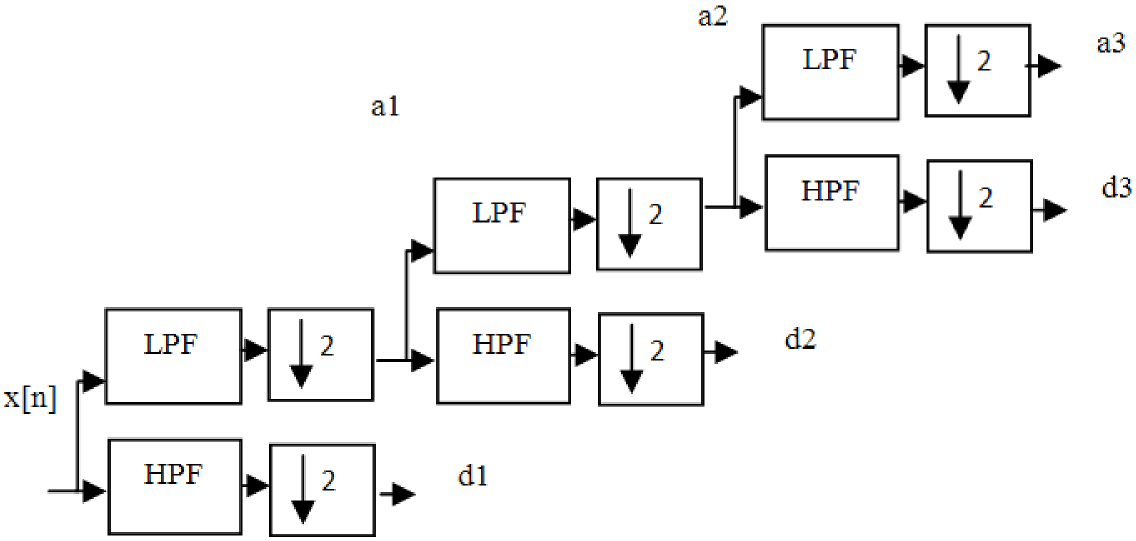
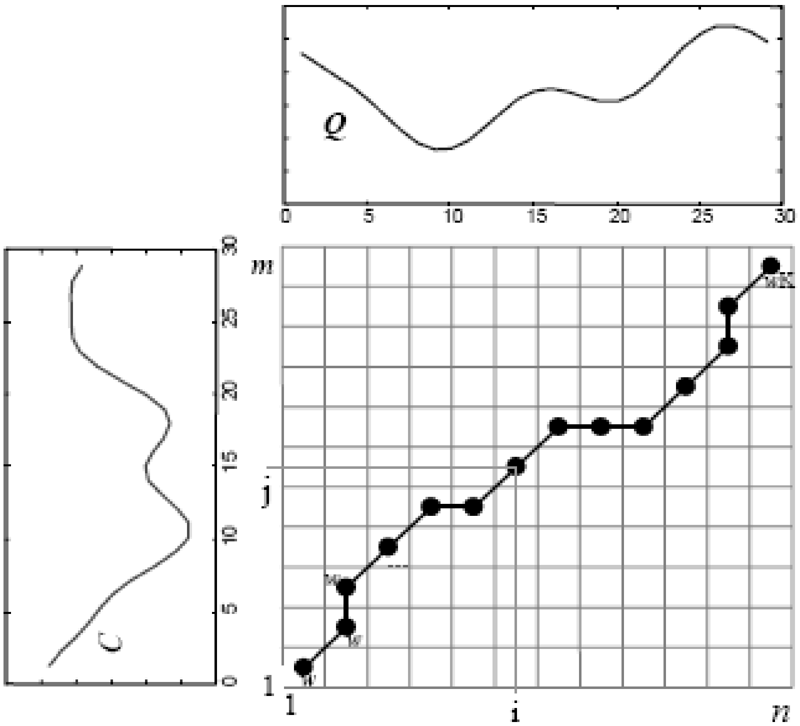


Fig.3:- 3-Levels of DWT



time series *Q* of length *n*, Q = q1,q2,…,qi,…,qn
time series *C* of length *m*, C = c1,c2,…,cj,…,cm
Fig.4:- An example of warping path

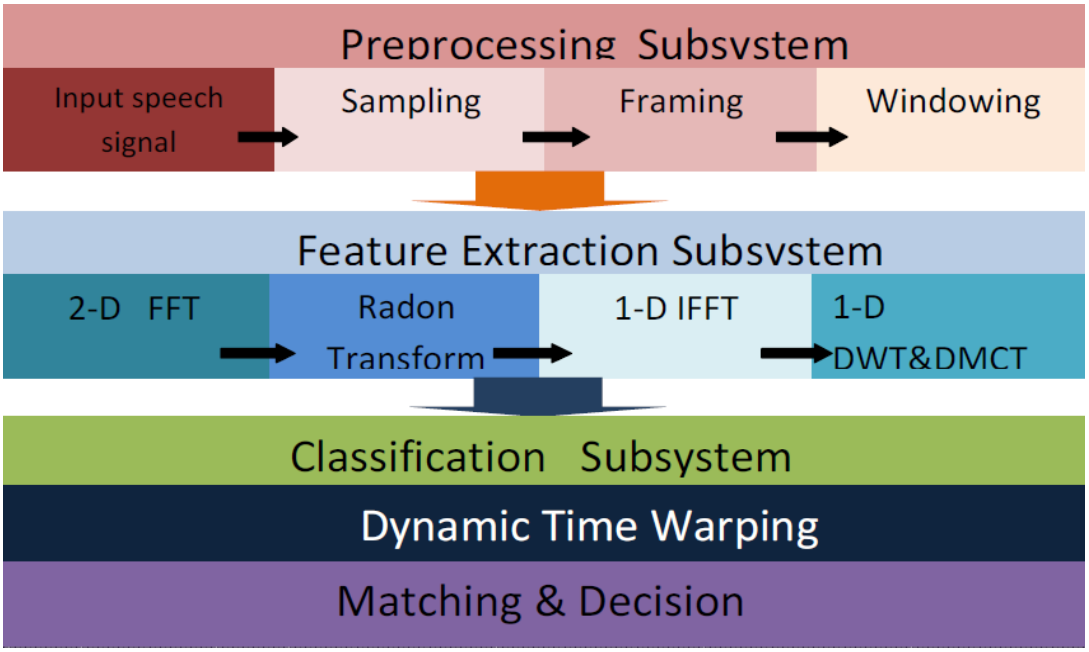


Fig.5:- Proposed speech recognition system

L1L1	L1L2	L1H1	L1H2
L2L1	L2L2	L2H1	L2H2
H1L1	H1L2	H1H1	H1H2
H2L1	H2L2	H2H1	H2H2

Fig.6:- The resultant MCT bands

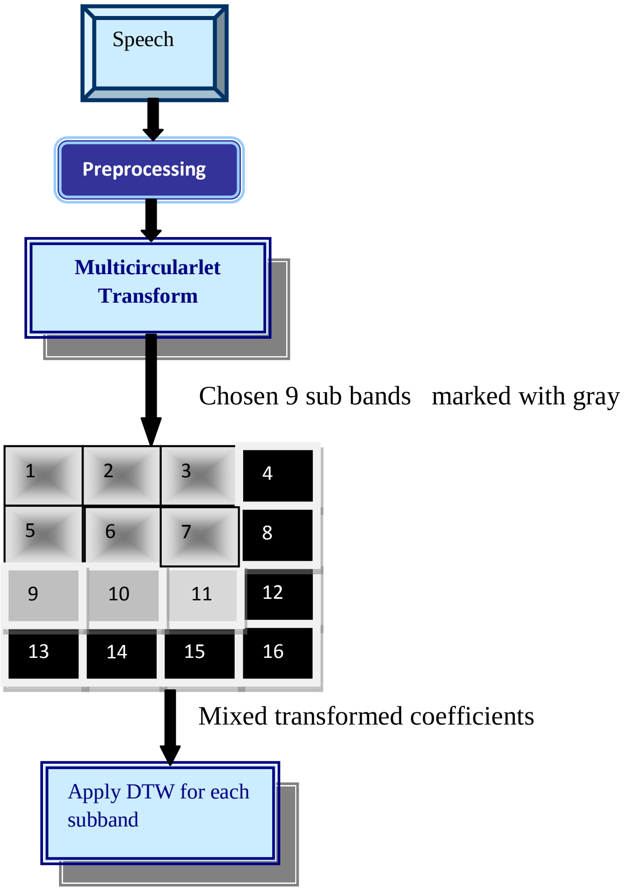


Fig.7:- The recognition steps using second mixed transform

I

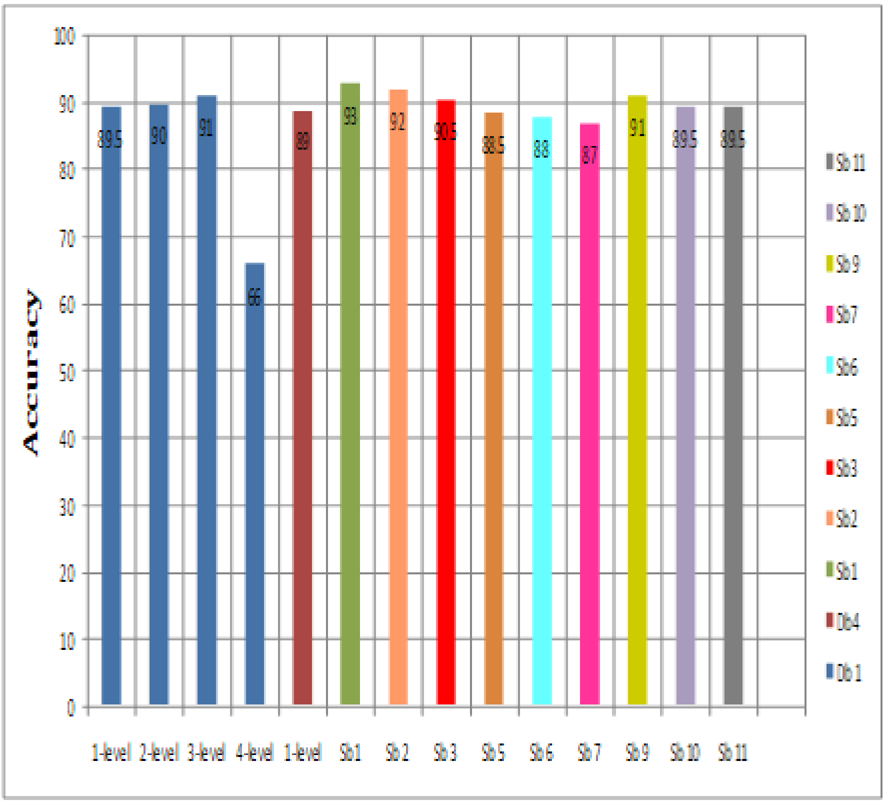


Fig.8:- The percentage of accuracy of each recognition system

Table 1: Comparison between the accuracy of different recognition system

Classification system	Total number of testing version	Total number of correct recognition	Accuracy
First model			
1-level Db1	200	179	89.5 %
2-level Db1	200	180	90 %
3-level Db1	200	182	91 %
4-level Db1	200	132	66 %
1-level Db4	200	178	89%
Second model			
sb1	200	186	93 %
sb2	200	184	92 %
sb3	200	181	90.5%
sb5	200	177	88.5%
sb6	200	176	88%
sb7	200	174	87%
sb9	200	182	91%
sb10	200	179	89.5%
sb11	200	179	89.5 %