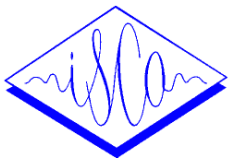

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

MANDARIN SYLLABLES RECOGNITION BY SUBSYLLABLES DYNAMIC NEURAL NETWORK

Dao Wen Chen, Xiao Dong Li, San Zhu, Dong Xin Xu & Tai Yi Huang

National Lab of Pattern Recognition
Institute of Automation,Chinese Academy of Sciences
Beijing 100080,China

ABSTRACT

This paper studied Mandarin syllables recognition dynamic neural network called state detector neural network (SDNN), which had minimum number of state detectors. We described the speech signal with a sequence of states and each state statistically characterized a period of speech vector. In each state detector which constituted the dynamic neural network, input vector nonlinearly approximated to a desired vector that was based on a forcing learning algorithm that has the merit of less training epoch and higher convergence accuracy. We compare this model to three other neural network (NN) models including the model of NN/HMM-based warping on the same database. Experiments show that our method has higher recognition accuracy than the others.

1. INTRODUCTION

Up to now, speech recognition has been a hard problem because of the high variability in the acoustic signal. The neural network appears a very promising model for speech recognition and also signal approximation. However, research in this domain has so far mostly been on small scale and static pattern classification. It seems that the sequential nature and temporal variation of the speech signal remains difficult to handle in neural network. The problem is necessary to be concerned in applying neural network to the processing of large vocabulary and continuous speech.

In speech recognition, speech is dealt with as a time sequence of short time spectral feature vectors. Information in the speech signal is encoded in the time sequence of its short duration vectors. Typically, if a speech signal is represented by time and spectral patterns, for different utterance of the same word, any point in such pattern exhibits more variance in temporal than in spectral direction. In classical pattern recognition match technique, these time-axis distortion are taken into account by nonlinear time-alignment of test and reference pattern, such as the most efficient dynamic programming (DP) technique.

Ken-ichi Iso proposed a neural prediction model (NPM) that used a multi-layer perceptron (MLP) sequence as a separate nonlinear prediction for each class and combined this NPM with DP that normalized the temporal distortion of speech. Using NPM , Iso reported very good results on speaker-independent isolated Japanese digit and 5000 words recognition. This paper propose a new system called SDNN that has less number of states detectors and much fast of convergence than the old one.

2. REPRESENTING SPEECH FEATURE VECTORS BY STATES

A critical fact about neural network is that they are statistical associative models. A typical network model has a set of input patterns and a set of output patterns. The role of the network is to perform a function that associates each input pattern with an output pattern. A learning algorithm, such as back-propagation (BP), uses the statistical properties of a set of input/output pairs, called the learning set, to generalize,that is, generate outputs from novel inputs. In a neural network model, the history of the system determines the system's response to a new stimulus. In a generalization neural network,the output distribution probability of a new input is determinate. According to mean-square criterion, the output of MLP approaches MAP and the probability of output class are conditioned by input data. MLP with context information and additional feedback input units can be seen as it were the generalization of Markov model. Like any dynamic system , the output including the all history of the system is just the "instantaneous" activation of each class and there is no any segmentation information in it. So the each output probability of MLP or HMM only gives the local contribution The sequence processing , that the global probability measured from on total observation vectors ,must rely on the time-alignment method as DTW (DP) or Viterbi algorithm to divide the signal into segments. Speech feature vectors can be represented as a sequence of states by the neural network added DTW that well model and solve the speech spectral variation and temporal variation.

3. STATE DETECTOR NEURAL NETWORK MODEL (SDNN)

Speech signal can be described as an non-stationary Markov chain that states transition is time-dependent. The DTW divides the transition process of the speech signal in N segment, and the signal in each segment is represented as a state by state detector. A SDNN model that is constituted of N detectors is used as a nonlinear approximation to a particular class of speech. The speech feature vector sequence $a_1, a_2, \ldots, a_T$ appear on the input of a SDNN model, and this sequence is divided in N segments by the model. then the n-th segment will be statistically described by the n-th detector.

The best segmentation n(t) depends on the segments that the accumulated prediction error D is as:

$$D = \underset{n(t)}{Min} \sum_{t=1}^{T} \left\| \hat{a}_t(n(t)) - a_t \right\|^2$$

here $\| \cdot \|$ is the Euclidean distance measure, n(t) is the t-th speech signal that belongs to n(t)-th segment. $a_1, a_2, \ldots, a_T$ considered as the best segments and they satisfy the above distance measure.

The optimization principal has the character that it is nothing how the initial state and the initial decision to be made since then the decision must be optimized DP is just the optimal algorithm. DP transfers a N step decision process into a N single step decision process,i.e., into N sub-problem that separately make decision , then the computation is more simplified.

In SDNN model , the training and recognition procedure are handled with BP and DP algorithm according to the optimal principal of the best segmentation and minimum accumulated prediction error.

In Iso NPM model's training algorithm , the BP and DP iteration take turns ,i.e., following a DP is a once BP. This will always change the input vector on the input of the detector and makes convergence hard. If the number N of state detectors is larger or equal to one half of the number of training frames,then the training data in each segment will be relatively stationary. But the training data will be terribly variable in the case when the N is much less than one half of the number of frames, and this will make the system convergence more difficulty. The SDNN convergence is based on individual detector's convergence and DP is based on the convergence of these detectors. We propose a modified Iso algorithm called Forcing Convergence Algorithm (FCA). The FCA is described as : the SDNN learning procedure (see Fig. 1) is divided into multiple stationary learning period. In a stationary period , each detector learns with a segment of data samples by BP algorithm and trends to converges while global error decreasing along with the variant of global errors being small value. The training of the detector pauses when it converges to a given threshold value of the global error. The next DP executes segmentation once again after all detectors satisfy the requiring error value. Whole training procedure can be considered as final end if the segmentation is no longer changed within two successive DP, i.e., the input training data of each segment for each detector doesn't change over. The learning curves showed in Fig.2 for our new algorithm FCA comparing with the old one.

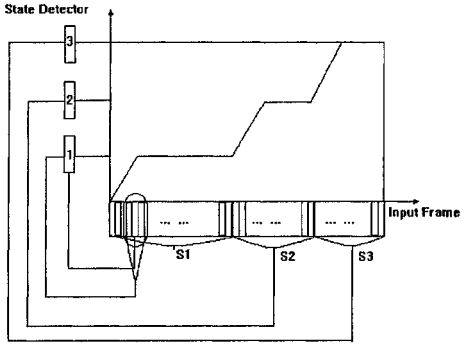

Fig.1 Configuration of SDNN model

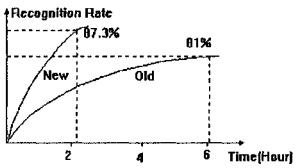

Fig.2 Comparison of two learning algorithm

4. EXPERIMENT RESULTS

4.1 Data base

1600 Chinese syllables for training data, other 400 syllables for test data.

4.2 Acoustic unit

According to the computation cost and complexity, memory consuming, training data amount and the stationary of acoustic unit in speech stream, we choose the initial subsyllable and final subsyllable as acoustic unit. Segmenting initial and final in a syllable by using zero crossing and energy ,and adding transition data between initial and final, we get 22 categories of initial

and 38 categories of final and the recognition accuracy of initial will improve for it bringing in transition information.

4.3 Number of detectors

We consider that more detectors can describe the details of speech signal then the SDNN model may be easy to training . But it will be very sensitive to the variant of same one word and in this case of existing large length differentiate the model converges hardly. We choose four detectors for initial syllable and five detectors for final. Table 1 show the average recognition accuracy of 22 initial subsyllable with three different number N of detector.

recognition candidate	1	2	3	4	5
N=3	87.3	96.2	98.2	98.5	99
N=4	88.9	97.7	99.0	99.2	99.7
N=5	88.4	97.2	99.0	99.0	99.7

Tab. 1 Average recognition accuracy of initial subsyllable with N

4.4 Recognition accuracy

Based on the above database and model parameter, the recognition accuracy of SDNN is as Table 2

recognition candidate	1	2	3	4	5
initial subsyllable	91.6	98.5	99.5	99.7	99.7
final subsyllable	98.5	99.7	100	100	100
whole syllable	90.5	98.0	98.7	99.0	99.2

Tab. 2 Average recognition accuracy of Mandarin syllable by SBNN model

4.5 Comparison with other dynamic neural network model

We studied other three models.

Method 1: Neural network with nonlinear segmentation

The method is to divide the sequence into N equal distortion segment and feed them onto input of N neural network correspondingly.

Method 2: Neural network with HMM-warping

According Viterbi algorithm, HMM divides the observation vector into N segment then normalize each segment to a fixed length number of frames. The N-MLP accept the data from N fixed length correspondingly.

Method 3: HMM

All three models use the same database of 38 final subsyllables which contain 8 vowels,14 compound vowels and 10 nasalized vowels. The vocabulary is very confusing.

The recognition result is in Table 3.

Method	Recognition accuracy (%)
SDNN	98.5
NN with HMM-based warping	91.2
NN with nonlinearsegmentation	75
HMM	95

Tab. 3 Comparison of four kinds of recognizer for final subsyllables

5. CONCLUSION

- The SDNN combine the neural network with DP warping can process temporal signal of speech efficiently. Chinese spoken language is constitute of 1400 single syllable ,so improving the initial and final subsyllables recognition accuracy is most important for Mandarin.
- From our experiment the performance of SDNN is comparative to HMM and much higher than NN/HMM.
- The SDNN is easy extending for large vocabulary and continuous speech recognition.
- The SDNN is a promising model and the higher recognition accuracy will be obtained by adding discriminate ability into the dynamic neural network.

ACKNOWLEDGE

The authors wish to thank the database support of Prof. Bo Xu & Mr. Bin Ma, also the valusble discussees of Mr. Dong Yu.

REFERENCE

(1) Readings in Speech Recognition, Edited by Alex Waibel & Kai-Fu Lee, Morgan Kaufmann Publishers,Inc., San Mateo California.

(2) Review of Research on Neural Nets for Speech, R.P.Lippmann, MIT, March 1989

(3) Links Between Markov Models & Multilayer Perceptrons, H.Bourlard, C.Wellekens, IEEE Transaction on P.A.&M.I. Vol.12,No.12 Dec.1990

(4) Large Vocabulary Speech Recognition Using Neural Prediction Model, Ken-ichi Iso & Takao Watanabe, ICASSP-91

(5) HMM-Based Warping in Neural Networks, Y.Q.Gao,T.Y.Huang & D.W. Chen, ICASSP-89