A Faster Recognition of Similar Acoustic Sounds Using Wavelets

Jalal Karam
Department of Mathematics and Computer Science
Arab Open University
Beirut Lebanon
email: *drkaram_aou@hotmail.com*

Abstract

This paper presents a detailed treatement of the training and testing phases of a Radial Basis Functions Neural Network (RBFNN) used for the recognition of the similar acoustic sounds of the letters a, j, k of the English alphabets. It is shown that the framing and windowing techniques used in the traditional Fourier approach for speech coding is absolete in comparison with the wavelet analysis approach. The subwords of the speech signals were generated by identifying spectral changes of the waveforms. The modeling of each subword was accomplished using a Wavelet Packet Scale (WPS) which has a higher resolution of low frequency components and lower resolution of high frequency components than that of the Mel scale. The analyzing wavelet function used is the wavelet $db6$ which belongs to the Daubechies family of orthogonal wavelets.

1 Introduction

The pattern recognition approach avoids explicit segmentation and labeling of speech. Instead, the recognizer uses the patterns directly [1]. Figure 1 illustrates the operation of the pattern recognition approach used in this paper. It is based on comparing a given speech pattern with previously stored ones. The way speech patterns are formulated in the reference database affects the performance of the recognizer. In general, there are two common representations: The Hidden Markov Model (HMM) and the templates.

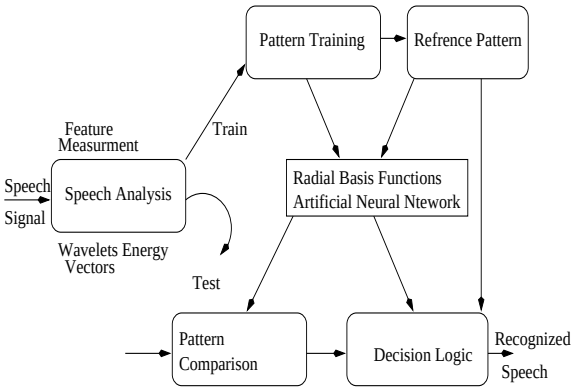

Figure 1: The pattern recognition approach.

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 [1].
Next section contains an account on the pattern recognition engines with emphesis on the RBFNN approach to the recognition of speech. Section four contains the implementations of the recognition systems, the training and testing phases of the RBF network that was built and employed as the recognition engine. the speech signals parameterization using the WPS is described in section Four. Section Five contains the experimental resuls and section Six contains the conclusion.

2 Pattern Recognition Engines

Pattern recognition algorithms such as the one described by Rabiner and Wilpon in [9], 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 [8]. Two new approaches submerged in the late 1970's and early 1980's to accommodate the small and medium size vocabulary recognition paradigms. The first one is the HMM and the second is the ANN [6] and [5].

2.1 Artificial Neural Networks

A neuron is defined as the fundamental processing unit of the human brain. Figure 2 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 [2]. This output is calculated by the formula:

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

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 [8]. In general, an ANN is a network of several simple computational units such as the one in Figure 2. 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 [6] and [5].

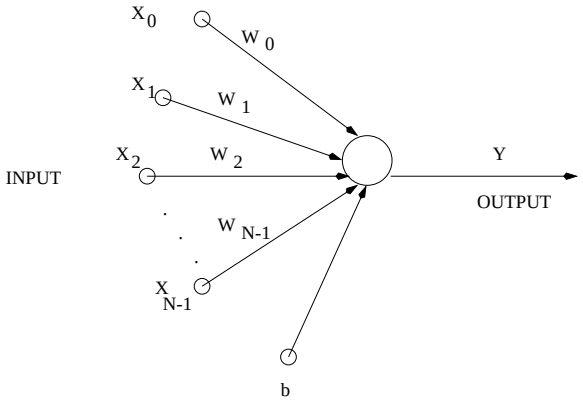

Figure 2: A computational element or node of a neural network.

2.2 Radial Basis Neural Networks

The core of a speech recognition system is the recognition engine. The one chosen in this work as Figure 1 suggests is the RBFNN. 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, as depicted in Figure 3. L_1 consists of kernel nodes that compute a localized and radially symmetric basis functions as in Figure 4.

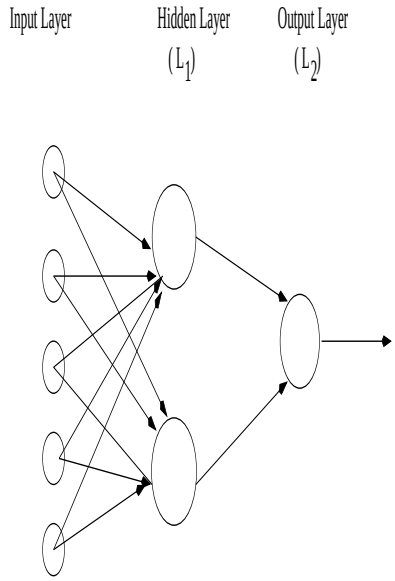

Figure 3: A multi-layer neural network.

The output y of an input vector x to a RBFNN with H nodes in the hidden layer is governed by:

$$y = \sum_{h=0}^{H-1} w_h \phi_h(x). \qquad (2)$$

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].

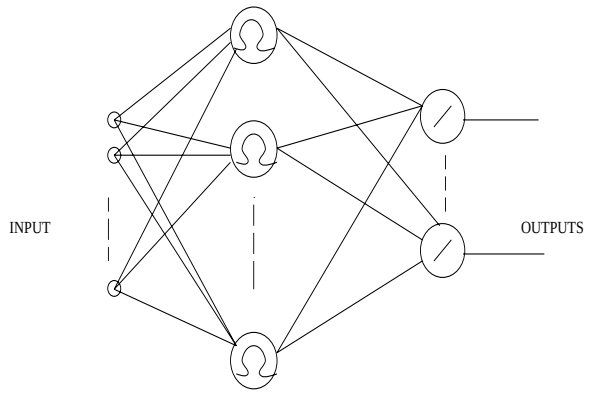

Figure 4: Radial basis functions neural network.

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 [2]. 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:

$$\phi_h = exp(||x - c_h||/2\sigma_h^2). \quad (3)$$

3 Formulation of Word Vectors

The segmentations of the speech signals into subwords is accoplished by visually inspecting the waveforms and their corresponding spectrograms. Their boundaries are characterized by the occurrence of important spectral changes. The three time marks selected correspond to two subwords for each signal as described in table 1.
A letter z is placed in the case where the word has less than two word vectors or subwords.
The following list of abbreviations used in Table 1:
b : begin
c : closure
ch : changes

Table 1: Selection of subwords in the time domain based on visible changes in the spectrograms

Word	TM1	TM2	TM3
a	**wb**	**we**	**z**
j	**wb**	**ch**	**we**
k	**wb**	**ch**	**we**

e : end
w : word
z: zero subword
(i.e. wb in column 1 represents word begin).

To extract energy parameters of the WPS, we apply the wavelet analysis to each subword. The frequency bands are chosen according to the scale [3] The last step is to compute the average absolute values of the wavelets coefficients over the corresponding to $P = 20$ bands of the scale to obtain the energy values. These values are then scaled to a decibel scale of 0-60 dB.

$$E_{max} = max(E(p)) \quad 0 \leq p \leq P-1 \quad (4)$$

$$ES(p) = 20 * log10(E(p)/E_{max}) \quad 0 \leq p \leq P-1 \quad (5)$$

$$ES'(p) = ES(p) - E_{max} \quad 0 \leq p \leq P-1 \quad (6)$$

$$ES''(p) = max(ES'(p), -60dB) + 60dB \quad 0 \leq p \leq P-1 \quad (7)$$

4 Training and Testing

4.1 Network Training Phase

The RBFNN implemented is trained initially with the Matlab [2] Neural Network tootbox function `newrb()` which takes two input matrices, a goal matrix and a spread matrix, and returns a trained radial basis network. The first input matrix P is a $40 * Q$ matrix that contains a training set of Q word vectors. The 40 correspond to 20 coefficients per subword multiplied by two subwords per trained signal. If the network is being trained with 2 speakers then $Q = 60$ since each speaker is repeating each of the words ten times. The second input is a $Q * 3$ matrix of targets T. The rows of this matrix are targets vectors T_i that contain '1' in the targeted word position and '0' otherwise as shown in Table 2. 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 as in Equation 2.

Table 2: Target vectors of the A-set

Alphabet	ta	tj	tk
a	**1**	**0**	**0**
j	**0**	**1**	**0**
k	**0**	**0**	**1**

$$P = [v_1, v_2, ..., v_Q] \tag{8}$$

$$T_i = [t_1, t_2, t_3] \tag{9}$$

$$T = [T_1, T_2, ..., T_Q]^T \tag{10}$$

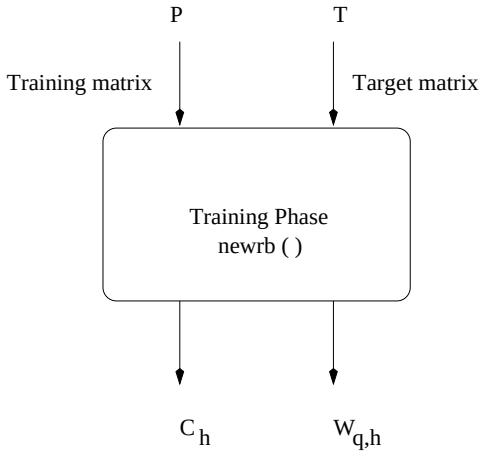

Figure 5: Training phase of the RBF network.

4.2 Network Recognition Phase

The Matlab [2] Neural Network toolbox function `sim()` is used to perform the recognition phase. This function accepts a matrix R (similar to P of the training phase) of unknown word vectors as an input along with the weights and bias vectors generated by the training phase as in Figure 6. Its output is a matrix similar to the target matrix of Table 2, where '1' is placed in the recognized index.

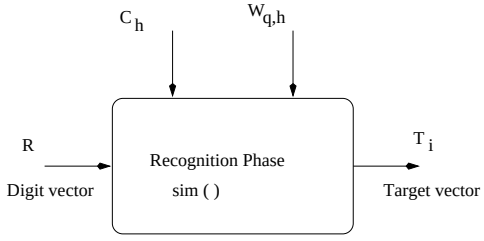

Figure 6: Recognition phase of the RBF network.

Table 3: The recognition percentage using the WPS and the Mel scale.

Train/Test	db6-WPS	Mel-scale
2/8	78	75
2/8	81	77
2/8	84	89
2/8	85	88
4/8	76	76
4/8	74	77
4/8	76	73
4/8	88	82
6/8	85	81
6/8	83	83
6/8	86	77
6/8	82	82
7/8	81	77
7/8	82	82
8/8	87	71

Table 4: Maximum, minimum, average and standard deviation of the experiments.

Scale	Max.	Min.	Ave.	σ
WPS(db6)	88	74	81.8667	4.2572
Mel	88	71	79.2667	4.9924

5 Experimental Results

All sixteen speakers of the database [7] are chosen. This set of speakers contains 8 males and 8 females and each speaker has 10 tokens per letter for a total of 540 speech files for the A-set containing the letters a, j, k. Four sets of 15 experiments were conducted. The recognition rates and their statistical results of these 60 experiments are shown in Table 3. Each row in this table contains the averaging of four experiments with different set of speakers but with the same number of speakers in the training and testing phases. The size of the test set was kept fixed at 8 as we increased the size of the training set from 2 to 8.
The normalized time required for the NN to train and test the experiments of conducted is illustrated in Table 5.

6 Conclusion

This paper gives a detailed description of the implementations of a RBFNN for the training and testing phases for the recognition of the A-set of the English

Table 5: Normalized time required for training and testing of the experiments.

Scale	Normalized Time
WPS	0.89
Mel	1

alphabets. The speech signals were parameterized using the WPS. A comparison in the performence of the recognizer between the wavelet model and that of the Mel was conduceted. The methods proposed and the experiments conducted in this work have shown that:

(1) Wavelet analysis can be applied to different length subwords or segments of speech signals This results in a reduction of the averaging operations that the components of the feature vectors of the Fourier based methods are subjected to.

(2) The WPS slightly outperformed the Fourier based Mel scale model.

(3) An increase in the size of the training set did not affect its performance. This was the case for the Mel scale and the WPS.

(4)The WPS model need less time than that of the Mel to train and test the RBFNN.

References

[1] Brewer, D.(1997). Speech Recognition Engines: [online], Available: http:// www.linfield.edu/ dbrewer/speech/, [1999, 22 August].

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

[3] Karam, J.R., Simulation and Analysis of Wavelet Based Speaker IndependentSpeech Recognition Systems for Isolated Spoken Words, PhD Thesis, Technical University of Nova Scotia, Halifax, 2000.

[4] Karam, J.R., Phillips, W.J. and Robertson, W., New Low Rate Wavelet Models For The Recognition Of Single Spoken Digits, IEEE, proceedings of ccece, Halifax, pp:331-334, May, 2000

[5] Lippman, R.P., An Introduction to Computing with Neural Net, IEEE ASSP Magazine, pp: 4-39, 1987.

[6] Lippman, R.P., Review Of Research On Neural Networks For Speech Recognition, Neural computation, Vol. 1, pp: 1-38, 1989.

[7] NIST, Speech Discs 7-1.1, TI 46 Word Speech Database Speaker-Dependent Isolated-Digital Corpus, LDC93S9, Texas Instruments, Sep. 1991.

[8] Rabiner, L. Juang, B., Fundamental of Speech Recognition, Prentice Hall, New Jersey, 1993.

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

[10] Rabiner, L.R. and Wilpon, J.G., Some Performance Benchmarks For Isolated Word speech recognition Systems, Computer, Speech and Language, Vol. 77, pp: 343-357, 1987.