DYNAMICALLY LOCALIZED SELF-ORGANIZING MAP MODEL FOR SPEECH RECOGNITION

KyungMin NA, JaeYeol RHEEM and SouGuil ANN

Department of Electronics Engineering
Seoul National University
San 56-1, Shinlim-dong, Kwanak-gu
Seoul 151-742, KOREA
Telephone: +82-2-880-7279
Fax: +82-2-882-3906
E-mail: ctlab@krsnucc1.bitnet

ABSTRACT Dynamically localized self-organizing map model (DLSMM) is a new speech recognition model based on the well-known self-organizing map algorithm and dynamic programming technique. The DLSMM can efficiently normalize the temporal and spatial characteristics of speech signal at the same time. Especially, the proposed model can use contextual information of speech. As experimental results on ten Korean digits recognition task, the DLSMM with contextual information has shown higher recognition rate than predictive neural network models.

1. INTRODUCTION

Speech is one of the most convenient way of communication. Lots of studies on speech recognition such as DTW (Dynamic Time Warping) [1], HMM (Hidden Markov Models) [2], and ANN (Artificial Neural Networks) [3-7] have been carried out. Especially in recent years, a lot of neural network models have proven successful in various speech recognition tasks. TDNN (Time-Delay Neural Network) [3], SOFM (Self-Organizing Feature Map) [4], and PNNM (Predictive Neural Network Models) [5-7] are representative ANN models for speech recognition.

Among those ANN models, the PNNM is superior to other neural rivals in that 1) it can effectively normalize the time variability of speech, 2) it is easily extended to continuous speech recognition, and 3) it needs not to be entirely retrained when new word classes are added. The PNNM also shows high recognition performance for various recognition tasks. Motivated by such successes of the PNNM, we propose a new speech recognition model, DLSMM (Dynamically Localized Self-organizing Map Model) and its effective training algorithm.

The DLSMM is based on dynamic programming technique coupled with localized self-organizing maps. Temporal and spatial distortions of speech are efficiently normalized by dynamic programming technique and localized self-organizing maps, respectively. The structure of the model is seemed to be same with that of the PNNM. So, the DLSMM also has all the above-mentioned merits that the PNNM has. But, the model uses an LSM (Localized Self-organizing Map) sequence as a separate template for each word class while the PNNM uses an MLP (Multi-Layer Perceptron) predictor sequence as a separate nonlinear predictor for each word class. Additionally, the DLSMM has the advantage of having smaller connections than the PNNM with a little higher recognition rate.

Basic operations of the DLSMM are seemed to be same with those of conventional DTW-based speech recognizers. But the model is different from those DTW-based speech recognizers in that 1) single reference template (DLSMM) is sufficient to realize speaker-independent speech recognizer, 2) relatively small computation is required for dynamic programming, 3) both time variability and nonlinearity of speech are efficiently normalized, and 4) additional procedure for obtaining reference templates is not required. The proposed model can use contextual information, which are important in speech recognition tasks. The experimental results have shown that the DLSMM with contextual information outperforms the DLSMM without them.

This paper is organized as follows. In section 2, the dynamically localized self-organizing map model with its efficient training and recognition algorithm is described. Experimental results on the isolated Korean digits recognition are presented in section 3. Finally, conclusions are drawn in section 4.

2. DYNAMICALLY LOCALIZED SELF-ORGANIZING MAP MODEL

2.1 Model Description

The DLSMM can be regarded as an ordered sequence of localized self-organizing maps. Fig. 1 shows a graphical representation for the DLSMM for word w. Numbered circles represent corresponding LSM's, and N_w is the total number of LSM. As you feel, the whole structure looks like the conventional left-to-right HMM and PNNM. In the DLSMM, however, transitions between the maps are determined by DP technique associated with each LSM's output.

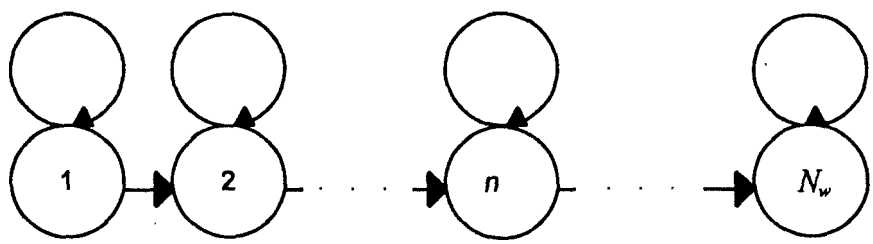

Fig. 1. The DLSMM for word w.

An LSM, a basic unit for the DLSMM, is created by Kohonen's self-organizing map algorithm [4]. The SOM (Self-Organizing Map) has the special property of effectively creating spatially organized internal representations of various features of input signals. The locations of the cells tend to be ordered as if some meaningful physical coordinate system were created over the map. Fundamentally, the SOM algorithm is based on the competitive learning and lateral interactions among the cells in the on-center off-surround manner. The whole procedure is as follows. The best-matching cell (the one whose weight vector most closely matches the input vector) is selected as a winning cell. All cells in the neighborhood that receive positive feedback from the winning cell participate in the learning process. As the learning proceeds, the size of the neighborhood is diminished until it encompasses only a single cell.

Let m_i be the weight vector of the i-th cell and x be the input vector. As with other competitive structures, a winning cell is determined for each input vector based on the similarity between the weight vectors and the input vector. In this paper, the Euclidean distance is used. The winning cell can be determined by

$$\|x - m_c\| = \min_i \{\|x - m_i\|\}. \tag{1}$$

Instead of updating the weight vector of the winning cell only, all cells within the defined neighborhood participate in the weight-update process. If c is the winning cell, and N_c is the list of cell indices that make up the neighborhood, the weight-update equations are given by

$$m_i(t+1) = \begin{cases} m_i(t) + \alpha(t)\big(x - m_i(t)\big) & \text{if } i \in N_c, \\ m_i(t) & \text{otherwise.} \end{cases} \tag{2}$$

The learning factor $\alpha(t)$ is written as a function of time to be reduced as the learning progresses.

The LSM has the same property and training strategy with the above. Apparently, it has the same structure with Kohonen's self-organizing map. Strictly speaking, however, the localized self-organizing map is different from Kohonen's original self-organizing phonetic map. Each localized self-organizing map is formed out of its corresponding local speech segments obtained by dynamic programming technique while Kohonen's original phonetic map is formed out of its global speech data. Fig. 2 shows a typical LSM.

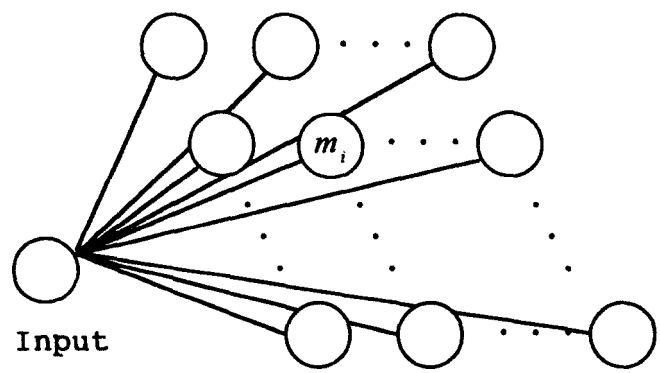

Fig. 2. Localized self-organizing map.

2.2 Recognition Algorithm

Kohonen's original map was used as a kind of vector quantization. His phonetic map emitted an index of the best-matching cell. But each trained LSM emits the smallest distance among the distances between the cells in it and the input feature vector. After all the input feature vectors are applied, a distortion matrix is formed. Input speech signal is optimally divided into N_w local segments by dynamic programming technique over the distortion matrix, and the n-th LSM emits the smallest distances for the n-th local segments. Fig. 3 illustrates a plane visualizing the dynamic programming computation. The horizontal and vertical axes represent the time direction for input vector s_t and the LSM sequence for word w, respectively.

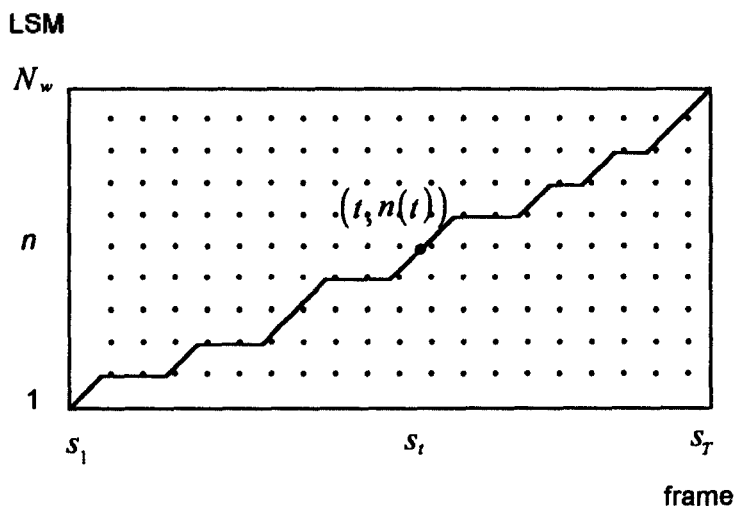

Fig. 3. Distortion minimization by DP computation.

The optimal segmentation of input vector sequence is performed by minimizing the accumulated distortion

$$D(w) = \min_{n(t)} \sum_{t=1}^{T} \min_{l} \| s_t - m^{w}_{n(t),l} \|, \tag{3}$$

where $\|\cdot\|$ is an Euclidean norm of a vector, and $m^{w}_{n(t),l}$ represents the l-th weight vector among the weight vectors of the n-th LSM for word w. Which LSM is assigned to the input vector at frame t is determined by $n(t)$.

The accumulated distortion $D(w)$ can be considered as the global distance between the input vector sequence and the DLSMM for word w. Consequently, the DLSMM that scores the smallest accumulated distortion should be chosen as the recognition result.

2.3 Training Algorithm

An effective training algorithm for the proposed model is presented below. It is easily applicable to continuous speech recognition tasks. By combining dynamic programming technique and self-organizing map algorithm, the DLSMM can be optimally trained. The algorithm is given as follows:

step 1. Initialize all the weights of each LSM.
step 2. Repeat the following steps for all training data until some conditions are satisfied.
step 3. Apply an input feature vetor sequence.
step 4. Create a distortion matrix by the outputs of each LSM.
step 5. Compute the accumulated distortion $D(w)$ by DP technique, and find its optimal trajectory $\{n^*(t)\}$ by the backtracking.
step 6. Update the weights for each LSM by the following formula along the optimal path $\{n^*(t)\}$.

$$m^w_{n(t),l}(t+1) = \begin{cases} m^w_{n(t),l}(t) + \alpha(t)\left(x - m^w_{n(t),l}(t)\right) & \text{if } l \in N_c, \\ m^w_{n(t),l}(t) & \text{otherwise.} \end{cases} \quad (4)$$

step 7. Reduce the size of the neighborhood if the predefined conditions are satisfied and the size of the neighborhood is not zero.

According to the above procedure, the optimal weights of each LSM are created without any teaching signal, which is a main difference between the predictive neural network models and the proposed model. Instead of the use of the nonlinear predictors, the localized self-organizing maps are adopted in the DLSMM. Each LSM can cluster corresponding speech segments in self-organizing manner.

3. EXPERIMENTAL RESULTS

The isolated Korean digits recognition experiments have been carried out in order to show the validity of the proposed recognition model. Speech data is uttered by 20 male speakers, and each speaker uttered each digit word once. The speech data was sampled at 10 kHz sampling rate, and analyzed with 25.6 ms Hamming window and preemphasis. The 16-order LPC analysis was performed, and the 12 cepstral coefficients without the 0-th order coefficient were derived and used as input feature vectors. The utterance data were divided into two sets. Only 80 data from the once utterances of 8 speakers were used for training, and the other 120 data were used for recognition test.

The DLSMM, composed of N_w LSM's, was prepared for each digit word w. The total number of LSM, N_w was determined as 12. Every LSM has 4 cells. An initial α was 0.03, and was decreased linearly. Total iteration was 300. The DLSMM without contextual information and the DLSMM with them were prepared. For contextual information, three consecutive speech vectors were used as an input vector. The neural prediction model [5] was adopted for comparison. An initial learning

coefficient was 0.01, and total number of predictor is 12. Total iterations were 300, and a second-order predictor was used.
The DLSMM with contextual information has scored 96.7 % in recognition rate while the DLSMM without contextual inflammations and neural prediction model have scored 93.3 %. So, the DLSMM with contextual information has outperformed the other two models.

4. CONCLUSION

In this paper, we proposed dynamically localized self-organizing map model for speech recognition, and showed its effectiveness by experiments. The proposed model is based on Kohonen's self-organizing map algorithm and dynamic programming technique. Especially, experimental results have shown that contextual information is important for speech recognition because speech is time-varying by nature.

The DLSMM based on subword units such as demi-syllable, phoneme, and triphone will be studied in the future. Using the conventional dynamic programming techniques such as one-stage dynamic programming and level-building dynamic programming, the DLSMM is applicable to the region of connected word recognition.

Additionally, discriminative training algorithms will be developed. Nonuniform weighting on dynamic programming plane can be considered. A generalized probabilistic descent method and other techniques will be used for obtaining the weighting function.

REFERENCES

1. P. C. Chang and B. H. Juang, "Discriminative training of dynamic programming based speech recognizers," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 2, pp. 135-143, April 1993.

2. W. Chou, B. H. Juang and C. H. Lee, "Segmental GPD training of HMM based speech recognizer," *Proc. ICASSP-92*, pp.473-476, 1992.

3. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, pp. 328-339, 1989.

4. T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-1479, September 1990.

5. K. Iso and T. Watanabe, "Speaker-independent word recognition using a neural prediction model," *Proc. ICASSP-90*, pp. 441-444, 1990.

6. J. Tebelskis and A. Waibel, "Large vocabulary recognition using linked predictive networks," *Proc. ICASSP-90*, pp.437-440, 1990.

7. E. Levin, "Hidden control neural architecture modeling of nonlinear time varying systems and its applications," *IEEE Trans. Neural Networks*, vol. 4, no. 1, January 1993.