CHAPTER 5

SPEAKER VERIFICATION SYSTEM USING WAVELET NEURAL NETWORK AND FUZZY WAVELET NETWORK

5.1 INTRODUCTION

Some of the recent works on speaker verification depend on various classical features, modelling techniques, and hybrid approaches for modelling the speaker. Neural Networks (NN) are particularly powerful for handling large scale problems. However, performance of NN, changes with the structure used for implementation and number of network parameters. The NNs have limited ability to characterize local features such as discontinuities in curvature, jumps in objective function, etc. (Bing Xiang & Toby Berger 2003).

The advantages of NNs are further strengthened with the introduction of a wavelet into NN structures. Wavelet theory is used to approximate nonlinear functions using two main properties: finite support and self-similarity (Shung-Yung & Lung 2008). Wavelet Neural Networks (WNN/Wavenet) are feed-forward neural networks using wavelets as activation function. Wavenet has the ability to extract the distinguishable and essential features in frequency rich signals. This is required in classification and identification problems such as speaker verification. The main advantage of WNN over similar architectures is the possibility to optimize wavelet network structure by means of efficient deterministic constructive algorithms (Zhang & Benveniste 1992). WNNs have important characteristics such as
faster convergence, avoidance of local minimum, easy decision, and adaptation of structure. Engin (2007) has used discrete wavelet neural network for word recognition.

The performance of wavelet neural network differs by the selection of wavelet functions. In (Shaban et al 2007), a hybrid approach of wavelet transform and neural networks is adopted for speaker identification. The same hybrid approach together with a number of other approaches is studied in (Abduladheem et al 2005) and their performances are compared for recognizing phonemes uttered by a single speaker. In these papers, the wavelet transform is applied on the features and the transformed features are fed into neural network to train it. But in some papers (Daniel & Ho 2001; Zhang 1997), the wavelet neural network (Thuillard 2000) is implemented by replacing the activation function of neural network by wavelet functions. Here, we have analysed both method of implementation of wavelet neural network for speaker verification system. The performance of SV system using these wavenets is compared with the ASV system based on Fuzzy wavelet neural network.

5.2 WAVELET NEURAL NETWORK

Since the first proposal by Zhang & Benveniste (1992), WNN has been widely researched and applied. It avoids the convergence of neural network at local minimum and it provides faster convergence than Radial Basis Function Neural Network (RBFN) (Zhang 1997). It is composed of a three-layer neural network in which the wavelet function is the activation function of the neurons in the hidden layer as shown in Figure 5.1. Wavelet function is a waveform that has limited duration and average value of zero (Zhang & Benveniste 1992). WNN has the ability to deal with the problems of “Curse of Dimensionality” (Thuillard 2000). The efficiency of WNN is in the effective learning of the activation function and its evaluation.
The general structure of the wavelet neural network used in this research is shown in Figure 5.1. It consists of input layer, hidden layer and output layer. The weight between the input neuron $k$ and hidden neuron $j$ is denoted by $w_{jk}$ and the connection weight between hidden neuron $j$ and output neuron $i$ is denoted by $v_{ij}$. The activation function of hidden neurons is wavelet functions of different resolutions and the output neurons have sigmoid activation functions.

Let us consider, $x_1, x_2, \ldots, x_L$ are the network input signals. The hidden layer of the wavelet neural network which has wavelet activation functions are called as wavelet layers. The wavelet activation functions are derived from a mother wavelet $\Psi(x)$, which satisfies the admissibility condition (Malik & Afsar 2009)

$$\int_{-\infty}^{\infty} \omega |\hat{\Psi}(\omega)|^2 d\omega < \infty$$

(5.1)

where $\hat{\Psi}(x)$ indicates the Fourier transform of $\Psi(x)$.

The output of the WNN is calculated as (Zhang & Benveniste 1992)
\[ y_i(t) = \sigma(x_n) = \sigma \left( \sum_{j=1}^{N} v_{jk} \psi'(z_i) \sum_{k=1}^{K} w_{jk} x_k(t) \right) \]  

(5.2)

where \( i = 1, 2, \ldots, N \) and

\[ \sigma(x) = \frac{1}{1 + e^{-x}} \]  

is sigmoid function used in the output layer.

\( Z_i \) is the input to the wavelet function and is calculated from

\[ Z_i = \sum_{j=1}^{N} x_i a_{ij} - b_i \]  

(5.3)

where \( a_{ij} \) and \( b_i \) are scaling and translation coefficients.

The WNN is trained using back propagation algorithm and the parameters of the wavelet and the weight parameters of the neural network are updated during training. Initial values of these parameters are selected at random and the parameters are tuned to minimize the least square error. The error is between desired target and the actual output of the WNN for the given input pattern. Thus the objective function to be minimized is given by (Zhang & Benveniste 1992)

\[ E = \frac{1}{2} \sum_{p=1}^{P} \sum_{n=1}^{N} (d_i^p - y_i^p)^2 \]  

(5.4)

The wavelet function used in this work is modified Morlet function (Teolis 1998; Malik & Afsar 2009)
\[ \psi_{a,b} \equiv \cos \frac{0.75t_z}{e^{2\theta^2}} \]  \hspace{1cm} \text{(5.5)}

where \( t_z = \left( \frac{t - b}{a} \right) \).

\( \theta \) - Wavelet width

b - Dilation coefficient

a - Scaling coefficient

The derivative of wavelet function is given by (Qian-jin Guo et al 2005)

\[ \psi'_{a,b}(t) = \left\{ -1.75 \sin 1.75t_z \exp \left( -\frac{t_z^2}{2\theta^2} \right) \cos 1.75t_z \exp \left( -\frac{t_z^2}{2\theta^2} \right) t_z \right\} \]  \hspace{1cm} \text{(5.6)}

The output of each neuron in the output layer is denoted by \( net_j = \sum_{k=0}^{K} w_{jk} x_k \). Thus the wavelet function and overall net output of the WNN are modified as

\[ \psi_{a,b}(net_j) = \psi \left( \frac{net_j - b_j}{a_j} \right) \]  \hspace{1cm} \text{(5.7)}

\[ y_i(t) = \sigma \left( \sum_{j=0}^{M} v_{ij} \psi_{a,b}(net_j) \right) \hspace{1mm} i=1,2,...N \]  \hspace{1cm} \text{(5.8)}

where the derivative of the output layer activation function is given by
\[
\sigma'(u) = \frac{\partial \sigma(u)}{\partial u} = \sigma(u) \left[ -\sigma(u) \right] 
\]

(5.9)

Using these relationships, the objective function is differentiated with respect to each of the parameter which is to be updated. The change in the parameters in each iteration in the algorithm is obtained as

\[
\delta_{v_{ij}} = \frac{\partial E}{\partial v_{ij}} = -\sum_{p=1}^{P} (d^p_i - y^p_i) y^p_i (1 - y^p_i) \psi'_{a,b} (net^p_j) 
\]

(5.10)

\[
\delta_{w_{ji}} = \frac{\partial E}{\partial w_{ji}} = -\sum_{p=1}^{P} \sum_{i=1}^{N} (d^p_i - y^p_i) y^p_i (1 - y^p_i) v_{ij} \psi'_{a,b} (net^p_j) x^p_i / a_j 
\]

(5.11)

\[
\delta_{b_j} = \frac{\partial E}{\partial b_j} = \sum_{p=1}^{P} \sum_{i=1}^{N} (d^p_i - y^p_i) y^p_i (1 - y^p_i) v_{ij} \psi'_{a,b} (net^p_j) / a_j 
\]

(5.12)

\[
\delta_{a_j} = \frac{\partial E}{\partial a_j} = \sum_{p=1}^{P} \sum_{i=1}^{N} (d^p_i - y^p_i) y^p_i (1 - y^p_i) v_{ij} \psi'_{a,b} (net^p_j) \frac{(net^p_j - b_j)}{a_j^2} 
\]

(5.13)

\[
\delta_{\theta} = \frac{\partial E}{\partial \theta} = \sum_{p=1}^{P} \sum_{i=1}^{N} (d^p_i - y^p_i) y^p_i (1 - y^p_i) \psi_{a,b} (net^p_j) \frac{(net^p_j)^2}{\theta^3} 
\]

(5.14)

Hence, the changes in the parameters with the learning rate \( \eta \) and momentum \( \mu \) are given by (Meyer 1993)

\[
\Delta w_{jk} (t + 1) = -\eta \frac{\partial E}{\partial w_{jk}} + \mu \Delta w_{jk} (t) 
\]

(5.15)

\[
\Delta v_{ij} (t + 1) = -\eta \frac{\partial E}{\partial v_{ij}} + \mu \Delta v_{ij} (t) 
\]

(5.16)
\[ \Delta a_j(t+1) = -\eta \frac{\partial E}{\partial a_j} + \mu \Delta a_j(t) \]  
(5.17)

\[ \Delta b_j(t+1) = -\eta \frac{\partial E}{\partial b_j} + \mu \Delta b_j(t) \]  
(5.18)

\[ \Delta \theta_j(t+1) = -\eta \frac{\partial E}{\partial \theta_j} + \mu \Delta \theta_j(t) \]  
(5.19)

Then the update equations are as follows:

\[ w_{jk}(t+1) = w_{jk}(t) + \Delta w_{jk}(t+1) \]  
(5.20)

\[ v_{ij}(t+1) = v_{ij}(t) + \Delta v_{ij}(t+1) \]  
(5.21)

\[ a_j(t+1) = a_j(t) + \Delta a_j(t+1) \]  
(5.22)

\[ b_j(t+1) = b_j(t) + \Delta b_j(t+1) \]  
(5.23)

\[ \theta_j(t+1) = \theta_j(t) + \Delta \theta_j(t+1) \]  
(5.24)

According to gradient descent algorithm, used in back propagation neural network, the error calculated at every \( t \) is used to adjust the parameters in the direction that reduces the error value. This procedure is continued until either the given number of iterations is reached or the criterion for minimum MSE is satisfied. Once the stopping criterion is met, the parameters of the WNN are used as model for the speaker and the same WNN will be used in the verification phase.
5.3 WAVELET TRANSFORM & NEURAL NETWORK

An important problem in speaker recognition systems is to determine a representation that is well adapted for extracting information content of speech signals. In general, to represent the speech signal in a better manner, the signal has to be transformed to a different domain. That means, any system can be able to classify the signals belong to different classes in the transformed domain.

Wavelets are mathematical functions that divide the data into various frequency components. Each frequency component is analysed with a resolution matched to its scale. The analysis of different frequency components of data at different scales or resolutions is called multi resolution (Meyer 1993). Wavelet functions localize a given signal in both space and scaling domains. A family of wavelets can be constructed from a mother wavelet. By stretching and compressing the mother wavelet, various versions of same mother wavelet can be generated. These dilated and translated wavelets are automatically adapting to both high frequency and low frequency components of the signal. Therefore, wavelets can provide approximation of any signal. Any small change in the wavelet representation produces a correspondingly small change in the original signal, which means local mistakes will not influence the entire transform. The wavelet transform is suited for non stationary signals. Hence it is useful in representing the speech signal. Wavelets are well-suited for representing data with sharp discontinuities.

Filters of different cutoff frequencies are used to analyze the signal. In Discrete Wavelet Transform (DWT), scales and positions are chosen as powers of two. An efficient way to implement this scheme using filters was developed by Mallat (1989). Given a signal \( f(t) \) of length \( N \), the DWT consists of \( \log_2N \) stages at most. Initially the signal is passed through LPF
and HPF whose bandwidths are half of the maximum bandwidth of the signal in order to obtain approximation coefficients CA1 and the detail coefficients CD1. These are the outputs of LPF and HPF with the signal input to obtain approximations and details respectively, followed by dyadic decimation. The next step splits the approximation coefficients CA1 into two parts using the same scheme, replacing the signal by approximation and producing CA2 and CD2, and so on. This technique is most effective when it is applied to the detection of short-time phenomena, discontinuities, or abrupt changes in the signal. Thus, wavelet transform of speech signal provides a very good time and frequency representation.

5.3.1 Wavelet Transformed MFCC

In speech signal processing, MFCC is a representation of the short-term power spectrum of a sound, based on a linear discrete cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. In MFCC, the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound. In this work, the wavelet transformed MFCC coefficients are obtained by decomposing the speech signal into different resolution levels and then MFCCs are extracted from the wavelet channels. This represents the frequency characteristics of speech signal at different resolution levels. Thus the combination of wavelet transform and MFCC can represent sound signals in an efficient way.

5.3.2 DWT & Neural Network

This approach is an alternate method of wavelet neural network. In this method, features are extracted by applying a Discrete Wavelet Transform (DWT), while a Neural Network (NN) is used for modeling the speaker and
for handling the task of decision making. Combination of DWT and ANN is used by Avci & Avci (2008) in the form of DWNN (Discrete Wavelet Neural Network) for digital modulation recognition. They have compared the performance of DWNN and DWANFIS (Discrete Wavelet Adaptive Network based Fuzzy Inference System) in digital modulation recognition.

The processes involved in this method are demonstrated with the block diagram shown in Figure 5.2. The input speech signal is preprocessed and the features are extracted from it. In feature extraction, different methods are followed.

1) Extracting MFCC features with DWT in the place of DCT (Hsieh et al 2003).

2) Applying DWT to speech signals and taking MFCC for the various channels results from DWT (Mahmoud Abdalla & Hanaa Ali 2010).

3) Applying DWT and then performing MFCC extraction for noisy speech (Shafik et al 2009).

In this research, wavelet transform is applied on the speech signal and the MFCC features are extracted for the wavelet transformed channels. These features are used to train a neural network to model the target speaker characteristics. The neural network used here is a back propagation neural network. It has input layer, a hidden layer and an output layer. The structure of the BPN is shown in the Figure 5.3 with log sigmoid as activation and output layer function. The network parameters are tuned to produce minimum mean square error between the target and the actual output. Gradient descent algorithm is used for training the network.
In the verification phase, the test feature vectors are applied to the neural network which is trained already. Based on the output of the neural network, the decision is made whether the claimed identity is correct or not. This step is performed in the block evaluation of neural network shown in Figure 5.2.

**Figure 5.2  Speaker verification system with wavelet transform and neural network**
Figure 5.3 Structure of Back Propagation Neural Network

5.4 FUZZY WAVELET NEURAL NETWORK

Recently, a combination of fuzzy technique and wavelet transformation was proposed in some papers. The method proposed by Thuillard (2000) gives the idea of multi-resolution fuzzy model, but it was not used for an experimental setup and for system classification. In Daniel & Ho (2001) proposed a FWN, which combined WNN into the consequents of the rules in Takagi-Sugeno (TS) fuzzy model. As a result, the model accuracy and the generalizing capability can be improved. But the model structure became more complicated.

Lin & Wang (1996) constructed the fuzzy model for function approximation in which the wavelet function is used as activation function. Jiao & Liu (1998) also developed the FWN for interval estimation with interval learning algorithm.

Wang Jun et al (2005) proposed a fuzzy wavelet network (FWN) using B-spline wavelet as membership function for approximating nonlinear
functions. In our research, the multi resolution capability and the good generalization performance of WN is combined with the properties of scaling functions which are fuzzy membership functions. The scaling function can be interpreted as fuzzy membership functions only when they are symmetric, positive everywhere and with single maxima. The proposed FWN could overcome the disadvantage of ANN in low convergence rate and the disadvantage of the wavelet network in needing too many training data.

5.5 PROPOSED SYSTEM USING FUZZY WAVELET NETWORK

The wavelet theory is combined with the fuzzy based neural network which leads to construct Fuzzy Wavelet Network (FWN). The advantage of fuzzy wavelet network is that the membership functions can be easily merged or divided using the multi resolution properties and the rules can be evaluated during learning.

The features are extracted from speech and are used to form the rules of Fuzzy based wavelet network. The verification or classification block verifies whether the input test data is uttered by the claimed speaker or not. This is carried out by applying the test data to the claimed speaker’s FWN model and decision is taken based on the result.

5.5.1 Training Algorithm for Fuzzy Wavelet Network

In Takagi-Sugeno fuzzy model, a set of fuzzy rules can be described by

\[ R_i \quad \text{if } x_1 \text{ is } A_{1i}, \text{ and } ..., \text{ and } x_p \text{ is } A_{pi}, \text{ then } y \text{ is } b_i \]  \hspace{1cm} (5.27)

where \( A_i \) is the fuzzy set characterized by the Gaussian type or triangular type membership function. \( x_i \) is \( i^{th} \) input, \( y \) is output variable, \( p \) is the number of
input variables. In this rule, if wavelet functions with symmetry, positive and singular value, i.e., Daubechies wavelets are selected as membership function \( A_k \) with fine resolution. It can capture the local behaviour of the function accuracy. So, it leads to different fuzzy sub-model with variable dilation and translation factors. Assuming when dilation factor is \( j \), corresponding fuzzy sub-system is expressed by \( i \). \( k \)th fuzzy rule can be defined as

\[
R_i^{(k)} \text{ if } x_i \text{ is } A_{j_i, t_i^j}, \text{and...and } x_p \text{ is } A_{j_p, t_p^j}, \text{then } y \text{ is } b^k_i
\]  

(5.28)

where \( A_{j_i, t_i^j} \) corresponding to Daubechies wavelet membership functions with \( j \) scale factor and \( t_i^j \) translation factor of input \( x_i \); It is expressed as \( \mu_{j_i, t_i^j}, l = 1, ..., p; b^k_i \) is the output of local model for rule \( R_i^{(k)} \). A typical fuzzy subsystem can be obtained

\[
\tilde{y}_i = \frac{\sum_{k=1}^{c} b^k_i \prod_{i=1}^{p} \mu_{j_i, t_i^j}(x_i)}{\sum_{k=1}^{c} \sum_{i=1}^{p} \prod_{i=1}^{p} \mu_{j_i, t_i^j}(x_i)} = \sum_{k=1}^{c} b^k_i \mu^k_i(x)
\]

(5.29)

where \( c \) is the number of fuzzy sub system.

The whole output of the system is

\[
\hat{y} = \sum_{i=1}^{c} \hat{y}_i = \sum_{i=1}^{c} \sum_{k=1}^{N_i} b^k_i \mu^k_i(x)
\]

(5.30)

This output is a linear sum of dilated and translated wavelet functions. This means that under this form the Sugeno fuzzy model is equivalent to a multi resolution function mode.

Let
\[
\hat{\mu}_i^k(x) = \frac{\prod_{l=1}^{p} \mu_{j_{l},i_{l}}(x_l)}{\sum_{i=1}^{c} \sum_{k=1}^{N_i} \prod_{l=1}^{p} \mu_{j_{l},i_{l}}(x_l)} \tag{5.31}
\]

where \( 0 \leq \hat{\mu}_i \leq 1, \sum_{i=1}^{c} \hat{\mu}_i^k = 1 \) \( \tag{5.32} \)

Number of fuzzy rules in subsystems are \( N_{1},...,N_{c} \) respectively. According to fuzzy rule and (5.6) and (5.7), FWN system can be described as in Figure 5.4 by a multilayer forward network.

The \( c \) rules have been divided into a rule in which membership function is dilation function \( \varphi_{i}(x) \) and \((c-1)\) rules in which membership function is wavelet function \( \psi_{i}(x) \). Equation (5.30) can be described as follows:

\[
\hat{y} = \sum_{i=1}^{c} \hat{y}_i = \sum_{k=1}^{N_i} b^k \varphi^k(x) + \sum_{i=2}^{c} \sum_{k=1}^{N_i} b^k_{i-1} \psi^k_{i-1}(x) \tag{5.33}
\]

where

\[
\varphi_{j_{l},i_{l}}(x_l) = 2^{j_{l}/2} \varphi(2^{j_{l}} x_l - t^{kl}_{i_{l}})\\
\psi_{j_{l},i_{l}}(x_l) = 2^{j_{l}/2} \psi(2^{j_{l}} x_l - t^{kl}_{i_{l}}) \tag{5.34}
\]

5.5.2 Structure of Fuzzy Wavelet Network

The fuzzy wavelet network is composed of four layers. In the first layer the number of nodes is equal to the number of input signals. In second layer each node corresponds to one linguistic term. For each input signal
entering the system, the membership degree to which input value belongs in the fuzzy set is calculated. To describe linguistic terms scaling and wavelet functions of Daubechies wavelet are used. The third layer and fourth layer are consequent layer and output layer respectively. The fuzzy set rules are validated by these layers.

![Figure 5.4 Sugeno Adaptive Fuzzy Wavelet Network](image)

The network is trained using hybrid learning method. The network parameters in the consequent layer and output layer are updated in learning. In Figure 5.3, $y$ is the output signal of fuzzy wavelet network. The adaptation of parameters is based on the error. The error, which is the difference between the desired and original output is given by

$$ E = \frac{1}{2} (y - y^d)^2 $$

(5.35)

Here, $y$ and $y^d$ are current and desired output values of network. The hybrid learning procedure involves two steps. The first step is the forward pass, in which, functional signals go forward till layer 4 and the consequent parameters are identified by the least squares estimate. In the second step-backward pass, the error rates are propagated backward and the
premise parameters are updated by the gradient descent learning method. Correspondingly, the parameters of fuzzy wavelet network are calculated as given in Equations (5.27-5.34).

\[ w_j(t+1) = w_j(t) + \gamma \frac{\partial E}{\partial w_k} \]
\[ a_i(t+1) = a_i(t) + \gamma \frac{\partial E}{\partial a_i} \]
\[ b_k(t+1) = b_k(t) + \gamma \frac{\partial E}{\partial b_k} \]

(5.36)

Here \( \gamma \) is learning rate; \( k \) is number of neurons of hidden layer of wavelet neural network. Using (5.36), the correction of parameters of fuzzy wavelet neural network is carried out. The proposed adaptive fuzzy wavenet, which results from the direct adaptive approach, has the ability to tune the adaptation parameter in the THEN-part of each fuzzy rule during real-time operation.

![Figure 5.5 Speaker Verification System using FWN](image-url)
5.5.3 Implementation of the Proposed System

TIMIT Acoustic-Phonetic Continuous Speech data base is used in this work. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech waveform file for each utterance. The system’s performance is evaluated for 50 speakers.

The speech signal is divided into frames of 30ms with overlapping of 10ms. For each frame, 13 Mel Frequency Cepstral Coefficients and its first and second derivatives are calculated. Thus, totally 39 coefficients per frame are obtained. These feature vectors are used as inputs for the system. 80% of the input data are used as training data from which the fuzzy rules are formed and the remaining data are used as test data. The entire training data set has been classified by means of subtractive clustering. Subtractive clustering finds the cluster centers based on the density of surrounding data points. For that it requires the radius (=0.5) which specifies the size of cluster in each of the data dimension. The reject ratio is set as 0.15. The Sugeno fuzzy model for the proposed speaker verification system is generated for the input MFCC vectors of 39 dimensions.

An initial Fuzzy Wavelet network system is generated for training. This is accomplished by extracting a set of rules that models the data behaviour. The model is generated with Daubechies wavelet membership functions. The consequent equations are derived from the antecedent membership functions of the rules. In, the matching /verification block, the test feature vectors are applied to the fuzzy wavelet neural network. The output of the FWN indicates that the applied test input is corresponding to which speaker. The decision to accept or reject the claim is taken from the output of the FWN.
5.6 SUMMARY

In this chapter, the excellent properties of wavelets like multi-resolution and approximation of nonlinear functions are utilized along with the concepts of fuzzy logic and neural network. The speaker verification system is developed using wavelet neural network and fuzzy wavelet network. Both systems perform better than conventional neural network.