A New Voice Activity Detection Algorithm Based On SOM & LVQ

Yan Liu, and Changchun Bao

School of Electronic Information & Control Engineering
Beijing University of Technology, Beijing, 100022

sharonliu@emails.bjut.edu.cn, chchbao@bjut.edu.cn

Abstract

A new Voice Activity Detection (VAD) algorithm is proposed in this paper[1]. This algorithm is based on competitive networks including of Self-organized Feature Mapping (SOM) and Learning Vector Quantization (LVQ) network. A four-dimensional learning vector is used in this neural network VAD algorithm. The test results show that this VAD algorithm has a significant improvement over traditional methods such as ITU-T G.729B VAD algorithm and BP network VAD algorithm. And it is proved to be a robust and adaptive algorithm under multiplicity noise environments.
Keywords: Voice Activity Detection; Speech Coding; Learning Vector Quantization

I. INTRODUCTION

The voice activity detection (VAD) is used to detect speech from noise background, which is a key algorithm in a silence compression scheme. To distinguish speech from noise background accurately and to ensure the robustness of the VAD under the noise environments is a very challenging problem. The traditional VAD algorithms mainly include the ITU-T G.729B[1]and 3GPP TS 26.094[2] algorithms. These traditional algorithms are implemented with hard computing, the hard computing only gives a solution optimized in local range, so sometimes they are unsuitable in the variable environments. The neural networks algorithm overcomes the shortage of these traditional algorithms and is more robust under noise environments. A new voice activity detection algorithm based on competitive networks is proposed in this paper. This algorithm is based on Self-organized Feature Mapping (SOM)[3] and Learning Vector Quantization (LVQ) network[4]. A four-dimensional learning vector is used in this neural network. And this algorithm is tested in various noise backgrounds and it is proved to be a robust algorithm under noise environments.

The competitive network theory is introduced in this paper and then a new VAD algorithm based on competitive algorithm is detailed accounted. At the last some simulation results and analysis are given.

[1] This work was supported by National Natural Science Foundation of China under Grant 60372063, Natural Science Foundation of Beijing under Grant 4042009 and the Science and Technology Project of Beijing Municipal Education Commission (KM200310005024).

II. COMPETITIVE NETWORKS

A. *Competitive Learning Algorithm*

Competitive network is very effective for pattern recognition. It turns out that this is a crude approximation of biological competitive layers. In biology, when a neuron is activated, it not only reinforces itself, but also inhibits other neurons. Then the competition happens between these neurons and we call the phenomena as 'on-center/off-surround'. The result of the competition is the most active neuron wins and the others lose. The learning algorithm of competitive networks is just in this way. The Fig.1 is the structure of a competitive network.

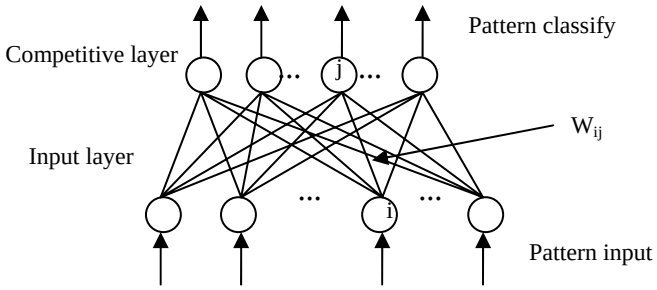

Fig.1. the structure of a competitive network

We define a transfer function that does the job of a recurrent competitive layer:

$$a = \mathrm{compet}(n) \tag{1}$$

It works by finding the index of the largest net input, and setting its output to 1. And all other outputs are set to 0.

$$a_i = \begin{cases} 1, & i = i^* \\ 0, & i \neq i^* \end{cases} (\text{where}, \ n_i{}^* \geq n_i, \forall i, \text{and} i^* \leq i, \forall n_i = n_i{}^*) \tag{2}$$

The prototype vectors are stored in the rows of W. The net input n calculates the distance between the input vector p and each prototype W_i (assuming vectors have normalized lengths of L). The net input n_i of each neuron i is proportional to the angle θ_i between p and the prototype vector W_i:

$$n = Wp = \begin{bmatrix} W_1{}^T \\ W_2{}^T \\ \vdots \\ W_s{}^T \end{bmatrix} p = \begin{bmatrix} W_1{}^T p \\ W_2{}^T p \\ \vdots \\ W_s{}^T p \end{bmatrix} = \begin{bmatrix} L^2 \cos\theta_1 \\ L^2 \cos\theta_2 \\ \vdots \\ L^2 \cos\theta_s \end{bmatrix} \tag{3}$$

The s is the number of neurons. So the competitive transfer function assigns an output of 1 to the neuron whose weight vector points in the direction to the input vector:

$$a = \mathrm{compet}(Wp) \tag{4}$$

B. *Self-Organized Map Learning Algorithm*

Self-Organized Map is an unsupervised learning algorithm and it's one of competitive networks. Assuming the training vector is X_p, which has p samples. The learning algorithm follows the below ways.

(1) Set the initial weight vector random $W_j(0)$, $j = 1 \square M$. And set the biggest calculate steps as $K(K \geq P)$.

(2) Let $k = 1, 2, ..., K$,then do the following calculation:

For each k, find an input sample vector $X(k)$ in turns or randomly. Then use the following formula and $W_j(k-1)$ to calculate the $W_j(k)$:

$$W_j(k) = W_j(k-1) + a(k)\Lambda(j, j^*(k), k)[X(k) - W_j(k-1)],\ j = 1 \square M \quad (5)$$

The $a(\square)$ and $\Lambda(\square)$ are defined as the step function and the neighborhood function.

(3) When $k = K$, the calculation finishes and $W_j(k),\ j = 1 \square M$ was out as the final weight vector of the neurons.

C. *Learning Vector Quantization*

The LVQ network is a hybrid networks. It uses both unsupervised and supervised learning to form classifications. A LVQ network has two layers. Each neuron in the first layer is assigned to a class, with several neurons often assigned to the same class. Each class is then assigned to one neuron in the second layer. The number of neurons in the first later, S^1, will therefore always be at least as large as the number of neurons in the second layers S^2, and will usually be larger.

III. THE VAD BASED ON COMPETITIVE NETWORKS

A. *The Classification Algorithm*

The SOM and LVQ algorithms are used in this new VAD algorithm, which combines with the classical G.729B algorithm. First, the training vectors are sent to the SOM network which be calculated as the input vector. Then the weight vectors of SOM are sent to the LVQ network, which are used as the initial weight vectors of the network. In this way, the errors coming from the effect of the initial weight could be avoided effectively. Figure.2 is the flow chart of the algorithm. The algorithm includes of the speech input, parameters calculation, SOM training, LVQ training and the hangover protection. The hangover protection is used to avoid the false detect. For intend, we set the hangover parameter to 4-5 frames, when the VAD detect the frame change from speech to silence, the VAD flag would not be set to 'noise' directly, until 4-5 frames later and if the detect result is keep to 'noise', the flag will change immediately and vice versa.

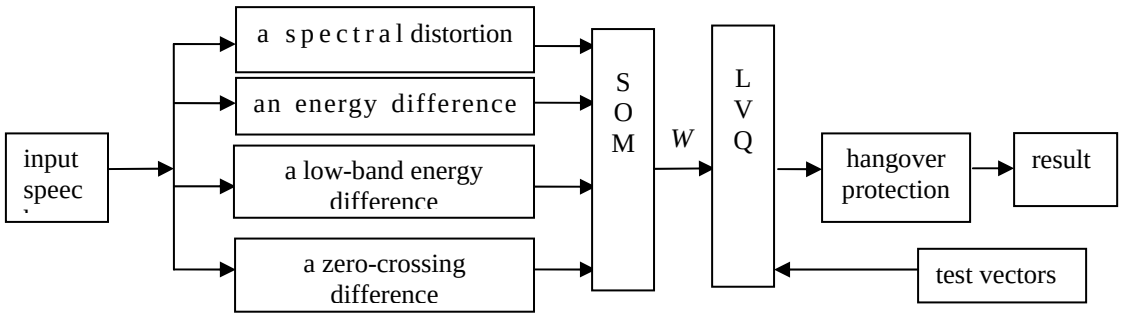

Fig.2. SOM&LVQ VAD algorithm flow chart

We use the following four parameters as the 4 dimensions input vectors, which are also used in the ITU-T G.729B:

A spectral distortion:

$$\Delta S = \sum_{i=1}^{P} (LSF_i - \overline{LSF})^2 \quad (6)$$

P refers to the LPC order and the upper bar notation stand for the moving average.
An energy difference:

$$\Delta E_f = \overline{E}_f - E_f \quad (7)$$

A low-band energy difference:

$$\Delta E_l = \overline{E}_l - E_l \quad (8)$$

A zero-crossing difference:

$$\Delta ZC = \overline{ZC} - ZC \quad (9)$$

The algorithm follows the below 5 steps:
(1) Choice a suitable speech database, transform the speech into the required speech sample form and the divide the database into two parts, the training data and the testing data;
(2) Calculate the parameters of the training data and make them into 4 dimensions input vectors;
(3) Use the SOM network trains the input vectors, then get the weights of each neuron;
(4) Use the weight vectors of SOM as the initial weight vectors of the LVQ network, then use the LVQ network trains the input vectors, then get the best competitive networks;
(5) Use the testing data to test the quality of the competitive networks.

B. Performance Analysis

The sampling frequency of the digital speech signals is 8 KHz, 16 bit and in PCM form, which are used as the input of the competitive networks VAD. The length of each frame is 10ms which equal to 80 samples. The training database contains 5 man speakers and 6 woman speakers whom aged from 11 to 59 (all the samples come from the speech database of speech and audio signal processing lab, Beijing University of technology). The training database has 8000 frames in totally.

The testing data comes from 2 male speakers and 2 female speakers and they are different with the training data speakers (all the samples come from the speech database of speech and audio signal processing lab, Beijing University of technology). The testing database has 1932 frames in totally. And the test environments include the silence background, the factory background, the babble background, the inner of car background and the white noise background. The signal noise ratio is also divided into different levels which are 20dB, 15dB, 10dB, 5dB and 0 dB. The artificial detection results are given firstly. Then we use the result of other detection algorithm to compare with the artificial result. When we mark on the noise speech we considered both of accuracy and practicality and tried to reduce the cutting speech and wrong detection.

The results of the competitive networks VAD are compared with the results of G.729BVAD and the BP network VAD algorithm. In the BP network VAD algorithm, we use the same parameters as the algorithm proposed in this paper. And these parameters are trained in the BP neural network. These test results under different environments fully prove the good quality of the VAD algorithm. The table 1 shows the comparison of competitive networks VAD and others algorithms under 5 different noise backgrounds (Clean, Factory, Babble, Car and

White). The figure 3 shows the detect accuracy of different algorithms under different SNR environments, which are for Clean background to 0db.

Table .1 Algorithms Accuracy under Different Kinds of Environments

test environment	frames	G.729B VAD	BP networks VAD	new VAD （no hangover）	new VAD （with hangover）
Clean	1932	95.24%	97.43%	97.31%	97.77%
Factory	1932	90.78%	86.85%	91.04%	94.15%
Babble	1932	90.78%	91.77%	92.36%	94.46%
Car	1932	89.28%	91.20%	94.31%	96.99%
White	1932	70.22%	93.06%	91.51%	94.46%

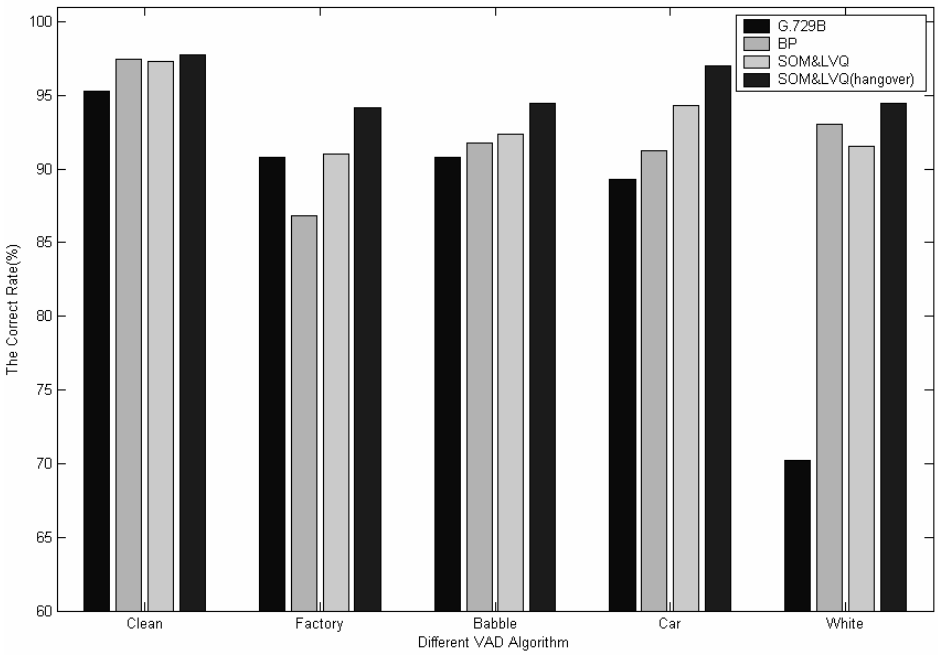

Fig.3. Algorithms Accuracy under Different SNR Environment

From the above analysis, we can see that the new VAD based on competitive networks has a better quality than the other algorithms and is especially more robust under noise environments. From the difference of the algorithm with the hangover protection and without that, we can see the hangover protection has a good effect to the algorithms and it make the accuracy increase by 2 per cents. So the new VAD algorithm in this paper is more accurate than G.729B VAD about 1 to 5 per cents.

What should be noticed is that when the noise environment is white noise, the detect accuracy of G.729B is only about 70 per cents. That means the G.729B VAD nearly classify all frames into speech frames. Because in our test database, the speech signal take up around 70 per cents and the noise signal take up nearly 30 per cents. And from the result of tests we can see that the G.729B VAD classify all frames into noise in faith. So this fact proves that the neural networks classify methods are more effective than the classical detection methods.

Then the contrastive results are showed in the figure 4. The upper one show a clean speech, the middle one is the same speech with 10db SNR and the lower is the detection result. The value 1 means the voice frame and the value 0 refers to the silence in the lower figure.

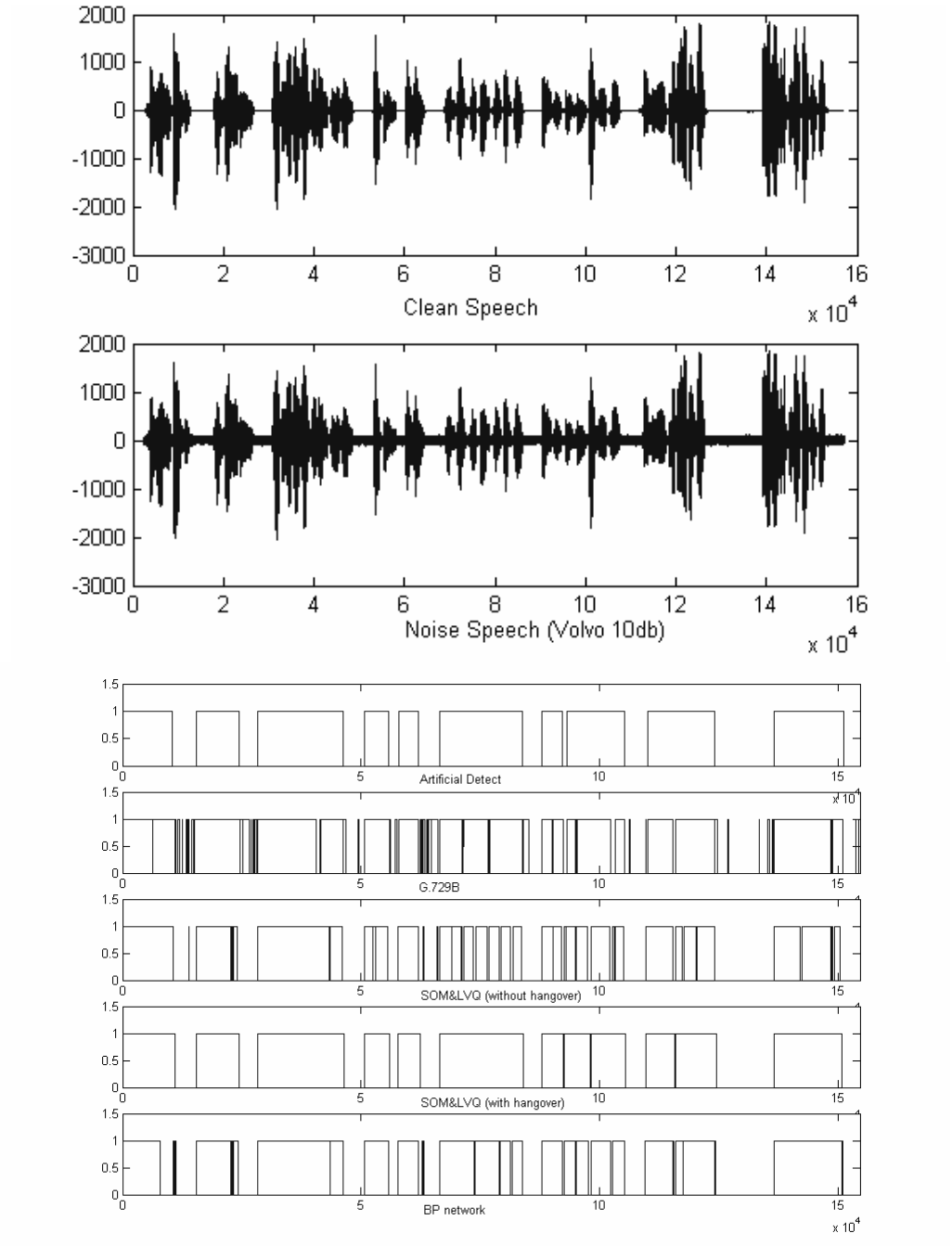

Fig.4. The speech (cleans and noise) and the flag result of different algorithms

From the above compare, we can see that the new algorithm is more close to the result of artificial detecting. Moreover the result of BP network is also better than the G.729B at a certain extent under the noise environment.

VI. CONCLUSIONS

From the network test results, we can see that the new VAD algorithm is effective in various noise environments. The combination of SOM and LVQ networks make the storage quantity and the calculate quantity decreased sharply. The LVQ networks can classify any kind of input vectors, no matter they are linear or not and this point is better than other neural networks. When we use the LVQ networks, we should take care that the number of the neurons must be big enough so that there are enough neurons for each class.

References

[1] ITU-T Recommendation G.729 Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70, 1996

[2] 3GPPTS 26.094 v5.0.0 Adaptive Multi-Rate (AMR) Speech codec, Voice Activity Detector 2002

[3] Xingjun Yang, Junli Zheng: Artifical Neural Network and Blind Signal Processing, Tsinghua Press, 2003

[4] Martin T. Hagan, Howard B. Demuth: Neural Network Design, China Machine Press, 2002

Yan Liu got bachelor's degree and master's degree from Beijing University of Technology in 2003 and 2005, respectively. Now she is working in Cisco Systems Co., Ltd as an Associate Sales Representative.

Changchun BAO is a Professor in the School of Electronic Information and Control Engineering of Beijing University of Technology. His research interests include speech signal processing and coding.