Hybrid Wavelet-Neural/FFT-Neural Phoneme Recognition

Dr. Abduladheem A. Ali Dr. Majid A. Alwan Abbas A. Jasim
(M.Sc. Computer Eng.)

University Of Basrah
College Of Engineering
Computers Engineering Department

Abstract

This work describes the implementation of a hybrid approach for phoneme recognition that mixes two elementary approaches (wavelet-neural and FFT-neural). Each of the two elementary approaches is developed separately. As a second stage the two elementary approaches are merged together to produce the hybrid approach. Results proved that the hybrid approach has higher accuracy than each of the two elementary approaches.

1. Introduction

Automatic speech recognition has been a goal of researches for more than five decades [1]. Speech recognition systems can be used to automate many tasks that previously required hand-on human interaction. These systems can be built based on word, syllable, or phoneme, which are called units of speech recognition [2]. Word based speech recognition systems suffered from several problems. One important problem is that: there is no training data sharing for words and each word has its own model in the system [3]. Since languages contain very large number of words, it is impossible to use word as a unit of large vocabulary speech recognition systems. Syllables can be shared among words, but there are also a large number of syllables, for example there are 10000 syllables in English [1].

On the other side each language has its distinctive set of phonemes, which are the smallest units in the speech system [4]. Typically the number of phonemes per languages are between 30 and 50. All words can be constructed from this relatively small set of units (phonemes). Then phoneme is the best choice for large vocabulary speech recognition systems. A problem is encountered in phoneme recognition is that low performance of the recognizer due to co- articulation effect. Generally, for the best phoneme recognition systems the recognition accuracy about 70%[5]. By delaying phoneme transcription until after word sequence is identified, poor level phoneme recognition performance can be converted to good word level performance using higher level constraints.

2. Wavelet-neural Phoneme Recognition

In this approach, wavelet packet transform is firstly applied to the phonemes signals. The wavelet packet transform is used to determine certain feature vectors. Neural networks use these feature vectors for the recognition process.

2.1 The Discrete Wavelet Transform

The wavelet transform is a technique for analyzing signals. It is relatively recent and computationally efficient technique for extracting information about non-stationary signals like speech [6].

There is a fast algorithm for computing Discrete Wavelet Transform (DWT) is called Mallat algorithm. In Mallat algorithm firstly the signal is filtered by high pass filter (HPF) and low pass filter (LPF), then the signal is down sampled by two levels, this operation is repeated on the low pass section of the signal and so on as shown in Figure 1 . $g(n)$ and $h(n)$ are the HPF and LPF respectively. The decomposition equations are:

$$x^{(j-1)}(n)=\sum_k x^{(j)}(k)h(2n-k)$$

$$d^{(j-1)}(n)=\sum_k x^{(j)}(k)g(2n-k)$$

Where $x^{(j)}$: is the signal. $x^{(j-1)}$: the signal at courser scale (approximation), and $d^{(j-1)}$: the details of the signal.

While in Wavelet Packet Transform(WPT) the high pass section of the signal is splited as with low pass section.

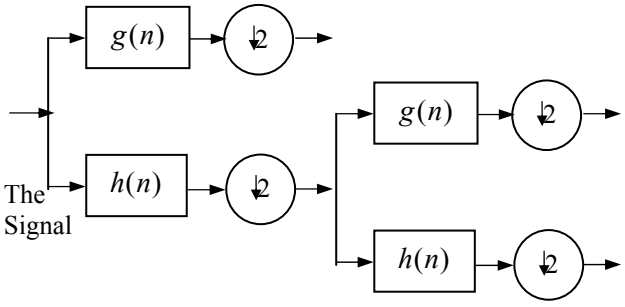

Figure 1:Two level Wavelet transform

2.2 The Procedure of Wavelet-neural Phoneme Recognition

The training procedure steps are:

1. Phoneme signals are to be normalized so as to make their energies equal to 1. This can be done by applying the following equations to speech signals that represent the Arabic phoneme [2].

$$S_{normalized}(n) = \frac{s(n)}{\sqrt{\sum_m [s(m)]^2}}$$

where s : is the signal.

The energy (E) of the obtained signal is:

$$E = \sum_n \left[S_{normalized}(n) \right]^2$$

$$= \sum_n \left[\frac{s(n)}{\sqrt{\sum_m [s(m)]^2}} \right]^2 = 1$$

The normalization step is very important since the incoming phonemes are usually having different intensities due to speaker loudness or speaker distance from microphone. If this variability is not adjusted, recognition mistakes will be occurred [2].

2. Five levels Wavelet Packet Transform WPT is obtained. Figure 2 shows speech signal with the first level WPT.

3. The energy of selected signals of WPT decomposition tree are calculated and then used as inputs to each Neural Network NN. The energy of the signal in WPT tree is a measure of the signal in that sub band of frequencies. All nodes in level 2 and level 3 are used to decompose the signal in equal bands (uniform frequency bands). The other selected nodes represent further splitting of the low frequency band of the signal. That is a simulation of human perception non-linearity in which narrow frequency bands are taken for low frequencies. Figure 3 shows five-level WPT decomposition tree. The shaded nodes represent selected sub band signals and for each splitting the low pass cell is in the left and the high pass one is in the right.

4. A Multi Layer Perceptron MLP NN is trained for each phoneme using backpropagation learning algorithm. Each phoneme is taken in different contexts. The number of input neurons in the input layer is 16 which are the energy values for the sub band signals selected in step 3. If there are few nodes in the hidden layer, then the neural network either does not learned and saturate with unaccepted error or take more time to be learned. On the other hand if the number of units in the hidden layer is high, then NN will take more learning time. Several experiments show that the suitable number of hidden neurons is equal to 40 . .There is one neuron in the output layer.

For each one of the 33 Arabic phonemes one neural network is learned. The phoneme N.N. is learned on 12 token of that phoneme. Phoneme N.N. gives one on its output if the feature vector that is presented at the input layer is related to one token of that phoneme. Else the output will be zero.

In testing phase twelve token for each phoneme are taken. The steps that are applied on the test phoneme are normalization, WPT, computing the energy values of the selected nodes, and then these values are applied to all the 33 neural networks. The output of each neural network is computed. The neural network that has minimum error is determined. The error is measured between the output of the neural network and the target output of it if the corresponding phoneme's feature vector is presented at its input layer. The phoneme corresponds to minimum error neural network is

said to be recognized. The recognition Accuracy of this approach is found to be 56.06.

3. FFT-Neural Phoneme Recognition

This approach applies recognition procedure on phonemes based on the feature vectors determined using Fast Fourier Transform(FFT).

3.1 The FFT and DFT

The Discrete Fourier Transform DFT can be defined as a Fourier representation of finite-length sequence, which is itself a sequence rather than a continuous function, corresponds to samples equally spaced in frequency of the Fourier Transform of the signal [7]. DFT can be obtained from the equation:

$$X(k)=\sum_{n=0}^{N-1}x(n)[\cos(2\pi kn/N)-j\sin(2\pi kn/N)]$$

Where $k=0,1,...,N-1$

When DFT is computed in more efficient algorithms it would be called the Fast Fourier Transform.

3.2 The Procedure of FFT-neural Phoneme Recognition

The training procedure takes the following steps:

1. Energy normalization of the phoneme signal.
2. A section of 256 point from the middle of each phoneme signal. Then 256-point FFT is computed for that section. A 256 point FFT is taken since some phonemes are not exceeding in length 512 point, but most phonemes have more than 256 in their sampled version on 16 kHz sampling rate and then more than 16 msec.
3. Sixteen features vector is measured for each phoneme. Each of these features measured by summing more than one value from the FFT magnitude (frequency domain information samples).

   If $S(k)=magnitude\ (FFT\,(s(n))$

   where $s(n)$: the time domain sequence of the signal of time index n.

   $S(k)$: the magnitude of FFT of $s(n)$ of frequency index k.

   The feature vector (F) is measured from $S(k)$ as:

$$F(1)=\sum_{k=3}^{4}S(k),\ F(2)=\sum_{k=5}^{8}S(k),\ F(3)=\sum_{k=9}^{14}S(k),$$

$$F(4)=\sum_{k=15}^{20}S(k),\ F(5)=\sum_{21}^{26}S(k),\ F(6)=\sum_{k=27}^{32}S(k),$$

$$F(7)=\sum_{k=33}^{38}S(k),\ F(8)=\sum_{k=39}^{44}S(k),\ F(9)=\sum_{k=45}^{50}S(k),$$

$$F(10)=\sum_{k=51}^{56}S(k),\ F(11)=\sum_{k=57}^{62}S(k),\ F(12)=\sum_{k=63}^{68}S(k)$$

$$,\ F(13)=\sum_{k=69}^{74}S(k),\ F(14)=\sum_{k=75}^{80}S(k),$$

$$F(15)=\sum_{k=81}^{88}S(k),\ F(16)=\sum_{k=89}^{98}S(k).$$

4. MLP neural network is learned for each phoneme by using backprobagation learning algorithm in the same manner that is in wavelet-neural method.

This method is tested on 12 token from each phoneme and it gives 54.79 % recognition accuracy.

4. Hybrid Wavelet-Neural/FFT-Neural Phoneme Recognition

The hybrid wavelet-neural/FFT-neural phoneme recognition approach is developed to improve the accuracy of the system. It merges the wavelet-neural approach with FFT-neural approach. A condition is applied then. Each of these two elementary approaches is modified to produce the best three candidate phonemes(corresponding to the minimum error networks). The decision rule is then deals with these two sets that are generated from each approach separately.

4.1 The Procedure of Hybrid (Wavelet-Neural/FFT-Neural) Phoneme Recognition

The procedure of the hybrid approach contains steps that are combination of more than

one step of the elementary approach procedures. These steps are:

1. Wavelet-neural approach is applied to the training phonemes by normalization, WPT, computing the energies, and the training of NNs.
2. FFT-neural approach is applied for each of the training phonemes by normalization, computing the 256-FFT, computing the networks input, learning the 33 NNs separately.

For testing the phonemes in the recognition phase the following will be done:

1. Wavelet-neural approach is applied to generate the feature vector as with the training phoneme. The feature vector is applied to the input of each NN and the error of each NN is computed. The best three phonemes are determined and labeled as w_1, w_2, and w_3 according to the errors in the network outputs.
2. FFT-neural approach is applied to the test phoneme to generate the feature vector in the same manner that was used with the training set of phonemes. The feature vector is applied to the input layer of each neural network and the outputs of the NNs are computed. The best three phonemes are determined and labeled as f_1, f_2, and f_3.
3. The following rules are applied on the two tests of recognized best phonemes for each approach as:

   if (w_1=f_1) then the recognized phoneme is w_1

   if (w_1=f_2) then the recognized phoneme is w_1

   if (w_2=f_1) then the recognized phoneme is w_2

   if (error of w_1<0.02) then the recognized phoneme is w_1

   if (error of f_1<0.02) then the recognized phoneme is f_1

   if (w_2=f_2) then the recognized phoneme is w_2

   if (w_1=f_3) then the recognized phoneme is w_1

   if (w_3=f_1) then the recognized phoneme is w_3

   if (w_2=f_3) then the recognized phoneme is w_2

   if (w_3=f_2) then the recognized phoneme is w_3

   if (w_3=f_3) then the recognized phoneme is w_3

   else the recognized phoneme is the one that has minimum error

The block diagram of the hybrid wavelet-neural/FFT-neural approach is shown in Figure 4 . After testing this approach the recognition accuracy was 67.93% which is higher than both elementary approaches.

5 . Results

Firstly for each phoneme there are 24 token for each Arabic phoneme are recorded by single speaker. Cool Edit Pro software is used for the recording process, the sampling rate was 16000 Hz. For each phoneme 12 token that represent the phoneme in different context are used for training the recognizer. The other 12 are used for testing. Table 1 shows the Arabic phonemes, IPA symbols, and the recognition accuracies in Wavelet-neural, FFT-neural, and the Hybrid. Table 2 shows the test tokens of the phoneme [u:] with their recognized phonemes in three approaches.

6.Conclusion

It is noted from this work that the recognition accuracy of the hybrid Wavelet-neural/FFT-Neural approach is higher than each of the two elementary approaches that are used as building blocks. When both elementary approaches recognize the same phoneme then the decision rule of the hybrid approach would select that phoneme to be recognized as with [u:]1 in Table 2. Some times when one of the two elementary approaches recognize the correct phoneme that phoneme would be recognized by the hybrid approach such as [u:]2 and [u:]3. That is because the same phoneme is found in the

best three of the other elementary approach even when it is not the first. Finally there is correct recognition case even when both elementary approaches fail to find the correct phoneme that is occurred when each of elementary approaches candidates that phoneme .

7. References

[1] L. Rabinar and R. W. Schafer, " Fundamental of Speech Recognition ", Prentice Hall 1993.

[2] Mohammad J. Haider, " Speech Compression and Recognition Using: Wavelet Transform ", Msc. Thesis, Electrical Engineering Department, Baghdad University, 1999.

[3] S. H. Amin, " Speech Recognition Using Neural Networks ", Msc. Thesis, Control and Computer Engineering, University of Technology, 2000.

[4] J. N. Holmes, " Speech Synthesis and Recognition ", Van Nostrand Reinhold (uk) co. Ltd, 1988.

[5] Mark Huckvale, " 10 Things Engineers Have Discovered About Speech recognition ", Presented at NATO ASI Workshop Speech Pattern Processing, Jersy, 1997.

[6] G. Tazanetakis, " Audio Analysis Using the Discrete Wavelet Transform", http:// www. cs. princeton. edu/ ~gessl/ papers/ amta 2001. pdf.

[7] A. V. Oppenheim, " Digital Signal Processing ", Prentice Hall, 1975.

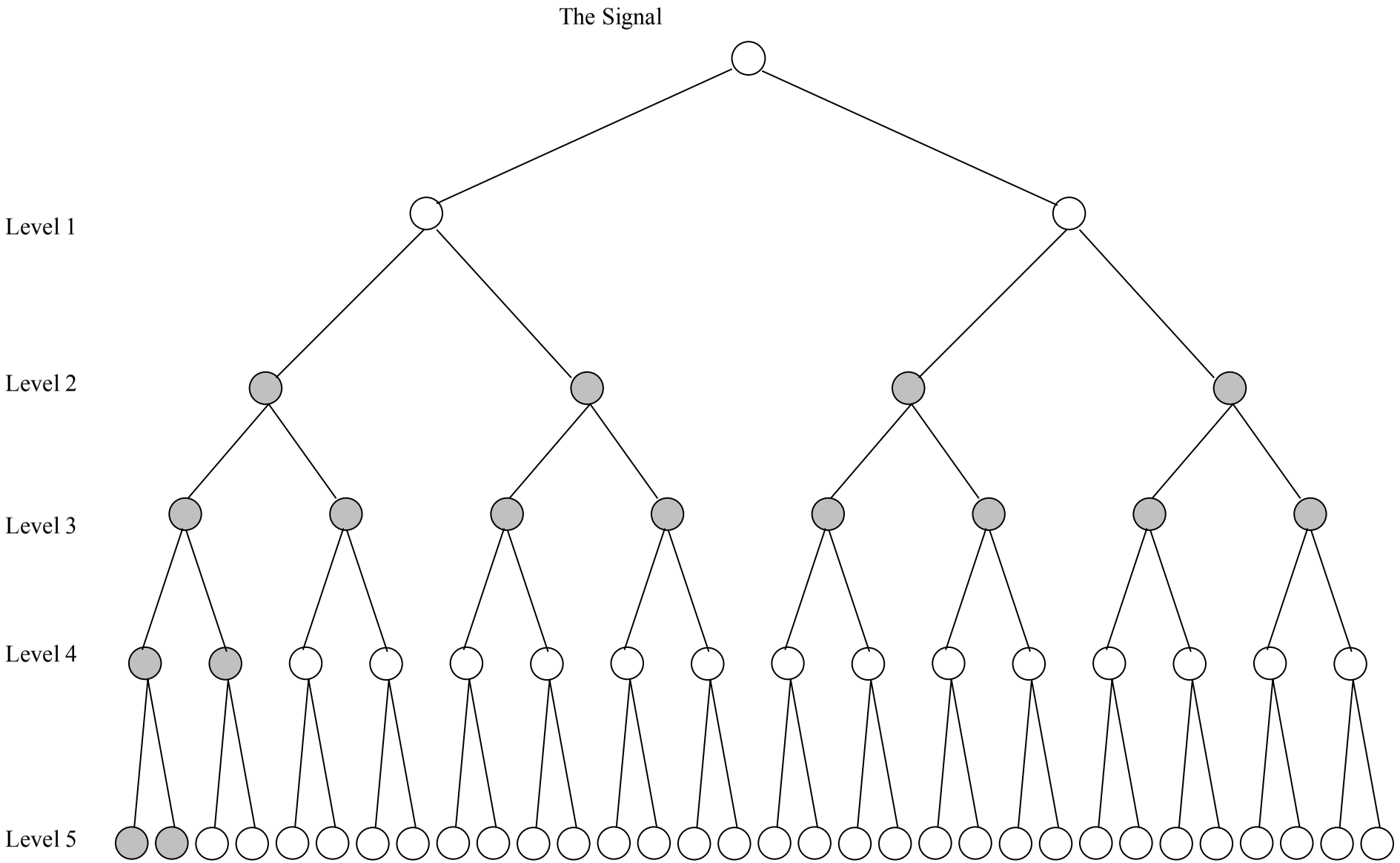

Figure 3: Five level WPT decomposition tree with shaded selected nodes

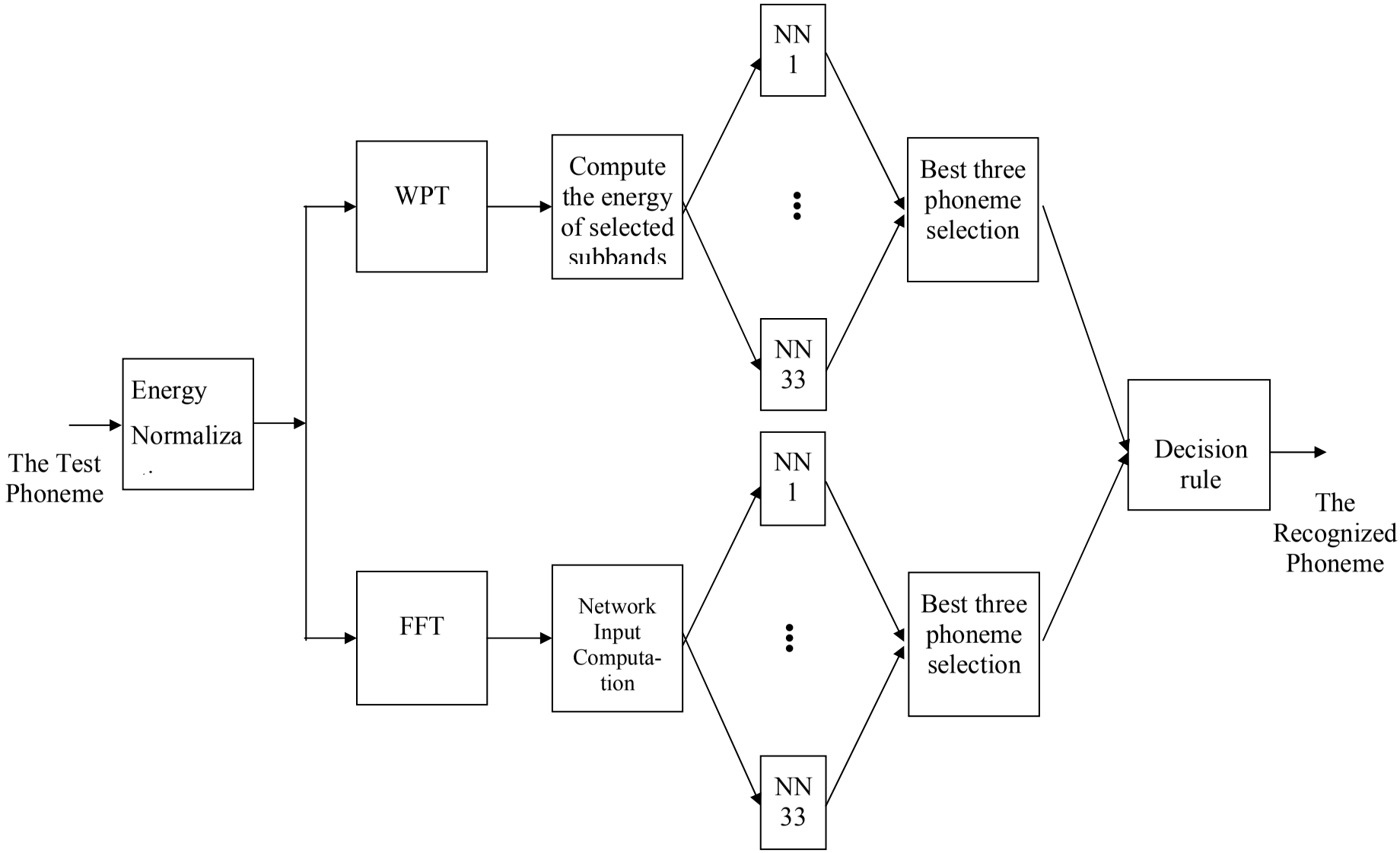

Figure 4: The Block Diagram of Hybrid Wavelet-neural\FFT-neural Phoneme Recognition Approach

Phoneme	IPA Symbols	Phoneme recognition accuracy % by		
		Wavelet-Neural	FFT-Neural	The Hybrid
ء	[ʔ]	75	75	91.66
آ	[a:]	75	83.33	83.33
ب	[b]	25	41.66	33.33
ت	[t]	50	83.33	100
ث	[θ]	25	41.66	41.66
ج	[dʒ]	16.77	50	50
ح	[ħ]	83.33	91.66	91.66
خ	[x]	83.33	33.33	83.33
د	[d]	50	25	41.66
ذ	[ð]	50	50	41.66
ر	[r]	41.66	25	50
ز	[z]	66.6	41.66	66.66
س	[s]	75	25	33.33
ش	[ʃ]	100	91.6	91.66
ص	[s̲]	66.66	25	41.66
ط	[t̲]	50	83.33	75
ض	[d̲]	33.3	50	41.66
ع	[ʕ]	41.66	58.33	58.33
غ	[γ]	83.3	91.66	91.16
ف	[f]	83.33	50	66.66
ق	[q]	33.3	50	66.66
ك	[k]	50	41.66	58.33
ل	[l]	58.33	25	41.66
م	[m]	41.66	66.66	58.33
ن	[n]	75	58.33	6.66
هـ	[h]	25	33.33	50
و	[w]	66.66	83.33	91.66
ي	[j]	83.3	75	83.33
fateha	[a]	58.33	58.33	66.66
dhama	[u]	33.3	41.66	41.66
kasrah	[i]	25	33.33	58.33
Vowel (و)	[u:]	50	66.66	75
Vowel(ي)	[i:]	75	66.66	83.33
Average		**56.06**	**54.79**	**67.93**

Table 1: Phoneme recognition accuracies in the three approaches.

Test phoneme	[u:] 1	[u:] 2	[u:] 3	[u:] 4	[u:] 5	[u:] 6	[u:] 7	[u:] 8	[u:] 9	[u:] 10	[u:] 11	[u:] 12
Recognized. ph. By Wavelet-Neural	[u]	[ð]	[u:]	[i:]	[u:]	[u:]	[u:]	[ð]	[u]	[u]	[u:]	[u:]
Recognized ph. By FFT-NN	[u:]	[u:]	[θ]	[u:]	[u:]	[i]	[u:]	[θ]	[u:]	[i]	[u:]	[u:]
Recognized ph. By hybrid	[u:]	[u:]	[u:]	[u:]	[u:]	[i]	[u:]	[θ]	[u:]	[i]	[u:]	[u:]

Table 2 : Test tokens of [u:]and the corresponding recognized phonemes in three methods