Biorthoganal Wavelet Packets and Mel Scale Analysis for Automatic Recognition of Arabic Speech via Radial Basis Functions

JALAL KARAM
Alfaisal University
Faculty of Science and General Studies
Riyadh
KINGDOM OF SAUDI ARABIA
email: jkaram@alfaisal.edu Website: www.alfaisal.edu/faculty/cv/jalal

Abstract: In this paper, a Neural Network (NN) approach for the recognition of the Arabic digits is presented. The two phases of training and testing in a Radial Basis Functions (RBF) type network is described. Biorthogonal Wavelets are constructed and used for analysis of generated subwords of the digits. This approach decomposes spoken Arabic digits based on the acoustical information contained within the speech signals. The procedure locates the boundaries between subwords by finding the peaks in the function representing the spectral changes between consecutive speech frames. The Frame-based energy parameters derived from a Wavelet Packet Scale (WPS) are used in deriving the Spectral Variation Function (SVF). Three Biorthogonal wavelets are used as analyzing functions and their performances are compared with that of their Orthogonal counterpart and with that of the traditional Fourier based Mel scale approach.

Key–Words: Biorthogonal Wavelets, Radial Basis Functions, Recognizing Arabic Speech.

1 Introduction

A two dimensional signal processing tool that remedies problems arising from time frequency domain methods such as trade off in time frequency resolutions and limitations in analyzing non-stationary signals is the time-scale representation. The Wavelet Transform (WT) accomplishes such representation. It partitions the time-frequency plane in a non-uniform fashion and shows finer frequency resolution than time resolution at low frequencies and finer time resolution than frequency resolution at higher frequencies. This type of transform decomposes the signal into different frequency components, and then analyzes each component with a resolution that matches its scale [9]. The Continuous Wavelet Transform (CWT) [4] of a signal $x(t)$, is given by :

$$CWT_{(a,b)}(x(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right)dt \quad (1)$$

Where a and b are the real numbers that represent the scale and the translation parameter of the transform respectively. The function $\psi(t)$ is called the mother wavelet and has to have the following two properties:

(1) $\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty$. This is equivalent to having $\psi(t) \in L^2(\Re)$ the space of finite energy functions.

(2) $\int_{-\infty}^{\infty} \psi(t)dt = 0$. This is equivalent to having the Fourier Transform of $\psi(t)$ null at zero (i.e., $\psi(t)$ has no dc components).

One can interpret this integral operation in two ways [3]:

(1) It evaluates the inner product or the cross correlation of $x(t)$ with the $\psi(t/a)/\sqrt{a}$ at shift b/a. Thus it evaluates the components of $x(t)$ that are common to those of $\psi(t/a)/\sqrt{a}$. Thus it measures the similarities between $x(t)$ and $\psi(t/a)/\sqrt{a}$.

(2) It is the output of a bandpass filter of impulse response $\psi(-t/a)/\sqrt{a}$ at b/a of the input signal $x(t)$. This is a convolution of the signal $x(t)$, with an analysis window $\frac{1}{\sqrt{a}}\psi(t/a)$ that is shifted in time by b and dilated by a scale parameter a.

The second interpretation can be realized with a set of filters whose bandwidth is changing with frequency. The bandwidth of the filters is inversely proportional to the scale a which is inversely proportional to frequency. Thus, for low frequency we obtain high spectral resolution and low (poor) temporal resolution. Conversely, (This is where this type of representation is most useful) for high frequencies we obtain high temporal resolution that permits the wavelet transform to zoom in on singularities and detect abrupt changes in the signal [9]. This leads to a poor high frequency spectral resolution. The Discrete Wavelet Transform and the Fourier Transform are modified versions of the Continuous Wavelet Transform. They can be derived from the CWT for specified values of a and b. For example, if the mother wavelet $\psi(t)$ is the exponential function e^{-it} and $a = \frac{1}{w}$ and b=0 then, the CWT is reduced to the traditional Fourier Transform with the scale representing the inverse of the frequency [29]. The advantages that this new representation has over the STFT can be noticed in its efficiency in representing physical signals since it isolates transient information in a fewer number of coefficients and also in overcoming the time frequency trade off induced by STFT [9]. The properties of the CWT for real signals include: linearity, scale invariant, translation invariant, real and has an inverse. For a detailed discussion about the properties of the CWT and their proofs, refer to [4]. Some of the applications of the CWT in speech processing include:

(1) Analysis, synthesis and processing of speech and music sound [17],

(2) Analysis of sound patterns [18],

(3) Formant tracking [8],

(4) Speech recognition [7] [10] [11] [12].

(5) Speech compression [13] [27].

1.1 The Biorthogonal Analyzing Function Bior3.9

A two channels filter bank has a low-pass and a high-pass filter in the decomposition (analysis) phase. Let H_0 and G_0 denote the low-pass filter coefficients and the high-pass filter coefficients respectively. Given the coefficients of H_0, it is shown in [25] and [26] that the coefficients of the filters H_1, G_0 and G_1 that lead to orthogonality can easily be derived from the coefficients of H_0. Biorthogonal filter banks produce biorthogonal wavelets. This calls for a new scaling function $\tilde{\phi}(t)$ and a new wavelet function $\tilde{w}(t)$. Here, one needs the conditions: $H_1(z)$ = $H_0^{-1}(z)$ and $G_1(z)$ = $H_1^{-1}(z)$ [28]. The wavelet filters for analysis banks are derived [25] from the scaling filters using the relations:

$$h_1 = (-1)^{n+1} g_0(n) \tag{2}$$

$$g_1 = (-1)^n h_0(n) \tag{3}$$

The analysis scaling and wavelet equations thus become:

H_0	H_1	$\tilde{H}_0$	$\tilde{H}_1$
-0.0007	0	0	-0.0007
0.0020	0	0	-0.0020
0.0051	0	0	0.0051
-0.0206	0	0	0.0206
-0.0141	0	0	-0.0141
0.0991	0	0	-0.0991
0.0123	0	0	0.0123
-0.3202	0	0	0.3202
0.0021	-0.1768	0.1768	0.0021
0.9421	0.5303	0.5303	-0.9421
0.9421	-0.5303	0.5303	0.9421
0.0021	0.1768	0.1768	-0.0021
-0.3202	0	0	-0.3202
0.0123	0	0	-0.0123
0.0991	0	0	0.0991
-0.0141	0	0	0.0141
-0.0206	0	0	-0.0206
0.0051	0	0	-0.0051
0.0020	0	0	0.0020
-0.0007	0	0	0.0007

Table 1: Coefficients of the filters for bior3.9.

$$\widetilde{\phi(t)} = \sum_{0}^{\tilde{N}} 2h_0^r(k)\tilde{\phi}(2t-k) \tag{4}$$

$$\widetilde{w(t)} = \sum_{0}^{N} 2g_0^r(k)\tilde{\phi}(2t-k) \tag{5}$$

where h_0^r and g_0^r are the reverse of the original filters h_0 and g_0 respectively. The construction of $\phi(t), w(t), \tilde{\phi}(t)$ and $\tilde{w}(t)$ starts with imposing the biorthogonality conditions on the filters. The lowpass analysis coefficients $h_0^r(k)$ are double shift biorthogonal to the lowpass synthesis coefficients $h_1(k)$:

$$2\sum h_1(k)h_0^r(k+2n) = \delta(n) \tag{6}$$

$$2\sum g_1(k)g_0^r(k+2n) = \delta(n) \tag{7}$$

And the highpass filter is biorthogonal to the lowpass filter:

$$\begin{array}{l} \sum h_1(k)g_0^r(k+2n) = 0 \quad and \\ \sum g_1(k)h_0^r(k+2n) = 0 \end{array} \tag{8}$$

Figure 1 shows the frequency responses of the decomposition and reconstruction filters and, the decomposition and reconstruction scaling and wavelet functions of the biorthogonal (bior3.9) [16] are displayed in Figure 2 and Figure 3. This wavelet is smooth, has a linear phase and short length filters. Also, Table 1 displays the coefficients of the lowpasses and highpasses filters of bior3.9.

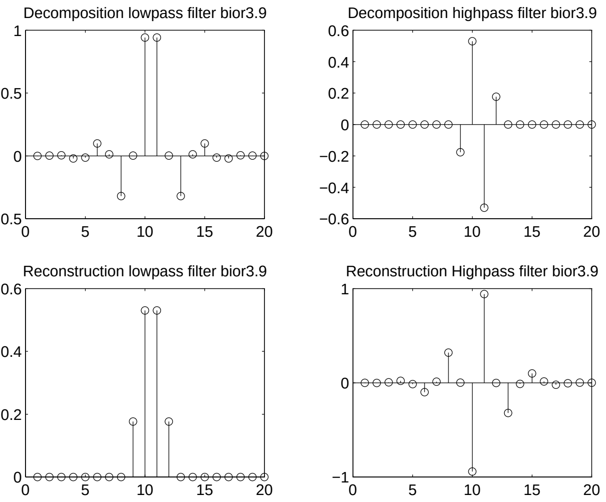

Figure 1: Impulse response for the construction and decomposition filters.

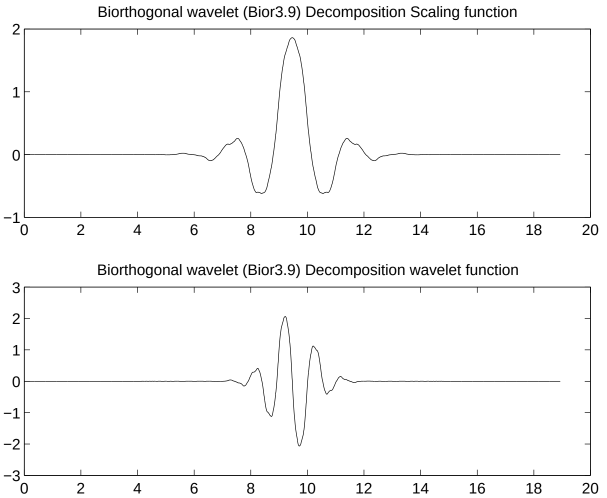

Figure 2: The Decomposition Scaling and Wavelet Functions for the Bior 3.9 Wavelet.

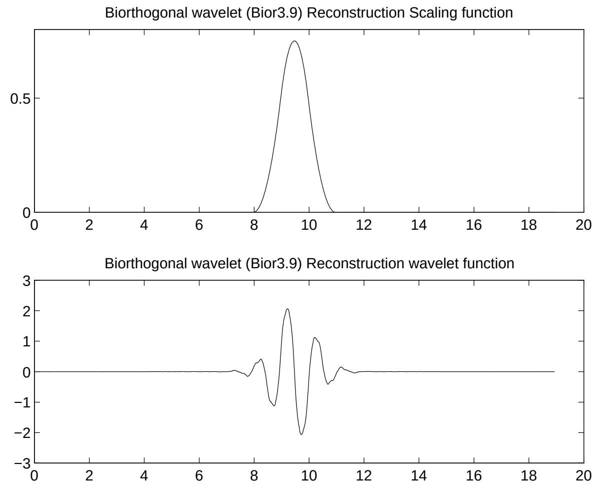

Figure 3: The Reconstruction Scaling and Wavelet Functions for the Bior 3.9 Wavelet.

Selected Nodes	Bandwidth In Hz	Centers WPS Bands	Band in Hz	Centers Mel Bands
[7,1]	78	117	(78-156)	100
[7,3]	78	195	(156-234)	200
[7,2]	78	273	(234 - 312)	300
[7,7]	78	351	(312 -390)	400
[7,6]	78	429	(390 -468)	500
[7,4]	78	507	(468-564)	600
[7,5]	78	585	(564 - 625)	700
[7,15]	78	663	(625 - 703)	800
[7,14]	78	741	(703 -781)	900
[7,12]	78	819	(781 -859)	1000
[7,13]	78	897	(859 -937)	1149
[7,8]	78	975	(937- 1015)	1320
[7,9]	78	1053	(1015 -1093)	1516
[6,5]	156	1171	(1093 -1250)	1741
[5,7]	312	1406	(1250 -1562)	2000
[5,6]	312	1718	(1562-1875)	2297
[4,2]	625	2187	(1875 -2500)	2639
[4,7]	625	2812	(2500 -3125)	3031
[4,6]	625	3437	(3125 - 3750)	3482
[3,2]	1250	4375	(3750 -5000)	4000

Table 2: The WPS Nodes Selection.

2 Wavelet Packet Scale

Experiments in human perception have shown that frequencies of a complex sound within a certain bandwidth of some nominal frequency cannot be individually identified. But if one of the components of this sound falls outside of the bandwidth then it could be distinguished. This bandwidth is commonly referred to as the critical bandwidth [19]. In this section we explore the flexibility of the Wavelet Packet analysis for the construction of a Mel-like scale for speech perception along with its banwidth and centers of frequencies. We drive a relationship between these centers of frequencies and the Wavelet Packet tree and construct a Wavelet Packet Scale (WPS).

Accurate representation of speech signals is the primary goal of digital speech processing. The Wavelet Packet (WP) representations provide a local time-spectral description which reveals the non-stationary nature of speech. This is a direct consequence of the the WP capability of arbitrary multiresolution time-spectral decomposition of speech [6]. The WP analysis is normally implemented using perfect reconstruction filter banks where only the lowpass filter was iterated. This leads to a one sided tree decomposition.

At each node of the WP tree we have the option to analyze the signal with the lowpass or the highpass wavelet or both. Two familiar options are the logarithmic tree of the DyWT with lowpass iteration only and the complete tree analogue to the STFT [25]. The computation scheme for wavelet packets generation is easy in the case of an orthogonal wavelet [16]. The idea is to start with two filters of length 2N denoted by h(n) and g(n) that correspond to a wavelet. They are the reversed versions of the lowpass decomposition filter and the highpass decomposition filter divided by

$\sqrt{2}$ respectively [16]. By induction, one can define the following two sequences of functions ($W_n(x)$, n = 0,1,2,...) by:

$$W_{2n}(x) = 2\sum_{k=0}^{2N-1} h(k)W_n(2x-k) \tag{9}$$

$$W_{2n+1}(x) = 2\sum_{k=0}^{2N-1} g(k)W_n(2x-k) \tag{10}$$

where $W_0(x) = \phi(x)$ is the scaling function and $W_1(x) = \psi(x)$ is the wavelet function. The $W_n(x)$ are called the Wavelet Packets functions.

2.1 Physical Interpretation

Consider the three-indexed family of analyzing functions:

$$W_{j,n,k}(x) = 2^{-j/2}W_n(2^{-j}x-k) \tag{11}$$

Where n is a positive integer and k and j are integers representing the time-localization and the scale parameter respectively [16]. For fixed values of the parameters k and j, the physical interpretation of $W_n(x)$ is that it analyzes the fluctuations of an input signal around the time $2^{-j}K$ at the scale 2^{-j} and at various frequencies for the different values of the parameter n [16]. The drawback in the way that the functions $W_n(x)$ are generated is that, for $m > m'$ it does not imply that $W_m(x)$ oscillates more than $W_{m'}$. To restore the property that the frequency increases with the order one should sort the functions $W_n(x)$ in their increasing order of frequency [16].
The (WPS) needs a level seven wavelet tree decomposition,

3 Speech Recognition Approaches

As classified by Rabiner and Juang [21], the three popular approaches in speech recognition are:

(1) The acoustic phonetic approach,

(2) The pattern recognition approach,

(3) The artificial intelligence approach.

Since the second approach is chosen for the recognition systems introduced in this paper, it is discussed in details in the next subsection while a brief description of the first is presented. The third approach is included for completeness.

3.1 The Acoustic-Phonetic Approach

The acoustic-phonetic approach [21] and [22], is based on the idea of phonemes which constitute the basic sound units of a verbal language. Unlike the letters in an alphabet, phonemes have distinct pronunciations. For example, the diphthong (aw) is always pronounced as (ou) in the word ‘out’. An acoustic-phonetic speech recognizer first processes a speech signal in short intervals and extracts a set of features for each of these interval. The features can include one or a combination of the parameters such as Formant, pitch, or energy. In a process known as segmentation and labeling, the recognizer attempts to divide the signal into regions corresponding to phonetic units by using the features extracted. In the last step a word that matches best the sequence of phonemes is chosen.

3.2 The Pattern Recognition Approach

The pattern recognition approach avoids explicit segmentation and labeling of speech. Instead, the recognizer uses the patterns directly [2]. This is based on comparing a given speech pattern with previously stored ones [21]. The way speech patterns are formulated in the reference database affects the performance of the recognizer. In general, there are two common representations, namely, the Hidden Markov Model (HMM) and the templates. The second one is chosen here where the patterns are stored as a sequence of speech units that have similar spectral characteristics called subwords. Using a number of samples, the recognizer uses averaging techniques to build a reference pattern that encodes the important and unique features of each pattern. When the recognizer receives new input, it compares it directly with the patterns in the database in an attempt to find the best match [2].

Generally speaking, there are four steps involved in this approach [21], namely,

(1) *Feature Measurement*, where an analysis is performed to represent or model a time segment of the speech signal.

(2) *Pattern Training*, where the recognizer creates exemplars or templates for speech sounds of the different classes.

(3) *Pattern Classification*, where the recognizer compares the incoming speech pattern with all of the reference patterns. For each comparison a similarity measure is computed.

(4) *Decision logic*, is to make a final decision based on the distance measures computed in the previous step.

The Artificial Neural Network (ANN) which is discussed in detail in the next section, is employed in this paper as the pattern recognition engine.

4 Pattern Recognition Engines

Pattern recognition algorithms such as the one described by Rabiner and Wilpon in [24], use dynamic programming or Dynamic Time Warping (DTW) for isolated words systems. These algorithms are computationally proportional to the size of the vocabulary involved in a given recognition system, i.e., the templates stored for matching [21]. Two new approaches submerged in the late 1970's and early 1980's to accommodate the medium and large size vocabulary recognition paradigms but are as effective for the digits recognition systems. The first one is the HMM [20], and the second is the ANN [15] and [14].

4.1 Artificial Neural Networks

A neuron is defined as the fundamental processing unit of the human brain. Figure 4 shows a model of a neuron that has N inputs (the X's), N weights (the W's), a bias b and an output Y [5]. This output is calculated by the formula:

$$Y = f(\sum_{i=0}^{N-1}(W_i X_i - b)). \tag{12}$$

where b is an internal threshold or offset, and f is a non-linear function chosen from one of the ones below:

(1)*Hard limiter*, where

$$f(x) = \begin{cases} +1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

or,

(2)*Sigmoid functions*, where

$$f(x) = \begin{cases} \tanh(\beta x) & \text{if } \beta > 0 \\ or & \\ 1/1 + e^{-\beta x} & \text{if } \beta > 0. \end{cases}$$

The Sigmoid nonlinearities are used often since they are continuous and differentiable [21]. In general, an ANN is a network of several simple computational units. It has a great potential for parallel computation since the processing of the units is done independently and are widely used in pattern classification, matching and completion [15] [14].

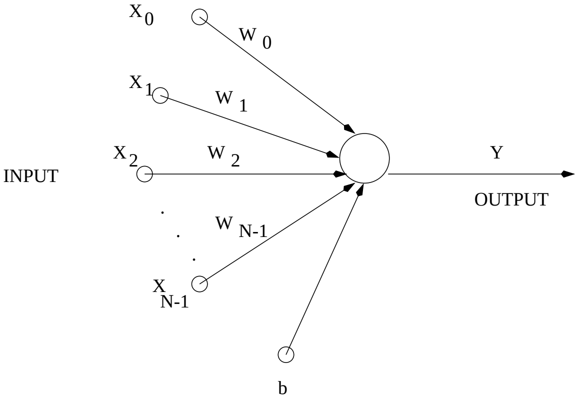

Figure 4: A computational node of a neural network [14].

4.2 Radial Basis Neural Networks

The core of a speech recognition system is the recognition engine. The one chosen in the paper is the Radial Basis Functions Neural Network (RBF). This is a static two neuron layers feed forward network with the first layer, L_1, called the hidden layer and the second layer, L_2, called the output layer. L_1 consists of kernel nodes that compute a localized and radially symmetric basis functions. A multi-layer (RBF) Neural Network is depicted in Figure 5. The pattern recognition approach used in this work is based on comparing a given speech pattern with previously stored ones [21]. The way speech patterns are formulated in the reference database affects the performance of the recognizer. In general, there are two common representations,

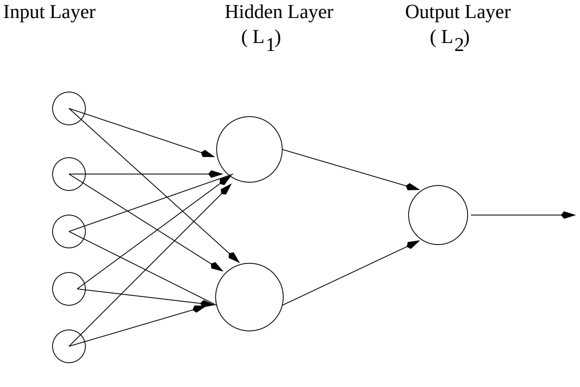

Figure 5: A multi-layer neural network [14].

The output y of an input vector x to a (RBF) neural network like the one in Figure 6, with H nodes in the hidden layer is governed by:

$$y = \sum_{h=0}^{H-1} w_h \phi_h(x). \tag{13}$$

where w_h are linear weights and ϕ_h are the radial symmetric basis functions. Each one of these functions is characterized by its center c_h and by its spread or width σ_h. The range of each of these functions is [0,1].

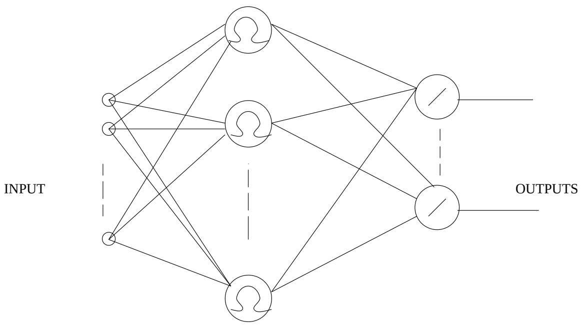

Figure 6: Radial Basis Functions Neural Network.

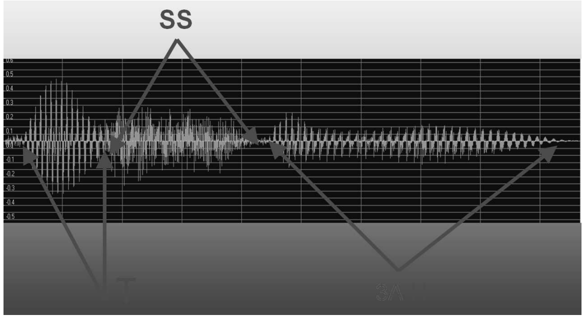

Figure 7: Segmentation of Tesaa = 9.

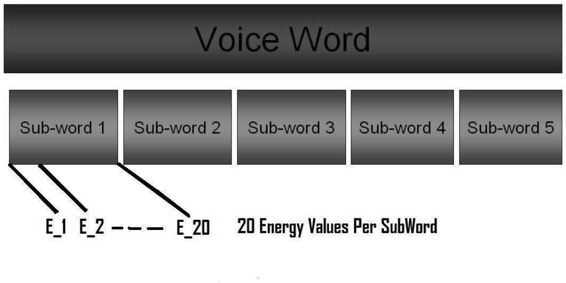

Figure 8: WPS Representation of Subwords

Once the input vector x is presented to the network, each neuron in the layer L_1 will output a value according to how close the input vector is to its weight vector. The more similar the input is to the neuron's weight vector, the closer to 1 is the neuron's output and vice versa. If a neuron has an output 1, then its output weights in the second layer L_2 pass their values to the neurons of L_2 [5]. The similarity between the input and the weights is usually measured by a basis function in the hidden nodes. One popular such function is the Gaussian function that uses the Euclidean norm. It measures the distance between the input vector x and the node center c_h. It is defined as:

$$\phi_h = exp(||x - c_h||/2\sigma_h^2). \tag{14}$$

5 Digit Vector Generation

Tracking changes of the acoustic properties in a speech sample is realized in a form of a (SVF) [23] whose peaks are indications of subword boundaries. The introduced system depicted finds boundaries that do not match those identified by the manual system. Due to the spurious peaks usually found in the (SVF), additional measures were taken to locate time boundaries. These measures include smoothing the (SVF) using spline polynomials to remove most of the false peaks in the function [1]. Moreover, some rules are applied to restrict the number of acoustic subwords of the digits. These rules depend on the knowledge of the phonetic decomposition of digits and also on energy levels and zero-crossing rates [1]. The performance of this part of the work was measured on the basis of its ability to locate the boundaries accurately and consistently. The ideal boundaries were assumed, in all cases, to be at the edges located manually as in Fig. 7 for the digit $Tess3a = 9$.

5.1 End-Point Detection

The algorithm used, is a modified version of that proposed in [1] which uses the signal energy to determine the most likely locations of the signal ends by finding Mel based energy vectors. The detection process performed here follows the (WPS).

5.2 Background Noise Estimation

This step ensures the algorithm's adaptation to the environment by estimating the noise level and set decision threshold values used by the subsequent steps. The energy levels of the first and last 160 ms of the signal are used for the noise threshold estimation. If the ratio between these two noise levels is outside the range [0.2 - 5], the algorithm terminates and manual end-point detection is requested.

5.3 Locating the First and Last Voiced Sounds

The starting-point of the first voiced sound and end-point of the last voiced sound are located using the thresholds set. These points serve as reference for the subsequent search for the actual end-points.

5.4 Locating Low Energy Areas

The algorithm moves backward (and forward) from the location of the previously determined points towards the ends of the signal to find the edges of low-energy sounds, The final endpoints are assumed to be located within these areas.

5.5 Final Endpoint Detection

The exact location of the endpoints is decided by locating the points where the slope of energy-time function has the highest value inside the low energy areas.

5.6 Frames and Windows

This stage is the front-end of most speech recognition systems where a speech signal is analyzed to obtain the speech mode1 employed by the system. The sampled speech signal is divided into fames of T_f sec. duration. The duration is chosen to be 10 ms. We define F_s as the sampling frequency and T_s as the sampling rate. $F_s = 20000$ KHz or $T_s = \frac{l}{F_s} = \frac{1}{20000} = 0.05$ m.s Therefore, the size of each frame N_f is: $N_f = \frac{T_f}{T_s} = 200$ samples. To smooth the change of parameters between frames, each frame is multiplied by the Hamming analysis window $w()$ of duration T_n. The window duration is chosen to be larger than the frame's (20ms) to provide an overlap of 50% between adjacent frames [19]. $N_w = \frac{T_w}{T_s}$ which is 400 samples. To extract energy parameters each windowed frame is analyzed with the DWTS. The speech frequency range 0-5 khz is then divided into 20 bands according to the (WPS). The last step is to compute the logarithm of the average absolute values of the coefficients over each of the 20 bands to get the energy value in that band encoded in decibel scale. The energy computed in each band is then scaled to a decibel scale of 0-60dB. Finally, the speech sample is a time sequence of these frame vectors. A speech sample has the form $[F_1, F_2, ..., F_{20}]$. The recognition part of the system was performed by a Radial Basis Function (RBF) neural network as specified by the manual system. The network was trained first by reference digit patterns in the form of parameter vectors. The vectors were created by partitioning the signal along the boundaries to isolate the acoustic subwords. Energy parameters from the (WPS) were extracted in each subword then concatenated to produce fixed-sized vectors thus relaxing the need for time-alignment. In recognition mode, the network was used to identify an unknown digit vectors. The fully automated system was tested using 10 speakers (with 1000 digits) from where signals are chosen from the Arabic digit speech data base. The digits were spoken by different speakers and recorded in the studio at the Lebanese American University, Byblos. The small size studio is designed to minimize noise and equipped with a multi directional microphone made by Neumann to collect the speech signal with the best quality. The signals are then recorded and transformed into wave sounds (.wav). The tool used is the PRO-Tools Control 24 Dig-Design Device which is a computerized digital mixer.

Arabic Digit	Phonemes
0	SE-F-R
1	WA-HA-D
2	E-TH-NA-NN
3	TH-A-LA-THA
4	A-RR-BA-A
5	KHA-MM-SA
6	SI-TT-A
7	SA-BA-AA
8	TH-A-MA-NI-YA
9	TE-SS-AA

Table 3: Phonetic of Arabic Digits.

5.7 Design Considerations

Speech is considered non-stationary, however, the properties in the signal can remain constant for periods between 5 and l0 msc. The frequency content may range up to 15 kHz or higher, but speech is sufficiently intelligible when bandlimited to frequencies bellow 3500Hz. A sampling rate of 20kHz or higher is required to accurately represent al1 speech sounds however, for commercial telephone applications, a sampling rate of 8kHz is sufficient [22][23].

5.8 Digit Vectors

Frame vectors are grouped into 5 subword vectors between the selected boundaries as depicted in Figure 8. Each subword unit consists of a variable number of frame vectors. Finally, a digit vector is generated by concatenating all sub-word vectors in the speech sample. Zero vectors are added in digit vectors that have less than the maximum number five of subword units to create fixed length vectors of 20 parameters.

5.9 Constructing the Feature Vectors

To extract energy parameters of the (WPS), we apply the wavelet analysis to each subword of speech Arabic digits selected for the testing phase. In this phase of the analysis, the constructed wavelet $bior3.9$ was used. The frequency bands are chosen according to the (WPS) that is similar to the Mel scale with frequency bands. The next step is to compute the average absolute values of the wavelets coefficients over the corresponding bands of the scale to obtain the energy values. These values are then scaled to a decibel scale of 0-60 dB.

$$E_{max} = max(E(p)) \quad 0 \leq p \leq P-1 \tag{15}$$

$$ES(p) = 20 * log10(E(p)/E_{max}) \quad 0 \leq p \leq P-1 \tag{16}$$

$$ES'(p) = ES(p) - E_{max} \quad 0 \leq p \leq P-1 \tag{17}$$

$$ES''(p) = max(ES'(p), -60dB) + 60dB \quad 0 \leq p \leq P-1 \tag{18}$$

There are 20 bands in the (WPS) corresponding to $P = 20$ of the Mel scale bands. Once the feature vectors were constructed, a Radial Basis Functions Neural Network is employed for recognition. Table 4 displays the recognition rates of the experiments conducted. The second step after selecting the speakers in the training set is to train the RBF network with the word vectors constructed.

6 Word Recognition

Next we present a description of the RBF ANN for the Arabic digits modeled using the (WPS) and a description of the RBF ANN are also included. The network contains two procedures, a training phase and a test phase, which are discussed in the next two subsections.

6.1 Network Training Phase

The RBF network implemented in this paper is trained initially with the Matlab [5] Neural Network tootbox function `newrb`$()$ which takes two input matrices, a goal matrix and a spread matrix, and returns a trained radial basis network. It is displayed in Figure 9. The first input matrix P is a $100 * Q$ matrix that contains a training set of Q digit vectors. The 100 correspond to 20 coefficients per subword multiplied by Six subwords per trained signal. If the network is being trained with 2 speakers then $Q = 40$ since each speaker is repeating each of the digits twice. The second input is a $Q * 10$ matrix of targets T. The rows of this matrix are targets vectors T_i that contain '1' in the targeted digit position and '0' otherwise. The output of the training function `newrb`$()$ consists of the centers and the weights C_h and $W_{q,h}$ for the hidden and output layers respectively.

$$P = [v_1, v_2, ..., v_Q] \tag{19}$$

$$T_i = [t_1, t_2, ..., t_{10}] \tag{20}$$

$$T = [T_1, T_2, ..., T_Q]^T \tag{21}$$

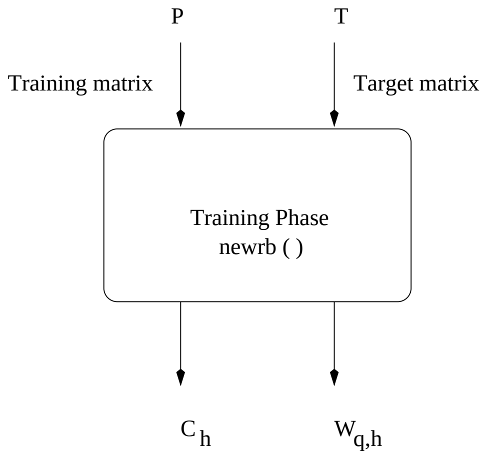

Figure 9: Training phase of the RBF network.

7 Experiments and Results

Ten speakers were chosen to test and evaluate all the intermediate ideas, concepts, and implementations that compose the final state of the algorithm. Moreover, digits in this set were used to measure the algorithm's accuracy in detecting the boundaries. This set has the following properties:

(1) All speakers belong to the database.

(2) Speakers represent both genders, 8 males, and 2 females.

The ABD'S accuracy evaluation was made by comparing the boundary locations generated by the proposed algorithm and those located manually by the visual inspection of the signal waveform and its spectrogram. For the purpose of this evaluation, it was assumed that the boundaries should ideally lie at the edges between phonemes in each digit. The boundary is decided to be accurately located if it lies within three frames (30ms) from the actual phoneme edge. Other possible States are shifted boundary, or missed boundary.

Scale	Recognition Rate
Mel-Fourier	87%
WPS(bior 3.9)	93%
WPS(bior 3.5)	89%
WPS(bior 6.8)	89%
WPS(db6)	92%

Table 4: Overall Recognition Average of Experiments.

7.1 Results of the Recognition Phase

The endpoint detection algorithm rejects any signal with a noise level higher than an experimentally determined threshold to prevent miss-allocation of endpoints. A manual Setting of the endpoints is required for the rejected signal before the subsequent processing can take place 15% of the digits processed by the algorithm were rejected. 64the allowed limit in addition to 1The algorithm detects an ”extra” boundary. This boundary always precedes the back endpoint of the digit and creates an extra subword. The boundary is considered an extra because it has not been classified manually. A successful recognition rate of 69.7% to 95.7% was achieved depending on the number of’ speakers used to train the RBF network. This result is lower than what was accomplished by the manual system due to the 20% error in locating some boundaries and the change of boundaries definition in both systems.

The Matlab [5] Neural Network toolbox function `sim()` is used to perform the recognition phase which is displayed in Figure 10. This function accepts a matrix R (similar to P of the training phase) of unknown digit vectors as an input along with the weights and bias vectors generated by the training phase. Its output is a a unit diagonal matrix where ‘1’ is placed in the recognized digit index.

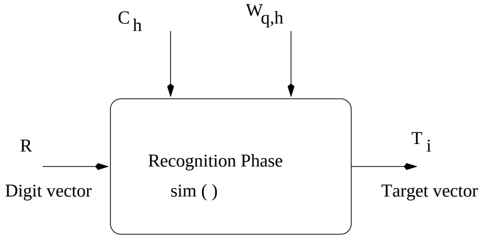

Figure 10: Recognition phase of the RBF network.

8 Conclusion

In this paper we construct Biorthogonal wavelets via the two channel perfect reconstruction filter bank. A detailed description of the implementations of the recognition system for the Arabic digits are examined and discussed. The speech signals parameterization using the Wavelet-based scales is also examined. The approach considered, decomposes spoken Arabic digits based on the acoustical information contained within each speech signal. The procedure locates the boundaries between subwords by finding the peaks in the function representing the spectral changes between consecutive speech frames. The Frame-based energy parameters derived from a Wavelet Packet Scale (WPS) are used in deriving the Spectral Variation Function (SVF). Three Biorthogonal wavelets are used as analyzing functions and their performances are compared with that of the orthogonal mother wavelet

9 Acknowledgment

The author would like to thank Alfaisal University and its Office of Research for securing the time, environment and funds to complete this research project. This work was also supported by the Alfaisal University Start-Up Fund (No. 410111410091).

10 Dedication

This paper is dedicated to my friend Dr. Maen Artimy.

References:

[1] M. Artimy, W.–J. Phillips and W. Robertson, Automatic Detection Of Acoustic Sub-word Boundaries For Single Digit Recognition Proceeding IEEE Canadian Conference on Electrical and Computer Engineering, 1999.

[2] D. Brewer, (1997), Speech Recognition Engines: [online], Available: http:// www.linfield.edu/ dbrewer/speech/, [2010, 22 September].

[3] Y.–T. Chan, Wavelet Basics, *Kluwer Academic Publisher*, Boston, 1995.

[4] I. Daubechies, Ten Lectures on Wavelets, *Philadelphia, SIAM*, 1992.

[5] H. Demuth and M. Beale, M., *Matlab Neural Network Toolbox*, Math Works, Natick, MA, 1997.

[6] A. Drygajlo, New Fast Wavelet Packet Transform Algorithms For Frame Synchronized

Speech Processing, *ICSLP 96. Proceedings, Fourth International Conference on Speech and Language Processing* Volume: 1, pp: 410–413, 1996.

[7] R.–F. Favero and R.–W. King, Wavelet Parameterization for Speech Recognition: Variations in Translations and Scale parameters, *Int. Symp. Speech Image Processing and Neural Networks Hong Kong*,Volume 2 pp: 694–697, April 1994.

[8] H. Goldstein, Formant tracking using the wavelet-based DST, *IEEE, COMSIG*, pp: 183–189, 1994.

[9] A. Graps, An Introduction To Wavelets, *IEEE Computational Sciences and Engineering*, Volume 2, Number 2, pp: 50–61, Summer 1995.

[10] J. Karam, A new approach in wavelet based speech compression, *Proceedings of the 10th WSEAS International Conference on Mathematical Methods, Computational Techniques and Intelligent Systems*, pp. 228–233, Corfu, Greece, October 26-28, 2008.

[11] J. Karam, A Comprehensive Approach for Speech Related Multimedia Applications, *WSEAS TRANSACTIONS on SIGNAL PROCESSING*, Issue 1, Volume 6, pp. 12–21, January 2010.

[12] J. Karam, Radial Basis Functions With Wavelet Packets For Recognizing Arabic Speech, *The 9th WSEAS International Conference on Circuits, Systems, Electronics, Control and Signal Processing*, Vouliagmeni, Athens, Greece, pp. 34–39, December, 2010.

[13] J. Karam, On the Distribution of Zeros for Daubechies Orthogonal Wavelets and Associated Polynomials, *15th WSEAS International Conference on Applied Mathematics*, Vouliagmeni, Athens, Greece, pp. 101–105, December 29 - 31, 2010.

[14] R.–P. Lippman, An Introduction to Computing with Neural Net, *IEEE ASSP Magazine*, pp. 4–39, 1987.

[15] R.–P. Lippman, Review Of Research On Neural Networks For Speech Recognition, *Neural computation*, Vol. 1, pp. 1–38, 1989.

[16] M. Misiti, Y. Misiti, G. Oppenheim and J. Poggi, *Matlab wavelet tool box*, 1997.

[17] R. Kronland-Martinet, *The Wavelet Transform for analysis, synthesis and processing of speech and music sound*, Computer Music Journal, Vol. 12, pp. 11–20, 1988.

[18] R. Kronland-Martinet, J. Morlet and A. Grossmann, Analysis of sound patterns through wavelet transform, *Recognition and Artificial Intelligence*, Vol.1, pp. 273–302, 1987.

[19] J.–W. Picone, *Signal Modeling Techniques in Speech Recognition*, IEEE, Vol.81, No.9, September 1993.

[20] L. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *IEEE, PROC.*, Vol. 77, No.2, February 1989.

[21] L.–R. Rabiner and B. Juang, Fundamental of Speech Recognition, *Prentice Hall*, New Jersey, 1993.

[22] L.–R. Rabiner and M.–R. Sambur, An algorithm for determining the end points of isolated utterances, Bell Systems Technical Journal, Vol.54, pp. 297–315, February 1975.

[23] L.–R. Rabiner and R.–W. Schafer, Digital Processing of Speech Signals, *Prentice Hall*, New Jersey, 1978.

[24] L.–R. Rabiner and J.–G. Wilpon, A Modified K-Means Algorithm for use of in Isolated Word Recognition, *IEEE Transactions on ASSP*, 33(3) pp. 587–594, June 1985.

[25] G. Strang and T. Nguyen, Wavelets and Filter Banks, *Wellesley–Cambridge Press*, Wellesley, MA, 1996.

[26] M. Vetterli, Wavelets and Filter Banks: Theory and Design, *IEEE Transactions on Signal Processing*, pp. 2207–2232 Vol. 40 September 1992.

[27] C. Taswell, Speech Compression with Cosine and Wavelet packet near-best bases, *IEEE International Conference on Acoustic, Speech, and Signal Processing*, pp. 566–568 Vol. 1, May 1996.

[28] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice Hall, PTR, Englewood Cliffs, New Jersey, 1995.

[29] R.–K. Young, Wavelet Theory and its Applications, *Kluwer Academic Publishers*, Lancaster, USA 1995.