A Noise-Robust Speech Recognition System Based on Wavelet Neural Network*

Yiping Wang and Zhefeng Zhao

College of Information Engineering, Taiyuan University of Technology,
030024 Taiyuan, China
ping6998@163.com

Abstract. Aiming at the problem that the performance of speech recognition system will drop severely in noisy environments, this paper proposed a recognition system that has excellent anti-noise performance. The feature parameters of front-end are ZCPA (Zero-Crossings with Peak-Amplitudes) feature and the recognition network of back-end is wavelet neural network. Mexican Hat wavelet was used to replace the Sigmoid or Gaussian basis function of feed-forward neural network. The network structure was three layers. Training network weights, determining the position and scale factor of wavelet function were processed respectively. And the information of training samples was made full use of when estimating the position factor. Hence, the network could converge rapidly and it also avoided the problem of wavelet "dimension disaster". A 50 isolated-word person-independent speech recognition system using BP network or wavelet neural network as recognition network was simulated under different SNRs in this experiment. The experimental results showed that recognition rates using wavelet neural network are higher than using BP network.

Keywords: speech recognition, wavelet neural network, ZCPA feature, mexican hat wavelet.

1 Introduction

Wavelet neural network[1] with a lot of excellent performance is a new feed- forward neural network based on wavelet analysis theory. This paper presented the structure and design method of wavelet network, it also analyzed advantage of this network for speech recognition. Traditional features such as LPC, MFCC have got excellent recognition results under clean environment, but their performance will deteriorate severely in noisy condition. This paper introduced a new anti-noise speech feature parameter named ZCPA and combined with wavelet neural network, a 50 words person-independent speech recognition system can be built. The experiment results showed it has excellent performance with this system in both clean and noisy environments.

* This work was partially funded by Shanxi Province Foundation for Returness(2009-31), International Scientific and Technological Cooperation Projects in Shanxi Province(2008-081026).

H. Deng et al. (Eds.): AICI 2011, Part III, LNAI 7004, pp. 392–397, 2011.

2 Presentation of ZCPA Feature

ZCPA feature[2] represents the signal frequency characteristics via speech signal zero-crossings rate. Signals of different frequency have different zero-crossings rate, and the characteristics of frequency can be reflected by extracting the intervals between adjacent zero-crossings of signals. Fig.1 illustrates the principle of ZCPA system. This system consists of band-pass filters, zero-crossing point detectors, peak value detectors, nonlinear compression and frequency receivers. The band-pass filters are composed of 16 FIR filters, which are used to simulate the cochlear base membrane. While the others simulate the movement of audio nerve fiber in the cochlear, so such features have strong anti-noise performance. The intensity information is obtained from the peak value detector and then compressed nonlinearly. The frequency receiver compounds the frequency and peak value information. Finally, the information through 16 filters is combined to form the whole output with the name of ZCPA feature.

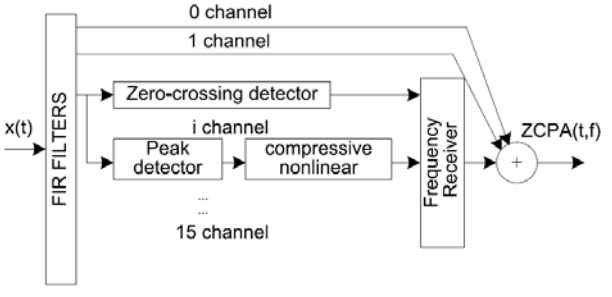

Fig. 1. ZCPA system diagram

3 Constructing Wavelet Neural Network Model for Recognition System

3.1 Network Topology

The diversity and complexity of constructing a wavelet function determine that of constructing a wavelet network. The network structure can vary with different applications, and it is not fixed structure. Present studies about network constructing rules include continuous wavelet transform, orthogonal wavelet transform, wavelet frame and wavelet basis fitting etc. The system in this paper constructs the network according to the wavelet basis fitting. It is known to us that a signal function $\hat{f}(t)$ can be fitted via linear combination of selected wavelet basis [3] :

$$\hat{f}(t) = \sum_{k=1}^{K} \omega_k \phi\left(\frac{t-b_k}{a_k}\right) \tag{1}$$

Where, b_k is position factor, a_k is scale factor and k is the number of basis function. The network topology can refer to Fig. 2:

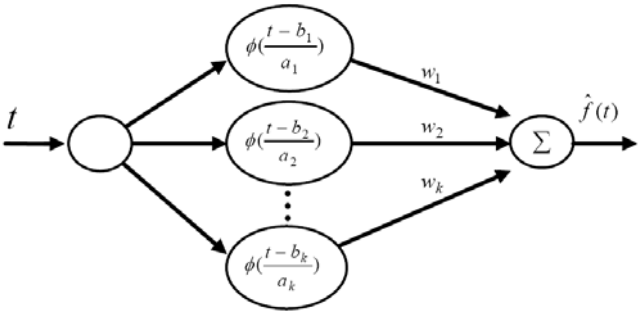

Fig. 2. Wavelet network topology with single input and single output

In Fig.2, a_k and b_k are variable, only w_k acts as the weight between the hidden and output layer, and $\phi\left(\frac{t-b_k}{a_k}\right)$ can be regarded as the output value of input nodes. The network structure is similar to the traditional forward multilayer perceptron, and the key difference exists in hidden layer function.

The system in this paper presented a three-layer neural network with multi- input and multi-output. The inputs are the feature parameters of every word, and their dimension is 1024, so the number of input nodes is 1024. The number of hidden nodes depends on words number recognized. Here, 10 nodes correspond to 10 words and 50 nodes correspond to 50 words. It is a direct connection with no weight between input layer and hidden layer, the same as that between hidden layer and output layer, except that they have weights. The number of output layer nodes equals to the classification number of words. So this system has the same nodes in hidden and output layer. The form of the basis function of every hidden node is uniform, while the scale and position parameters vary with nodes. Strictly speaking, this network is not based on the wavelet mathematical analysis theory, in fact, it is used to approximate function in wavelet combination of certain form. In this system, it is used for word recognition.

3.2 Selecting Basis Function

As for selecting method of wavelet basis function, there are no uniform rules in theories. Generally speaking, it can be determined by experiential and practical applications, and also can refer to the experience of wavelet analysis. In the experiments, several different wavelets have been tested. Finally, Mexican Hat[4] wavelet is selected.

3.3 Determining the Parameters of Network

Parameter of hidden nodes the parameters to be determined are number of hidden nodes, scale factor, position factor and connection weights between all hidden and output nodes. Estimation of the number of hidden nodes also has no uniform rules in theory.Because both scale and position parameters can be set as continuous values,

infinite number of wavelet functions $\phi_{a,\,b}$ can be obtained and they are uncorrelated, so the wavelet coefficients are redundant seriously. We can select part of wavelet functions according to practical application, so the created wavelet has less hidden nodes and can also get good generalization ability. In this paper, the hidden nodes number equals to the output layer number, (i.e. the word classification number) . Otherwise, a bias which has fixed value of 1 should be added to hidden layer. This bias factor also should be connected to all the output nodes in order to estimate weights.

Calculating scale factor and position factor Several methods can be used to estimate the parameters of wavelet network, such as BP network training method and orthogonal least square method which optimizes these parameters simultaneously . The network topology of this paper makes it possible that the training of scale and position can be separated from the weights training. One common method of estimating a_k and b_k is clustering. But clustering algorithm just assigns the given vectors into several finite classes according to certain distortion measure. However, this method does not take full advantage of the information of training samples. The clustering algorithm of K-Means has been used in this experiment to estimate the parameter b_k , but the recognition results are not satisfying. Because the given training samples have involved the corresponding classification information of every training feature, these information can be used to estimate the position parameter b_k . The hidden exciting function of wavelet network is a local function, and it has strong approximation ability to the function with large difference in localness. The same as the features with large difference, they can be classified properly. In recognition network, we hope the output of all training samples corresponding to a certain word after they get through the hidden node which is determined by its position factor can get the biggest value. In other words, the more adjacent to position factor b_k , the bigger the output of the kth node will be. Hence, for all the training samples corresponding to a certain word, their centroid can be calculated to be a position factor. Once a position factor has been estimated, there will be a scale factor corresponding to it. There are usually two methods to calculate scale factor and position factor:

- **Method 1.** $a_k = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left\|x_k - a_k\right\|^2}$, where x_k is feature vector, a_k and x_k have the same dimension.
- **Method 2.** experiential formula $b_k = \sqrt{\frac{1}{1+\sqrt{2}}\sum_{k=1}^{K}\left\|x_k - a_k\right\|^2}$ proposed by Stokbro is also available. Where x_k , a_k can refer to Method 1. In this experiment, the first method was adopted.

Training network weights LMS method was adopted to train the weights between hidden and output layer in this paper. Supposed ϕ_k as the output of hidden nodes. The basis function from hidden to output layer can be written as

$$\hat{f}(t) = \sum_{k=1}^{K} \phi_k c_k \tag{2}$$

and it can be denoted in form of matrix as $\boldsymbol{Y} = \boldsymbol{W\Phi}$. Here, $\boldsymbol{Y}$ is the output matrix of output layer, $\boldsymbol{W}$ is the weight matrix to be figured out, and $\boldsymbol{\Phi}$ is the output matrix of hidden layer. The aim of LMS method is to minimize the mean square error between desired and real output of network, that is to say $\|\boldsymbol{Y} - \boldsymbol{W\Phi}\|^2$ is least. According to the differential method, matrix W can be obtained from the formula

$$\boldsymbol{W} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{Y} \tag{3}$$

Hence, compared with BP network: firstly, the LMS method can avoid training the network iteratively, it is to say that it can avoid the network spending a longer training time to reach a certain precision. Secondly, it can avoid too much hidden nodes. Lastly, the local minimal problem that is result from the random initial value can be avoided. Training weights by LMS method only has one matrix multiplication, so it needs less training time. And by adding the hidden nodes it can meet the practical requirement. The additional nodes will have little effect on the training time. In this system, the number of input nodes is 1024, generally speaking, wavelet network with high dimension will lead to a “dimension disaster” , which means with the increasing of dimension of input and number of training samples, the converge rate of network will descend severely. In this paper, Training network weights, determining the position and scale factor of wavelet function are processed respectively, so it will avoid the “dimension disaster” effectively due to high dimensional wavelet network.

4 Process of System Implementation and Results

An isolated person-independent recognition system has been implemented in C++ in this experiment. Speech data of 50 words with the pronunciation of 16 persons were used in the experiments. For each word, each person will speak 3 times. Speech data of 9 persons were used to train the system, and the other speech data of 7 persons were used in recognition. One speech sequence of each word for one person was a training file, so there were 1035 files to train the network, and 1050 test files were for recognition. As for the feature extraction, the sampling rate of speech signals was 11.025kHz, and frame length was 10ms, so there were 110 samples for one frame. And the frame shift was 5ms. 1024dimensional speech features were obtained after the ZCPA feature was processed via time and amplitude normalization. These 1024-dimensional features were the inputs of the wavelet network, so the number of input nodes is 1024. After this experiment, a BP network speech recognition system has also been implemented, the files for training and recognition were the same as the wavelet network totally. Tab. 1 is the comparing results of recognition performance.

Table 1. Recognition rate of 50 words based on zcpa feature(%)

SNR(dB)	15	20	25	30	clean
Wavelet network	89.71	91.71	93.43	93.33	94.29
BP network	59.81	68.57	71.24	72.76	73.43

From the table, we can see that the recognition rate of the system constructed by ZCPA feature and wavelet neural network is high in clean condition, and even in noisy environment (in this experiment Gaussian white noise was added in the data) , the recognition rate descends little. So we can make a conclusion that the system has a very good anti-noise performance. And compared with BP network, training speed of the system is very fast.

5 Conclusion

Wavelet network uses wavelet function as the exciting function of neuron, and it integrates the advantage of neural network and wavelet decomposition in function approximation. And because of the localized characteristics of wavelet basis function, when given a certain number of samples, the wavelet network will have stronger discrimination ability and need less training time than other ordinary neural network. Applying wavelet neural network to anti-noise speech recognition is very promising.

References

1. Yong, X., Aina, Q.: Noise robust speech recognition based on improved hidden markov model and wavelet neural network. Computer Engineering and Applications 46, 162–164 (2010)
2. Zhang, X., Li, H.: A passwords recognition system based on zcpa feature parameter. Electronic Technology 7, 27–29 (2010)
3. Hou, X.: Noise-robust speech recognition based on wavelet network and RBF network. Computer Engineering and Applications 45, 150–152 (2009)
4. Steven, G., Pan, A., Seddeik, Y.M.: A feature based technique for face recognition using mexican hat wavelets. In: PACRIM, pp. 792–797 (2009)