ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 4, April 2016

Classification of Speech using Signal Processing and Artificial Neural Network

Dr.T.Jayasree[1], M.E, Ph.D, MISTE
D. Gowri Shankar[2], M.Murugan[3] M.Mathesan[4] S. Rajkumar[5]
[1] Assistant Professor, GCE, Dept of ECE, Tirunelveli, TN.
[2,3,4,5] Final year B.E students, Dept of ECE, GCE, Tirunelveli, TN

Abstract*:* A wide range of researches are carried out in signal processing for denoising. Stress management is important to improve disabled person speech. In order to provide proper speech practice for the disabled children, their speech is analyzed. Initially, the normal and pathological subject's speech are obtained with same set of words. In this project, classification of normal and pathological subject's speech is discussed. The relevant features are extracted using Mel Frequency Cepstrum Coefficient (MFCC) for both words of normal and pathological subject's speech. Dimensionality reduction of features is achieved by using Principle Component Analysis (PCA). Finally the features are trained and tested using Artificial Neural Network (ANN) for classification.

Keyword: speech signal, stress management**,** Mel Frequency Cepstrum Coefficients (MFCC), Principal Component Analysis (PCA), trained

I. Introduction

As the computer technology develops, it promotes the development of society. On the other hand the development of human society entails a higher challenge to the computer development. The communications between humans and computers are wider and deeper communication functions by using mouse, keyboard and touch screen cannot satisfy the quick, accurate and efficient interchange of information. How to send information in a more natural, more efficient and quicker way has become an urgent question. From technology research to daily life, computers are involved in every aspect of people's daily life. Computers are used to accomplish many tasks. Considering this situation, intelligent communication between computers and humans, human-computer interaction, becomes one of the most important research fields. Speech is one of the natural forms of human communication. Since childhood people can express themselves by speech, recognizing others by distinguishing their voices and under-standing others by their speech. People are very good at speaker and speech recognition. The human brain uses neurons and synapses, modified with experience and provides a distributed form of associative memory. Motivated by this speaker and speech recognition systems have been developed.

II. Feature Extraction

After pre-processing we could have fed each 'raw' audio sample to a neural network, to let the neural network learn to recognize a speaker or speech based on that. This could work for speaker recognition at least, but this was not feasible for a number of reasons:

- Once a signal has been sampled, we have huge amounts of data every second.
- Feeding each sample to a neural network, that is one input to the neural network for each sample, which means a huge number of inputs to the NN.
- We can only capture the properties of the signal that are important for recognition
- Instead of using all the samples.
- The spectrogram contains much of the information we need.
- By extracting the features, the same features can be used by multiple tasks.

A .Spectrogram

There is a better representation domain, namely the spectrogram. This representation domain shows the change in amplitude spectra over time. It has three dimensions:

X – axis: Time
Y – axis: Frequency
Z – axis: Color intensity represents magnitude

The complete sample is split into different time-frames (with a 50% overlap).

For every time-frame, the short-term frequency spectrum is calculated. Although the spectrogram provides a good visual representation of speech it still varies significantly between samples. Samples never start at exactly the same moment, words may be pronounced faster or slower and they might have different intensities at different times.

B. Cepstrum

We know that human ears, for frequencies lower than 1 KHz, hear tones with a linear scale instead of logarithmic scale for the frequencies higher that 1 KHz. The Mel frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The voice signals have most of their energy in the low frequencies. It is also very natural to use a Mel-spaced filter bank showing above characteristics. The following approximate formula is used to compute the Mel for a given frequency in Hz:

$$Mel\ (f) = 259510.log\ (1+f/700)$$

For each tone with an actual frequency f (in Hz), a subjective pitch is measured on a scale called the 'Mel' scale. The pitch of a 1 KHz tone, 40 dB above the perceptual hearing threshold is defined as 1000 Mel. The Cepstrum is the forward Fourier transform of a spectrum. It is thus the spectrum of a spectrum, and has certain properties that make it useful in many types of signal analysis. One of its more powerful attributes is the fact that any periodicities, or repeated patterns, in a spectrum will be sensed as one or two specific components in the Cepstrum. If a spectrum contains several sets of sidebands or harmonic series, they can be confusing because of overlap. But in the Cepstrum, they will be separated in a way similar to the way the spectrum separates repetitive time patterns in the waveform.

C. Mel Frequency Scale

It is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The name Mel comes from the word melody to indicate that the scale is based on pitch comparisons. It is a logarithmic scale similar to the way the human ear perceives sound.

As the picture shows, when the frequency is 1000 Hz, in the Mel scale it is also 1000 mel. When the frequency is below 500 Hz, the intervals are smaller compared with frequencies larger than 500 Hz. Generally speaking the Mel scale has a linear relationship with hertz when frequency is below 1000 Hz, and a logarithmic relationship when the frequency is bigger than 1000 Hz. There are many formulas to convert hertz into Mel, but the most popular is:

$$\text{Mel (f)}=2595*\log(1+f/700)$$

Where f is the linear frequency.

D. MFCC

"In sound processing, MFC which means Mel Frequency Cepstrum, is a representation of the short term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency". MFCC stands for Mel Frequency Cepstral Coefficients. The coefficients represent audio based on perception. They are derived form the Mel frequency Cepstrum. It is known that the human ear is more sensitive to higher frequency. The spectral information can then be converted to MFCC by passing the signals through band pass filters where higher frequencies are artificially boosted, and then doing an inverse FFT on it. This results in higher frequencies being more prominent. As the Mel frequency Cepstrum can represent a listener's response system better, MFCC is always considered to be the best available approximation of human ear.

As figure 4.4 shows the whole MFCC processing procedure can be divided into three main steps:

First the FFT is used to convert speech signal from time domain to frequency domain. Map the powers of the spectrum obtained above onto the Mel scale, usisng triangular overlapping windowing. These filters follow the Mel scale. In this thesis 24 filters are used. So the outputs after filtering are:

Signal k,k=1,2,3…..24

Using the filter bank essentially mimics the way the human ear perceives certain frequencies. Finally, take the logs of the outputs, and take the discrete cosine transform (DCT) of the list of Mel log Powers. The MFCCs are the amplitudes of the resulting spectrum. We use log because our ears work in decibels. Discrete Cosine Transform (DCT) will be applied to each Mel Spectrum to convert the values back to real values in the time domain. We take the DCT because it is good for compressing information. In voice box there is a function called Melcept that can calculate the Mel Cepstrum. Actually it can also do the framing and windowing, so the previously mentioned en frame function (described under pre-processing) is not even needed. Where 11025 is the sampling frequency, 'M' means Hamming window and 12 is is the number of Cepstral coefficients, 24 the number of filter banks, 256 the frame length and 256-64=192 is the frame overlap. All these numbers have been discussed earlier. The result MFCCs is an N*12 matrix of 12 MFCCs per each N frame. The number N will naturally vary for each voice sample. The MFCCs are our features that will be fed to the neural network. The figure 4.4 shows the mfcc of the word 'INCRIDIBLE INDIA'.

III. Principle Component Analysis

Principle component analysis (PCA) is one of the statistical techniques frequently used in signal processing to the data dimension reduction or to the data decorrelation. The application of PCA takes advantage of eigenvectors properties for determination of selected object orientation. Various methods can be used for previous object detection. Quality of voice segmentation implies to results of the following process of object orientation evaluation based on PCA as well.

Principal component analysis (Karhunen-Loeve or Hotelling transform) - PCA belongs to linear transforms based on the statistical techniques. This method provides a powerful tool for data analysis and pattern recognition which is often used in signal and image processing as a technique for data compression, data dimension reduction or their decorrelation as well. There are various algorithms based on multivariate analysis or neural networks that can perform PCA on a given data set.

IV. Artificial Neural Network

As we all know the brain perception information much faster than modern computers, like visual and auditory information. An artificial neural network or neural network is an information processing paradigm that is inspired by the way biological nervous systems, especially the brain, process information.

The key elements in this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements working in unison to solve specific problems. A neural network is the same as a human brain which learns by examples. A neural network is same as a human brain which learns by examples. A neural network is normally trained so that a configuration for specific application can be achieved.

A. Node Characteristics

Every node has multiple inputs from others via connections that have associated weights, analogous to the strength of a synapse in the brain. The neuron has m inputs. The bias bk can used to adjust the output to a suitable range before the activation function.

Mathematically the output of the linear combiner (summation element) is given by

$$Vk = \sum^{n}_{j=0} w_{kj}\ xj+bk$$

B. Network Topology

Usually there are the input layer, the output layer, and zero or more hidden layers. The network topology determines the numbers of nodes in each layers, the number of layers in the network, and the methods of connections among the nodes. Generally, there are two types of connections between nodes. In our thesis only the feed forward network, which is a one way connection without loop back is used. It is static, so one input is associated with one particular output. It is appears that the weights of a feed forward neural network are fixed after the training phase and it implanted fixed weight mapping from inputs to outputs.

So the state of any neuron is solely determined by the input-output pattern and not the initial and past states of the neuron, that is, there is no dynamics involved. Consequently, this type of feed forward architecture is classified as a static neural network. The advantage of the static feed forward neural network is that the network can easily be built with a simple optimizing algorithm. It is the most popular neural network architecture in use today. Nonetheless, the static-feed forward neural network also has several network also has several drawbacks for same applications.

- First, it may fail to produce a satisfactory solution because the training data are Insufficient in size.
- Second, the static-feed forward neural network cannot cope well with major changes that were never learned in the training phase.
- Finally, the static-feed forward neural network easily falls into a local minimum and the speed of convergence of the network can be very slow when number of input in the data set is large.

The other type of network architecture, the feedback network has a loop back connection in which the outputs of the nodes can be the input of previous or same level nodes. It is dynamic compared with the feed forward network, so one input produces a series of outputs. As the neurons have one or more feedback link, whose state varies with time, the feedback architecture is also called a dynamic neural network. The presence of a feedback link has a profound impact on the learning capability of the network and on its performance. Because the feedback neural networks have adjustable weights, the state of its neuron depends not only on the current input signal, but also on the previous states of the neuron. The advantage of the dynamic-feedback neural network is that it can effectively decrease the network's input dimension and therefore the training time. However, due to the nonlinear nature of unit activation output characteristics and the weight adjustment strategies, the network stability is often difficult to ascertain.

C. The Topology Used

In this thesis a feed forward network is used. Here a description of all the components of the used network will be given.

Inputs

As the extracted features are compressed by MFCC function, so we can get a fixed number of features or inputs. Here we use 12.

Hidden layer

In this thesis only one hidden layer is used, but different number of nodes or neurons in hidden layer are tested: 16, 32, 64, 128.

Activation function

The activation function used for hidden layer is tansig, and for output layer is satlins. Tansig is a sigmoid activation function and satlins is a stepwise linear function. Fig (1) shows the feed forward neural network.

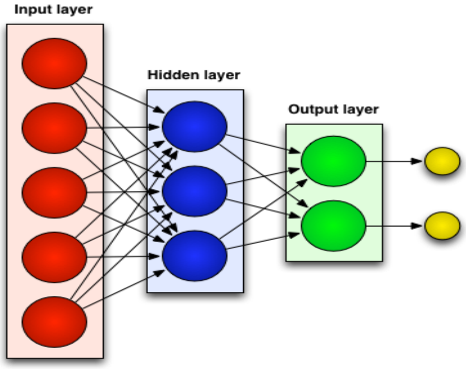

Fig 1. Feed forward Neural Network

Output layer

For different tasks different sizes of output layers are used. For speaker recognition, the output layer represents the speakers. There are tow ways to do this task. One way is that we use one output, with a pure linear activation function. So if speaker one is detected, then the output will be 1. If speaker two is detected the output will be 2 etc. another way is using three binary outputs, if we have three speakers, which are either 1 or -1. For example is speaker two is detected, then the output will be -1, 1, and -1. There is one problem with the first method: for example if we have one linear output, and during the training process we present a voice sample of speaker 1, but the NN detects speaker 3, then the error is 3-1=2, that is a bigger number than if the NN believed it was speaker 2! There is thus a bigger penalty in confusing speaker 1 and 3 than 1 and 2. This error directs the training process which will not work properly. With three binary outputs this problem will not occur. So in this thesis three outputs with satlins activation function (stepwise linear) is used to correspond to every speaker.

D. Neural Network Data Set Analysis

For the operation of ANN, the data are usually divided into three subsets:

Training phase

Before the neural network can realize a specific task, a learning process or training process must be done. During the training phase, the data is usually part of the whole database. After using all the training data once it is called learn cycle or one epoch. The training data will be given to the neural network repeatedly until the weight values are determined.

Validation phase

The main task doing the validation phase is to check the performance of the network and to determine the epoch at which training should be stopped in order to avoid over-fitting or over-training. Normally the best validation performance and validation check are measured when training the neural network. When the neural network is still learning or working correctly, the best validation performance normally decreases while the epoch increases. The validation checks is normally 0 when the neural network is still learning or working correctly, but will increase if the neural network is over-fitting or over-training. Then when the threshold is reached, the neural network stops. If the validation data set indicates that the network is over trained, then the network should be retrained using a different number of neurons and/or parameter values.

Testing phase

After the neural network is trained well, then a testing data set is fed to the neural network to evaluate the neural network performance. Normally the neural network uses all the data including “old data”, which has been used before and “new data”, which has not been used before or only use the “new data” which it has never used before as testing data set. With a different input data set, it will have different errors or errors rate. Here the inputs will be given to the neural network, and the output of the network is produced solely by input output mapping the testing phase.

E. Back Propagation Algorithm

The type of neural network model, known as a "feed-forward back-propagation network". This model is easy to understand, and can be easily implemented as a software simulation.The field of neural networks can be thought of as being related to artificial intelligence, machine learning, parallel processing, statistics, and other fields. The attraction of neural networks is that they are best suited to solving the problems that are the most difficult to solve by traditional computational methods.

The Feed-Forward Neural Network Model:

If we consider the human brain to be the 'ultimate' neural network, then ideally we would like to build a device which imitates the brain's functions. We change the connectivity between the neurons so that they are in distinct layers, such that each neruon in one layer is connected to every neuron in the next layer. Further, we define that signals flow only in one direction across the network, and we simplify the neuron and synapse design to behave as analog comparators being driven by the other neurons.

We now have a feed-forward neural network model that may actually be practical to build and use. Each neuron receives a signal from the neurons in the previous layer, and each of those signals is multiplied by a separate weight value. The weighted inputs are summed, and passed through a limiting function which scales the output to a fixed range of values.

The output of the limiter is then broadcast to all of the neurons in the next layer. So, to use the network to solve a problem, we apply the input values to the inputs of the first layer, allow the signals to propagate through the network, and read the output values. Stimulation is applied to the inputs of the first layer, and signals propagate through the middle (hidden) layer(s) to the output layer. Each link between neurons has a unique weighting value. The algorithm can be decomposed in the following four steps: i)Feed-forward computation ii)Back propagation to the output layer iii)Back propagation to the Hidden layer iv)Weight updates. The algorithm is stopped when the value of the error function has become sufficiently small.

V. Neural Network System Implementation In Matlab

Computer simulation is an experimental method that is always used to test the design of a theory or a method's validity and operational performance. For a speech recognition system the simulation experiment follows the sequence: Capturing original speech signal, processing signal, extracting feature, training neural network, recognition processing and result output. The whole processing sequences is illustrated in Figure 6.2 and all the steps are gone through below. Among all those procedures, there have many parameters which can tune in order to get good result:

- Frame lengths
- Frame overlaps
- Number of MFCC's
- Number of hidden layer neurons

In this project as the frame length and overlap are defined according to many of the previous researches results. So we can say these two are the best choices for this project. In this project 12 MFCC features are given as input to the neural network. Fig (2) shows the steps involved in this project.

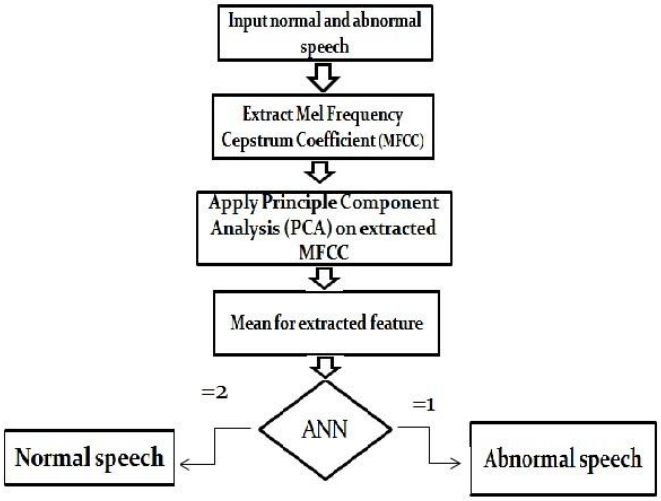
Input normal and abnormal speech
Extract Mel Frequency Cepstrum Coefficient (MFCC)
Apply Principle Component Analysis (PCA) on extracted MFCC
Mean for extracted feature
ANN
=2
=1
Normal speech
Abnormal speech

Fig 2. Neural Network emulation

V1. Simulation Results

A. Capturing the Signal

The word 'INCREDIBLE INDIA' has been recorded by different speakers and the waveforms of these speakers are given below.

Here we can see that at the beginning of the signal and the end of the signal silence takes a lot of space. But there is a very boundary between silence and non-signal. Fig (3) and (4) shows the normal and abnormal input speech signal.

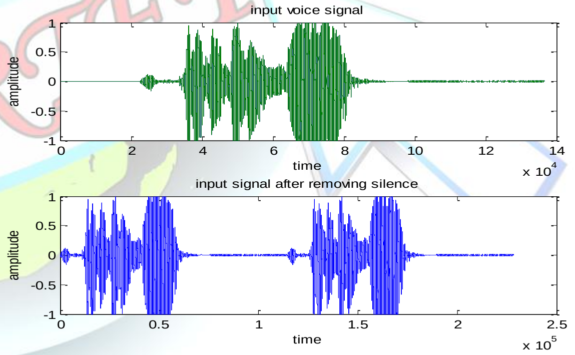

Fig 3. Normal speech input signal by speaker1

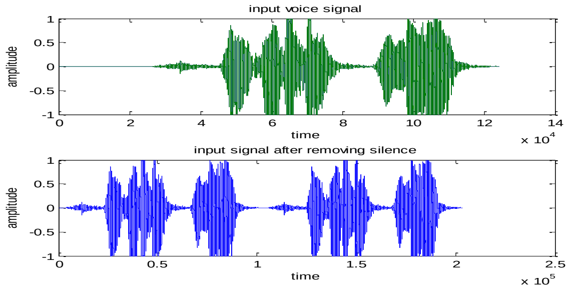

Fig 4. Abnormal speech input signal by speaker1

B. MFCC

The following are MFCC of the word 'INCREDIBLE INDIA' for two different speakers. They are the amplitudes of the resulting spectrum. The output of the MFCC will be given as the input to neural network. Depending on the sampling frequency, the MFCC is 12.Fig (5) and (6) shows the MFCC of normal and abnormal speech.

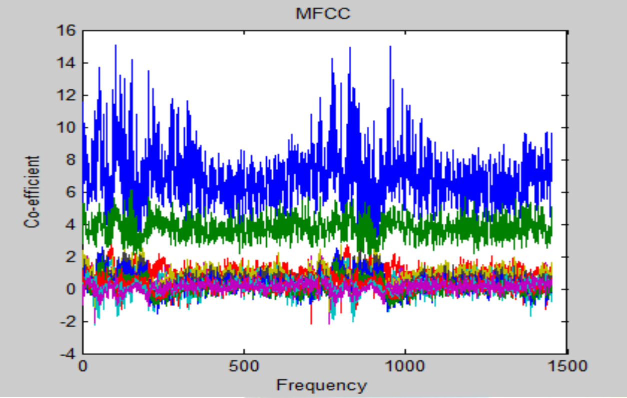

Fig 5. MFCC for normal speech signal by speaker1

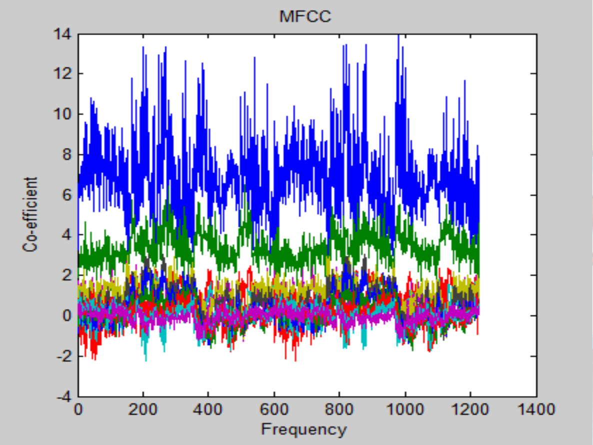

Fig 6. MFCC of abnormal speech by speaker 1

C. Neural Network

In this network, a feed forward neural network istrained by using Back Propagation Algorithm. The input layer consists of 12 mfcc as feature. Only one hidden layer is used with 15 nodes. For speaker recognition, the output layer represents number speakers. Fig (7) shows neural network output for given normal speech.

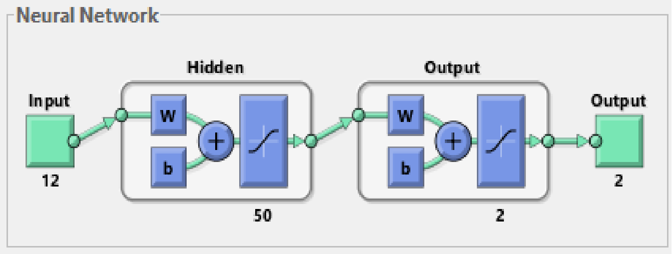

Fig 7. Neural Network

VII. CONCLUSION

Speech recognition using Neural Networks are a hotspot of international academic circles. This project is showing that neural networks can be very powerful speech signal classifiers. The pre-processing quality is giving the biggest impact on the neural network performance. The mel frequency cepstrum coefficients are a very reliable tool for the pre-processing stage, with good results they provide. The multilayer feed forward network with back propagation algorithm is used for training and testing producing better performance results.

This project can be further extended by choosing different feature vectors such as zero crossing variance, standard deviation of the speech signal and other Neural Networks such as radial basis function Neural Network.

REFERENCES

[1]. Han Yi, Wang Guo yin, and Yang Yong. Speech emotion recognition based on mfcc.Journal of Chongqing University of Posts and Telecommunications, 2014.

[2]. Scott Chin, Kelvin Lau, and Lindsey Leu. A speaker verification system. Department of Electrical and Computer Engineering ELEC499a, 2012.

[3]. B. Plannerer. An introduction to speech recognition. Tutorial, university of Munich, 2014.

[4]. Wen Lin. Based on retrofitted mfcc speech recognition system research and design. Master thesis, 2013.

[5]. Amit-Degada. Digital coding of analog signal. Lecture notes, Sardar Vallabhbhai National Institute of Technology.

[6]. Wang Wei-Zen. Research of speech recognition based on neural network. Smaster thesis, 2014.

[7]. MOHD ZAIZU ILYAS, Salina Abdul Samad, Aini Hussai khairul Anuar Ishak, "Speaker Verification using Vector Quantization and Hidden Markov Model".

[8]. Nur Shazreen Yussoff, "Speaker Recognition with Back Propagation Neural Network and Vector Quantization: A Comparative Study"

[9]. Niladri Sekhar Dey, Ramakanta Mohanty, K.L. Chugh, "Speech and speaker Recogniton System Using Artificial Neural Networks and Hidden Markov Model", pp. 322-315, 2012 International Conference on Communication Systems and Network Technologies, 2012.

[10]. Ji-Soo Keum, Yongin-si, Chan-Ho Park, Hyon-Soo Lee, “A new text-independent speaker identification using vector quantization and multi-layer perceptron”.

[11]. Zou Da-yong, “The research of the system of speech enhancement based on bp neural network”, Master thesis, 2006.

[12]. Wang Wei-Zhen, “Research of speech recognition based on neural network”, Master thesis, 2008.

[13]. He Xiang-Zhi, “Speech recognition researching and development”, Computer and Modernization, 3, 2002.

[14]. BRIAN C. LOVELL., “Speaker Verification Using Artificial Neural Networks”.

[15]. Rong Rong. Speech recognition using artificial neural networks. Master thesis, 2005.

[16]. Wu Wei-ye. Arthimetic research based on neural network speech recognition. Master theis, 2009.

[17]. Zou Da-yong. The research of the system of speech enhancement based on bp neural network. Master thesis, 2009.

[18]. Wang Wei-Zhen. Research of speech recognition based on neural network. Master thesis, 3, 2002.