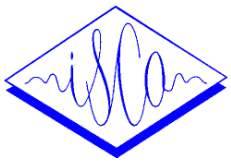

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A MULTI-STATE NN/HMM HYBRID METHOD FOR HIGH PERFORMANCE SPEECH RECOGNITION

Dong Yu, Taiyi Huang, Dao Wen Chen

National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
P.O. Box 2728, Beijing 100080, People's Republic of China

ABSTRACT

This paper proposed a novel Multi-State NN/HMM Hybrid Method named Multi-State Gaussion Competitive Neural Network (MSGCNN) for High Performance Speech Recognition. The basic idea of this new approach is integrating the Viterbi algorithm into a Gaussion Competitive Neural Network (GCNN). Comparing with Self-Aligning Network[3] it is more successful for it has both the Time alignment ability of HMM and the strong discrimination capability of Neural Network (NN). Moreover, because GCNN has better performance than basic Multilayer Perceptron (MLP), the novel system has many strong points such as faster training, more robust and noise immunity. An all Chinese syllable recognition system have been established based on MSGCNN and the comparative experiments confirmed the good characters stated above.

INTRODUCTION

Though Neural Network (NN) has been regarded as an efficient way for speech recognition because of its powerful discrimination ability, it cannot effectively deal with the problem of time variation caused by the variability of input speech signal. We have proposed several methods such as Self-Aligning NN[3] to enable NN the ability of time alignment in itself. These methods sure can better the results, they are not as good as expected. On the other hand, time alignment is a strong point of traditional methods such as Dynamic Programming (DP) and Hidden Markov Model (HMM). So an effective solution to the problem of time variation in a NN-based speech recognition system is to combine DP or HMM with NN. This paper describes a novel Multi-State NN/HMM hybrid system named Multi-State Gaussion Competitive Neural Network for high performance speech recognition. This new approach is better than other HMM/NN hybrid systems both on training time, recognition rate, robustness and noise immunity ability.

In the second part, we will introduce the architecture and the basic training algorithm of the novel HMM/NN hybrid method; In the third section, Gaussion Competitive Neural Network is described; In the forth part, we will introduce MSGCNN which is a integrated system combined the basic HMM/NN hybrid system and the GCNN. An all Chinese syllable recognition system is established and the comparative experiments is shown in the fifth part of this paper. At the last paragraph is the conclusion.

A NOVEL HMM/NN HYBRID METHOD

The architecture of this new HMM/NN hybrid system is shown in Fig. 1.

The basic idea of the new hybrid system is integrating the NN with Viterbi algorithm, the former is used to generate the score related to the emission probabilities and the later is used to find the best time alignment path. Unlike the classic MLP, a state layer is added into the network, its output value represents the score related to the emission probability of the corresponding state. In the state layer, utterances may have different number of states. Therefore, the whole system, or the whole network consists of four layers: an input layer, a hidden layer, a state layer and an output layer. The connections among input layer, hidden layer and state layer are the same as that in the conventional NN, and the activations for all hidden and state units are calculated by a feed-forward algorithm. When the windowed input utterance shifts frame by frame, a sequence of activation vectors is generated as the output of the state layer. Then the Viterbi Algorithm is performed to find out the optimal time alignment path through these states activations. The score of this optimal path represents the output score of the word. The training of such a network is accomplished by propagating the output error through the states in the optimal path down to the input layer. This network introduces time alignment power into NN as well as negative training into HMM. Furthermore, compared with MS-TDNN, it is more consistent with the articulation model of speech and so is a relatively better method for continuous speech recognition. Unlike other HMM/NN hybrid system witch establishes a network for each class, the new method integrating all classes in a single network and so will enhance the discrimination ability of NN.

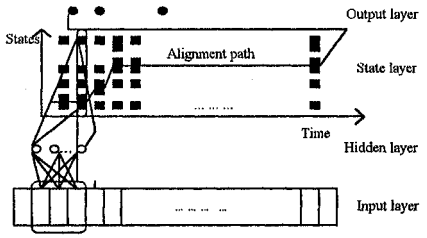

Fig. 1: Novel HMM/NN hybrid system

The basic training algorithm of this new model is as follows:

1. the input speech feature vectors are equally segmented into several states, and BP algorithm is used to find the initial weights;
2. Viterbi algorithm is performed on the speech feature vectors and a new optimal time alignment path is found;
3. BP algorithm is then used to revise the weights;
4. if error is larger than threshold value then go to 2;
5. end.

GAUSSION COMPETITIVE NEURAL NETWORK

Such a basic system described above is very time consuming, for, as all known, BP and Viterbi algorithm are both very slow when training. To make the training of such a hybrid system faster we introduce Gaussion function or Gaussion Competitive Neural Network into our system.

$$f(\bar{x}) = \exp\left(-w * \|\bar{x} - \bar{a}\|^2\right)$$

The Gaussion function is introduced because it has some good characters. It is a one peak function, and the descent speed of the function value can be adapted by changing the parameter w. So we can pre-select parameters of Gaussion function using other methods and then precisely modify them. Because gradient calculation and error Back propagation is no longer needed and the modifying is localized the training is much faster than classic BP algorithm.

The structure of GCNN is shown in Fig. 2.

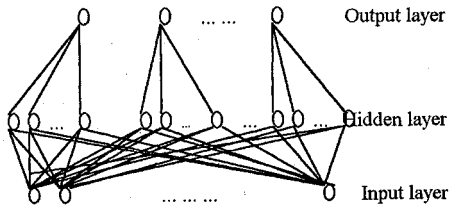

Fig. 2: Structure of GCNN

In this network, every output node is connected with a group of hidden nods and every hidden nods is connected with every input nods. The output value of GCNN is calculated in a competitive style. The maximum value of every hidden nod group is the representative of the corresponding group and the maximum value of the representative value of all groups is the output of the network.

Besides the strong point of fast training GCNN has also other good characters. First, it is a more robust system, because the output value of a hidden nod is changed slightly if the feature vector is enclosed in the domain of the peak point of the Gaussion function. Second, It is noise immnued. The noise data in the training data base will be automatically expelled by the training algorithm.

MULTI-STATE GAUSSION COMPETITIVE NEURAL NETWORK

Combining the novel HMM/NN hybrid system and GCNN rises a integrated network named Multi-State Gaussion Competitive Neural Network. The algorithms of this network is as follows:

Forward algorithm:

1. Calculate value of every output nod for each windowed part frame by frame.
2. Use Viterbi algorithm to find the optimal time alignment path and the corresponding score for each class.
3. The class whose overall score is the maximum is the recognition result.

Training algorithm is divided into two steps:

The first step:

1. Divide input speech evenly into several states and use VQ algorithm to obtain initial code book.
2. Calculate the output value of every window of input speech frame by frame. Then use Viterbi algorithm to find the optimal time alignment path and use VQ algorithm obtain new code book.
3. If error is larger than threshold value go to 2.

The second step:

1. Calculate the output value of every window of input speech frame by frame.
2. Use Viterbi algorithm to find the optimal time alignment path. If any windowed part of input speech is not classified correctly then modify parameters to reduce the number of errors.

$$dif = \ln o_j - \ln o_i$$

$$dis_i = \|\bar{x} - \bar{a}_i\|^2$$

$$dis_j = \|\bar{x} - \bar{a}_j\|^2$$

$$\bar{a}_i = \bar{a}_i - \eta_k * \alpha * (\bar{a}_i - \bar{x})$$

$$\bar{a}_j = \bar{a}_j + \eta_k * \alpha * (\bar{a}_j - \bar{x})$$

$$w_i = -(\ln o_i + dif * \eta_k) / dis_i$$

$$w_j = -(\ln o_j - dif * \eta_k) / dis_j$$

3. If the iteration is less than a pre-set value then go to 2

where o_i is the output value of corresponding output nod which the input speech belongs to. o_j is the output value of actual maximum output different with the correct state. η_k is the modifying step and $\lim_{k\to\infty} \eta_k = 0$.

It can be proved that such an algorithm is convergent.

CHINESE SYLLABLE RECOGNITION SYSTEM AND EXPERIMENT

Based on the MSGCNN an All Chinese Syllable Recognition System is established. The structure of this system is shown in Fig. 3. Vowels, consonants and transition segment of the input speech are recognized first and corresponding scores are obtained. Then the weighted sum of these three score is taken as the score of the syllable.

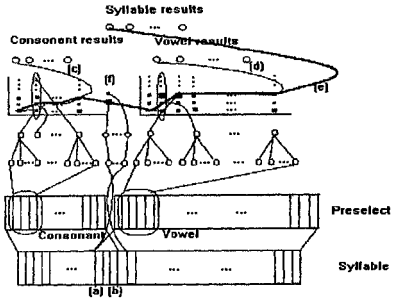

Fig. 3: Chinese Syllable Recognition System

(a): separate point

(b): transition segment

(c): optimal alignment path for consonant

(d): optimal alignment path for vowel

(e): optimal alignment path for syllable

(f): transition segment recognition

The performance of this system or MSGCNN is measured by a comparative experiment on speaker dependent all Chinese syllable database. We use 1024 syllables as training set and 405 syllables as test set. The results which is shown in table 1 approved our system.

Method	First Choice	Second Choice	Third Choice	Forth Choice	Fifth Choice
TDNN	84.20%	89.63%	93.09%	95.31%	96.05%
MSGCNN	92.35%	95.56%	97.04%	98.02%	98.52%

Table 1: Syllable recognition

CONCLUSION

This paper described a new HMM/NN hybrid system which is named Multi-State Gaussion Competitive Neural Network. The MSGCNN has many good points such as fast training, high recognition rate and strong robustness and is a high performance speech recognition system. Furthermore, continuous speech recognition and speaker independent recognition can be implemented with this MSGCNN.

REFERENCE:

[1]: Patrick Haffner, Michael Franzini, and Alex Waibel, "Integrating Time Alignment and Neural Networks for High Performance Continuous Speech Recognition", In Proc. ICASSP, IEEE, May 1991;

[2]: Patrick Haffner and Alex Waibel, "Multi-State Time Delay Neural Networks for Continuous Speech Recognition";

[3]: Dong Yu and Taiyi Huang, "A New Time-Alignment Approach for Robust Neural Network Based Speech Recognition", In Proc. ICSP, IEEE, 1993;

[4]: L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models", IEEE ASSP MAGAZINE, January 1986.