International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.5 (2014), pp.187-196
http://dx.doi.org/10.14257/ijsip.2014.7.5.16

Voice Activity Detection Algorithm based on Improved Radial Basis Function Neural Network

Bao-yuan Chen, Ya-qiong Lan, Jing-yang Liu, Zi-he Li and Xiao-yang Yu

The higher educational key laboratory for Measuring & Control Technology and Instrumentations of Heilongjiang Province Harbin University of Science and Technology,Harbin 150080,China
chenbaoyuan@126.com

Abstract

Voice activity detection (VAD) is the key of voice recognition, voice synthesis and speech-sound enhancement.For the sake of improve the accuracy and robustness of speech endpoint detection system. Combining the advantages of adaptive genetic algorithm (AGA) and improved radial basis function network (RBF) defects in existing learning methods. This paper presents a comprehensive detection method-- Adaptive genetic algorithm radial basis function network. This method uses adaptive genetic algorithm to simultaneously optimize the center, the width and the structure of RBF network. The method using wavelet analysis to extract the characteristics of the speech signal, use them as an input amount to the radial basis function networks. Establish voice detection system model, this method enhance the accuracy of the detection system and has better robustness.

Keywords: *Voice activity detection, radial basis function network, Wavelet analysis*

1. Introduction

Voice activity detection is that from a signal contains speech determine starting point and end points accurately, Thus only the storage and processing of speech signals effectively, To reduce the computation time and data processing, Enhance the system recognition accuracy, Improve the quality of voice. In nature, the existence of noise is the most common form of white Gaussian noise, which noise is inevitable [1]. The study found Gauss white noise is random and smooth noise. From the spectrums of speech signal and Gauss white noise can be seen: The speech energy is concentrated in 300-3400Hz, howeverthe spectrum of Gauss white noise changed little, the energy distribution of each frequency band is more uniform. This is the theoretical basis for the speech and non speech signal separation and speech endpoint detection.

Bell Laboratory for the first time put forward the technology of speech endpoint detection in 1959, It has hundreds of kinds of methods have been produced. The speech endpoint detection methods can be divided into two categories: based on the characteristics of the signals and based on pattern recognition [2]. The methods based on the characteristics of the signals, was widely research and application because of the advantages of simple and fast.But the existing methods have various characteristics based on limited; Based on the short time average magnitude endpoint detection method has the advantages of simple operation. But it is difficult to distinguish between weak fricatives and final nasals of different. Based on the short-time average zero crossing rate endpoint detection method was better detection of voiceless, but its anti-noise performance is poor. Based on spectral entropy endpoint detection method can effectively distinguish between speech and noise, but part of the detection results

ISSN: 2005-4254 IJSIP

for the voiceless was poor. In order to improve the accuracy of endpoint detection, many scholars proposed combination of speech endpoint detection method of multiple features. Such as the current wider application based on spectrum entropy, short-term zero-crossing rate and distance inverted spectrum endpoint detection method, it was combined features from three kinds of voice for endpoint detection. Trying to overcome the traditional single feature poor noise performance shortcomings, advantage of the characteristic parameters of each, improve the accuracy of endpoint detection. But in the case of low SNR, especially for non-stationary noise, its detection performance deteriorates rapidly [3].

With the development of the theory of nonlinear, neural networks and SVM (support vector machines) nonlinear speech endpoint detection algorithm gradual emergence. To promote voice activity detection accuracy rate gradually increased [4]. Radial basis function (RBF) neural networks, especially because of its simple structure, learning ability and nonlinear mapping ability, has been widely applied. In practical applications, RBF neural network performance parameters have a great relationship with. Therefore, when the use of RBF neural network model for speech endpoint detection, the need to optimize its parameters [5]. A Genetic Algorithm (GA) is a class of evolutionary, adaptive, stochastic algorithms involving search and optimization [6]. It does not contain the problem to be solved held form, it is from altering the genetic configuration to achieve overall optimization. It was optimization method of bottom-up. With less restrictions, global optimization, can adaptively adjust the search direction, etc. It was one of the key technologies of the modern intelligent computing.

In this paper, use genetic algorithm to optimize the RBF network structure. The topology of the network structure, the connection weights, hidden node centers parameters and width parameters as a whole, compiled chromosome, the overall optimization, ensure the generalization ability of RBF network. In order to improve the learning ability of RBF neural network. The simulation results show that this method compared to traditional methods. In common noise conditions effective to reduce the false negative rate, in complex detection performance in noisy environments is still good, strong robustness.

2. Adaptive Genetic Algorithm Radial basis Function Neural Network

The crossover and mutation probability ofadaptive genetic algorithm in the genetic operations were as different fitness value varies. The basic idea was that when the fitness value of each instance of population tends to converge or local optimum, crossover and mutation rate increased. When the value of the fitness function is relatively dispersed, crossover and mutation rate is reduced [7]. Meanwhile, the fitness value is higher than the population average fitness value of the individual, that was the best individual, used a lower crossover and mutation rate, so that it can enter the next generation. Otherwise considered undesirable individuals, improving crossover and mutation rate, so that the individual is eliminated, preventing poor individuals into the next generation [8].

The crossover rate P_c and mutation rate P_m were adjusted to examining the relationship between theaverage fitness value$f_{avg}$and the largest fitness valuef_{max}, P_c and P_m is expressed as:

$$P_c = \begin{cases} k_1 \frac{(f_{max}-f')}{(f_{max}-f_{avg})}, & f' > f_{avg} \\ k_3, & f' < f_{avg} \end{cases} \quad (1) \quad P_m = \begin{cases} k_2 \frac{(f_{max}-f)}{(f_{max}-f_{avg})}, & f > f_{avg} \\ k_4, & f < f_{avg} \end{cases} \quad (2)$$

$f'$was the larger fitness value of two strings for the exchange. f was the fitness valueof individual that will be mutates.General recommended values:k_1=k_3=1， k_2=k_4=0.5.In actual operation, appropriately adjusted k_1, $k_2$to ensure that the value of P_c, P_m in the range of a reasonable.

Neural network control is the use of computers using different connection methods to the individual neurons connected together to simulate the human visual thinking approach in the control process.It has an adaptive, self-organizing, self-learning function, feature identification and information to maximize the use of dynamic process control system provided, be inspired and logical reasoning to achieve the object of the lack of an accurate model for effective control [9].

In 1988, Moody and Darken proposed new neural network architecture, that radial basis function network, belongs to the forward neural network, it can be used to approximate arbitrary accuracy any continuous function. Radial basis function network is a kind of local approximation network, the output for a particular local area there are only a small number of input layer neurons are used to determine network.[10] RBF neural network is a kind of three layer feed forward network, is composed of input layer, hidden layer, the output layer.Network structure was shown in Figure 1.

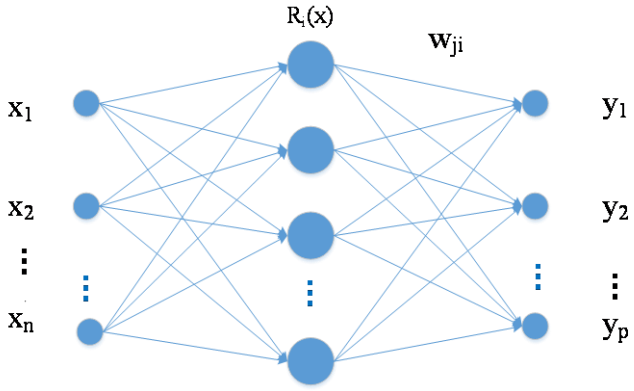


Figure 1. The Network Topology of RBF

The input layerwas composed ofsignal source nodes. The second layer is the hidden layer; the number of units was proportional to the problems described. The third layer is the output layer, it responds to the input mode of action. The basic function is a Gaussian function of the most common, can be expressed as:

$$R_i(x) = exp\left[-\frac{\|X-C_i\|^2}{2\sigma_i^2}\right], i = 1,2,3...,m \qquad (3)$$

$R_i(x)$ Was the output of the i-th hidden layer node? X is the input sample,$X=(x_1,x_2,\dots, x_n)^T$. $C_i$was the Center of Gaussian kernel function of the i-th hidden layer node,it has the same dimension with X. $\sigma_i$was the variable of i-th hidden layer node; it said the normalization constant or base width. m is the number of hidden nodes.

It can be seen from Figure 1:

$$y_j = \sum_i w_{ji} R_i(x) j=1,2,\cdots p \qquad (4)$$

w_{ji} was the hidden layer to the output layer connection weights. y_jWas the actual output sample corresponding to the input network of the j-th output node. P was the number of output nodes.

In RBF neural network, hidden layer node performs a nonlinear variation by radial basis function, it maps the input samples into a new space, and the output layer nodes can achieve linear weighted combination in the new space.

3. Create a Voice Activity Detection Model

The core issue of radial basis function neural network design was to determine the numberof hidden nodes, the center of base functions and other parameters, designed neural network to meet the requirements of the target error as small as possible. To ensure that the generalization ability of neural network, RBF neural network learning rise to a higher level [11].

Consider the case of only one node of the output layer, putformula3 into formula 4:

$$f(x) = \sum_{i=1}^{m} w_i \, exp\left[\frac{-\|x-c_i\|^2}{2{\delta_i}^2}\right] \tag{5}$$

Set the desired output of network is$y^{(d)}(x)$, the energy function of network was defined as:

$$E = \frac{1}{2}\sum_{j=1}^{n}(y^{(d)}(x^{(j)}) - f(x^{(j)}))^2 \tag{6}$$

Set the number of samples is L, put $f(x^{(j)})$ into the above formula:

$$E=\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{L}\xi(x_l^{(j)},\ c_i,\ \delta_i)^2 \tag{7}$$

In this formula,$\xi\left(x_l^{(j)},\ c_i,\ \delta_i\right) = y^{(d)}(x^{(j)}) - \sum_{i=1}^{m} w_i \, exp\left[\frac{-\|x-c_i\|^2}{2{\delta_i}^2}\right]$

When learningthe value of the center and the width parameters, considered $w_i$was a constant, you can get the central value and width parameters formula:

$$c_i(t+1) = c_i(t) - \lambda\frac{\partial E}{\partial c_i} \tag{8}$$

$$\delta_i(t+1) = \delta_i(t) - \beta\frac{\partial E}{\partial \delta_i} \tag{9}$$

λ, β were learning efficiency of central value and width parameters.

In this paper, we use wavelet analysis method. According short-stationary characteristics of the speech signal, each over a period of time the signal is divided into a frame that the characteristics of this period, the signal is stationary.Each frame of speech signal decomposition layer 5, broken down into six different wavelet sub-band. The method shown in Figure 2: The single frame of the speech signal *s(n)* is decomposed into $d_1(n)$ and $a_1(n)$,Then $d_1(n)$ is decomposed into $d_2(n)$ and $a_2(n)$,…,Until after 5 level decomposition to obtain $d_5(n)$ and $a_5(n)$.

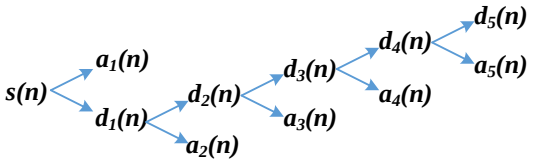


Figure 2. Signal Five Layer Decomposition Diagram

According to the principle of wavelet multi-resolution analysis to know, $d_1(n)$~ $d_5(n)$ and $a_5(n)$ can characterize all of the original signal frequency signals. Calculate the average energy of wavelet coefficients of each layer.

$$E_i = \frac{1}{N}\sum_{n=1}^{N} {s_i}^2\,(n) \tag{10}$$

In this formula,$s_i(n)$ was the wavelet coefficients of one of these sub-bands, N was the number of wavelet sub-band for each frame containing wavelet coefficients.

Calculationthe average of six wavelet sub-band average energy$E_m$and the average energy variance${\delta_m}^2$:

$$\begin{cases} E_m = \frac{1}{6}\sum_{i=1}^{6} E_i \\ {\delta_m}^2 = \frac{1}{6}\sum_{i=1}^{6}(E_{i-}E_m)^2 \end{cases} \quad (10)$$

Completion of the above steps to complete the extraction of the speech signal features. Received a total of eight wavelet sub-band average energyE_1~E_6、E_m、${\delta_m}^2$. The feature vector was composed of them: $X(n) = \left[E_1,\ E_2,\ E_3,\ E_4,\ E_5,\ E_6,\ E_m,\ {\delta_m}^2\right]^T$

4. Algorithm Steps

In order to solve the problem of RBF network structure optimization, we use Boolean vectors U^T=(u_1，u_2，…，u_M), Among them u_i={0，1}，u_i=1Indicate the presence of hidden nodes; u_i= 0Means the corresponding hidden node does not exist.Based on the above theoretical analysis, this section will give specific steps of the algorithm.

The first step randomly generated n group of $U^T$and corresponding center parameters c_i and width parametersδ_i as populations.

The second step, by using the gradient descent method to the N group of initial parameters are pre training, learning linear weights of networkw_i with the method of least square.

The third step, letfnew-max is the maximum value of population fitness function, if fnew-max>fmax, sofmax= fnew-max,and retain the individual corresponds to fnew-max as the best individual. Or else fnew-max<fmax, sofmin= fnew-max, retain the individual corresponds to fnew-max as the best individual. Keep fmax constant.

The fourth step, if the number of hidden nodes in most individuals were the same, this same hidden nodes denoted as m,choose the biggest fitness corresponding individual.Otherwise the implementation of the fifth step.

The fifth step,selection, crossover, mutation adaptive genetic operation on the N set of weights, turn to the second step.

The sixth step, to choose the biggest fitness corresponding individual continues to perform gradient descent method for several times, until meet the precision request center parameters and width parameters.

The seventh step, we put the wavelet analysis to extract feature values as the input of network, with the manual calibration corresponding frames endpoint is the most ideal results output network.Through these to train the neural network.

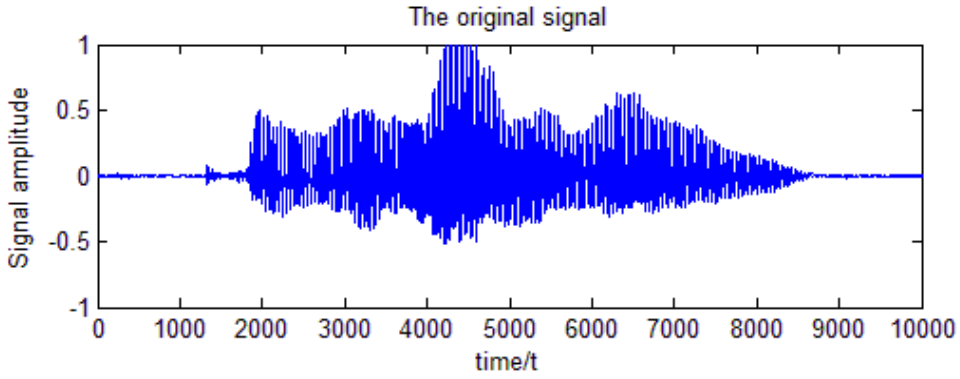
The eighth step, after the training is done, we still used training samples as the input of network. By 0.5 as the threshold, speech frames were greater than 0.5; otherwise, the non speech frames. The decision results compared with the manual calibration results, if the result is not ideal, need to re adjust the network.

The last step, use the trained RBF neural network todetected speech endpoint.Feature extraction with wavelet analysis as the input of the network.By 0.5 as the threshold, to obtainthe results of endpoint detection.

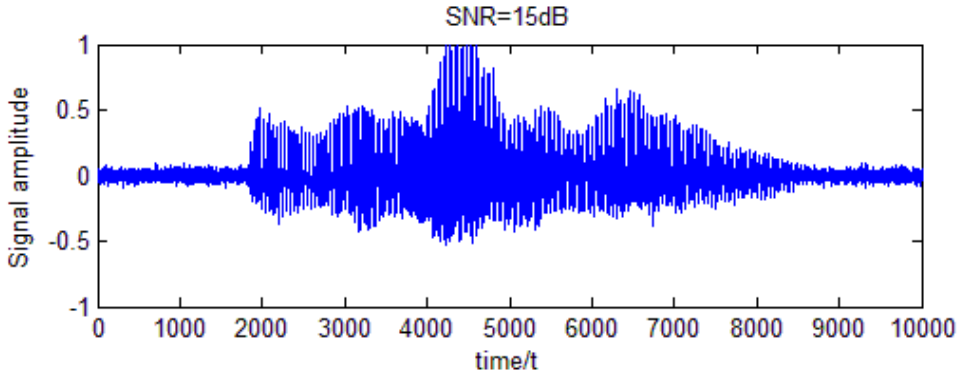
5. Simulation Experiments and Results

Experiments under different noise conditions in accordance with the optimization algorithm using MATLAB simulation.In all experiments, the sampling frequency is44.1KHz, the speech signal is divided into frames, each frame comprising 220 samples. Each voice file manually label to distinguish between speech and background noise,It can be used as a standard test of activity detection accuracy.Noise is artificially added, Gaussian white noise generated by computer simulation, SNR was-5dB, 0dB,5dB,15dB.The processed speech

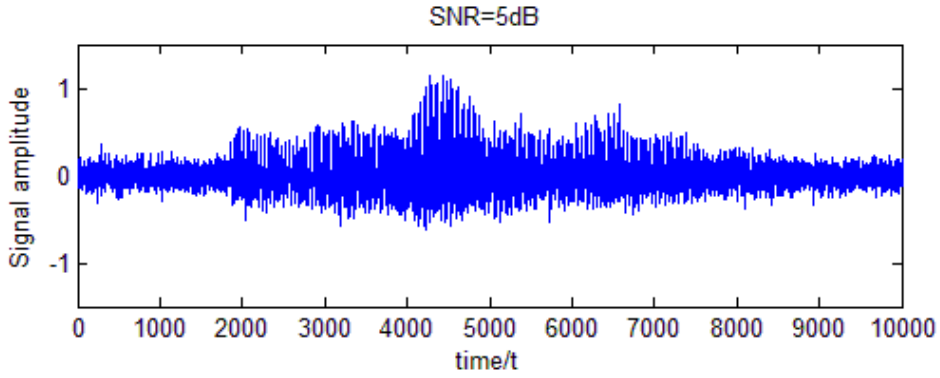
signal as the following Figure.



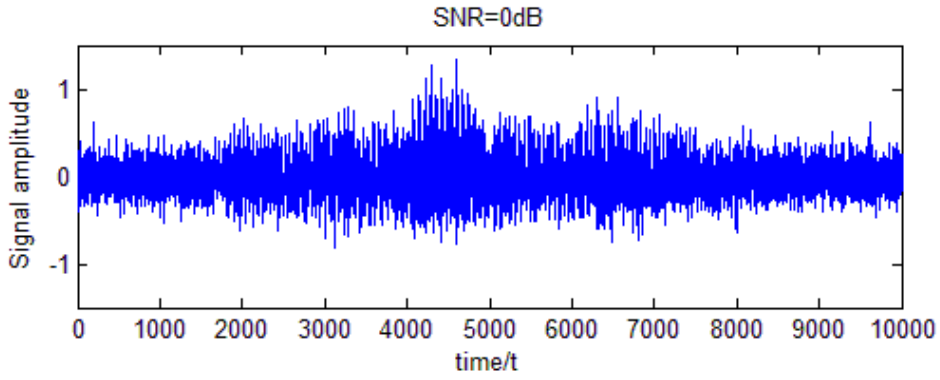
a. Signal without noise



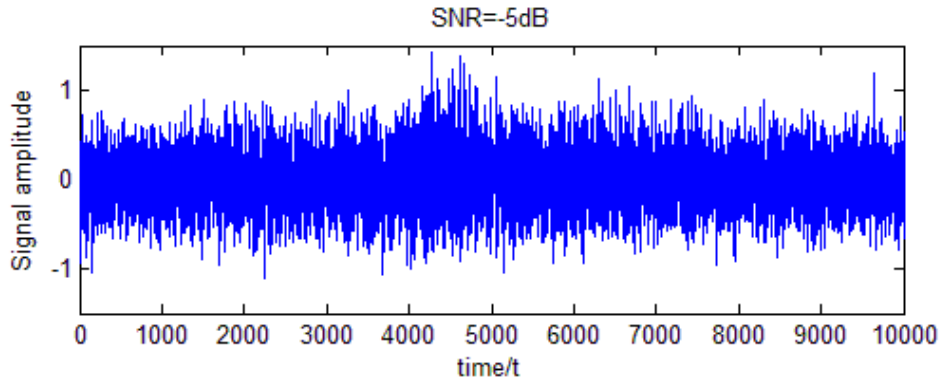
b.Signal with noise (SNR=15dB)



c.Signal with noise (SNR=5dB)



d.Signal with noise (SNR=0dB)



e. Signal with noise (SNR=-5dB)

Figure 3. Signal with Different SNR Noise

As can be seen, as the SNR increases, the speech feature gradually being submerged by noise, cannot even tell the difference.

The simulation result of each voice activity detectionswere shown in Table 1 on condition that the SNR was same.The results shown in the same noise environment,high SNR, all the methods have a good performance.But as the noise increases, the old method of performance decreased significantly.In the case of low SNR (0dB), short-time energy method was basically ineffective.Traditional RBF neural network methods and improved RBF neural network method all had a good detection performance.The improved RBF method has higher accuracy.

Table 1. Result of Voice Activity Detection

SNR/dB	Short-time energy method	Traditional RBFmethod	Improved RBFmethod
Without noise	96.21	97.92	99.24
15 dB	87.35	96.67	97.82
5 dB	72.13	95.95	96.49
0 dB	62.15	95.45	96.01
-5 dB	55.26	93.91	95.37

The non speech signal recognition for speech signal called false, the speech signal was identified as non speech signal called omission. In the practical application to ensure the correct rate at the same time, reduce the omission rate.

From Table 2 that the test results of improved RBF neural network method, the omission rate lower than the false rate, to achieve the desired results.

Table 2. Result of Voice Activity Detection based on Improved RBF Method

SNR/dB	Correct rate	Falserate	Omission rate
Without noise	99.24	0.76	0
15 dB	97.82	1.96	0.22
5 dB	96.49	2.24	1.27
0 dB	96.01	2.48	1.51
-5 dB	95.37	3.02	1.61

6. Conclusion

Through a large number of experimental data shown: Voice activity detection algorithm based onimproved RBFneural networkwas in terms of network optimization with crossover probability and mutation probability adaptively adjusting, speed up the search speed ofRBF neural network for global.

Under the conditions of the same type of noise environment, accuracy was higher than that of the traditional method.In the case of SNR decreases, the correct rate of the traditional method decreased rapidly, adaptive genetic algorithm radial basis function neural network training methods can still maintain good detection accuracy.The detection result is more stable, reliable, and has better practical value. But in the network learning speed is slightly slower than the traditional RBF method, it needs to be improved according to the actual application.

Acknowledgment

Instruments and equipment research project in 2014 of Harbin University of Science and Technology.

Heilongjiang Province College Education Engineering Projects Grant No: JG2013010302.

References

[1] L. Han, B. Wang and S. Duan, "Development of voice activity detection technology", Application Research of Computers, **(2010),** pp. 1220-1226.
[2] J. Fu, S. H. W. Wang, X. L. Cao, *etc.*, "The research on speech endpoint detection algorithm based on spectrogram row self-correlation", Computer Science and Network Technology (ICCSNT), 2nd International Conference on, **(2012)** December 29-31, pp. 212-216.
[3] Q. Guo, N. Li and G. Ji, "A Improved Dual-threshold Speech Endpoint Detection Algorithm", Computer and Automation Engineering (ICCAE), The 2nd International Conference on, vol. 2, **(2010)** February 26-28, pp. 123-126.
[4] L. Ye, Z. H. Yang and H. Y. Guo, "Low bit rate speech coding based on wavelet transform and compressed sensing", Chinese Journal of Scientific Instrument, vol. 31, no. 7, **(2010),** pp. 1569-1575.
[5] T. Wang, H. Wang and H. Xie, "Networked synchronization control method by the combination of RBF neural network and genetic algorithm", Computer and Automation Engineering (ICCAE), The 2nd International Conference on, vol. 3, **(2010)** Februaru 26-28, pp. 9-12.
[6] K. Gkoutioudi and H. D. Karatza, "A simulation study of multi-criteria scheduling in grid based on genetic algorithms", Parallel and Distributed Processing with Applications (ISPA), IEEE 10th International Symposium on, **(2012)** July 10-13, pp. 317-324.
[7] X. Yang, L. Yuan and Y. Wang, "Quantitative Detection for Gas Mixtures Based on the Adaptive Genetic Algorithm and BP Network", Industrial Control and Electronics Engineering (ICICEE), **(2012)** August 23-25, pp.1341-1344.
[8] B. Chen, "Study of FIR Filtering Algorithm Based on the Improved Genetic Algorithm Radial Basis Function (RBF) Network", Journal of Harbin University of Science and Technology, vol. 17, no. 6, **(2012),** pp. 97-101.
[9] J. Fei, Z. Wang, X. Lu, *etc.*, "Adaptive RBF neural network control based on sliding mode controller for active power filter", Control Conference (CCC), 32nd Chinese, **(2013)** July 26-28, pp. 3288-3293.
[10] R. Zheng, "A novel radial basis function neural network for discriminant analysis", Neural Networks, IEEE Transactions on, vol. 17, Issue 3, **(2006)** May, pp. 604-612.
[11] J. Su and Q. Zhong, "Research on the Prediction of Breath Period Signal Based on RFN Network of Self-Adaptive Genetic Algorithm", Electrical and Control Engineering (ICECE), International Conference on, **(2010)** July 25-27, pp. 1798-1801.