A New Speech Recognition Model : Dynamically Localized Self-organizing Map Model

새로운 음성 인식 모델 : 동적 국부 자기 조직 지도 모델

Kyung Min Na*, Jae Yeol Rheem*, Souguil Ann*

나 경 민*, 임 재 열*, 안 수 길*

ABSTRACT

A new speech recognition model, DLSMM(Dynamically Localized Self-organizing Map Model) and its effective training algorithm are proposed in this paper. In DLSMM, temporal and spatial distortions of speech are efficiently normalized by dynamic programming technique and localized self-organizing maps, respectively. Experiments on Korean digits recognition have been carried out. DLSMM has smaller Experiments on Korean digits recognition have been carried out. DLSMM has smaller connections than predictive neural network models, but it has scored a little high recognition rate.

요 약

이 논문에서는 새로운 음성 인식 모델인 동적 국부 자기 조직 지도 모델과 그 학습 알고리즘을 제안한다. 동적 국부 자기 조직 지도 모델은 음성의 시간적, 공간적 왜곡을 동적 프로그래밍 기법과 국부 자기 조직 지도로 각각 정규화 시킨다. 한국어 숫자음에 대한 실험 결과로 제안하는 모델이 예측 신경회로망 모델보다 적은 수의 연결을 갖고서도 약간 높은 인식률을 보여 효과적임을 알 수 있었다.

I. Introduction

Speech recognition technique is essential to man-machine interface. A lot of studies on speech recognition such as DTW(Dynamic Time Warping) [1], HMM(Hidden Markov Model) [2], and ANN (Artificial Neural Network) [3-7] have been continued. Especially in recent years, various ANN's have proven successful in speech recognition tasks. TDNN(Time-Delay Neural Network) [3], SOFM (Self-organizing Feature Map) [4], and PNNM (Predictive Neural Network Model) [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 efficiently normalize the time variability of speech, 2) it is easily extended to continuous speech recognition, and 3) it needs not to be entirely retrained in case new word classes are added. With those merits, PNNM also shows high recognition performance. Motivated by such successes of PNNM, we propose a new speech recognition mo-

*서울대학교 전자공학과

접수일자 : 1994년 1월 4일

del, DLSMM(Dynamically Localized Self-organizing Map Model) and its effective training algorithm.

DLSMM is based on DP(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 and basic idea of the model are similar to those of PNNM. So, DLSMM also has all the above-mentioned merits that PNNM has. But, the model uses and LSM(Localized Self-organizing Map) sequence as a separate template for each word class while PNNM uses and MLP(Multi-layer Perceptron) sequence as a separate nonlinear predictor for each word class. Additionally, DLSMM has the advantage of having smaller connections than PNNM with a little higher recognition rate. Basic operations of the DLSMM are similar to 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 in speaker-independent case and 3) additional procedure for obtaining reference templates is not required.

This paper is organized as follows. In section II, 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 III. Finally, conclusions are drawn in section IV.

Ⅱ. Dynamically Localized Self-organizing Map Model

2.1 Model Description

DLSMM can be regarded as an ordered sequence of localized self-organizing maps. Fig. 1 shows a graphical representation for DLSMM for a 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 DLSMM, however, transitions between LSM's are determined by DP technique associated with the outputs of each LSM.

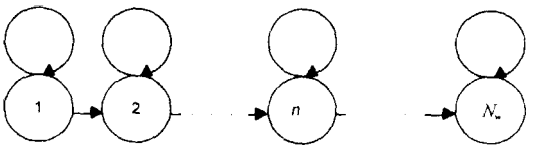

Fig. 1. DLSMM for word w.

An LSM, a basic unit for DLSMM, is created by Kohonen's self-organizing map algorithm. 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 competitive learning and lateral interactions among the cells in the on-center off-surround manner. First, the best-matching cell(the one whose weight vector most closely matchs 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 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. The winning cell can be determined by

$$\|x - m_c\| = \min_i \{\|x - m_i\|\}, \quad (1)$$

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

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

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

LSM has the same property and training strategy. Apparently, it has the same structure with Kohonen's self-organizing map. Strictly, 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 DP technique while Kohonen's original phonetic map is formed out of its global speech data. Fig. 2 shows a typical LSM.

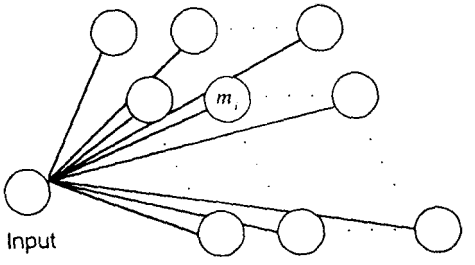

Fig. 2. Localized self organizing map.

2.2 Recognition Algorithm

Kohonen's original map was used as a kind of vector quantizations. 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. After all input vectors are applied, a distortion matrix is achieved. Input speech is optimally divided into N_w local segments by DP 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 DP computation. The horizontal and vertical axes represent the time variables for input vector s_t and the LSM sequence for word w, respectively.

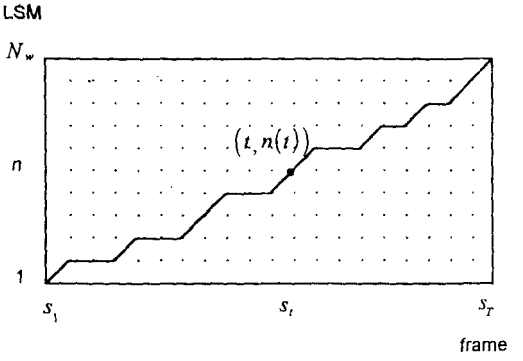

Fig. 3. Distortion minimization by DP computation.

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

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

where $\|\cdot\|$ is an Euclidian norm of a vector, and $m^w_{n(t),l}$ represents the l-th weight vector among the weight vectors of the n-the LSM for word w. Which LSM is assigned to the input vector at frame t is determined by $n(t)$. Actually, DP technique is applied under the following constraints :

$$n(1) = 1, \quad (4)$$

$$n(T) = N_w, \quad (5)$$

$$n(t) = n(t-1) \text{ or } n(t-1)+1, \ (1 < t \le T). \quad (6)$$

The sequence $\{n(1), \cdots, n(T)\}$ represents the optimal trajectory on the DP plane. Let the local distance measure $d_w(t, n)$ be $\min_l \| s_t - m^w_{n(t),l} \|$. Then the DP recursion formula is given by

$$g_w(t, n) = d_w(t, n) + \min[g_w(t-1, n), g_w(t-1, n-1)], \quad (7)$$

$D(w) = g_w(T, N_w n)$ is achieved at the end of recursive application of Eq.(7). The input speech segmentation is obtained by the backtracking for the optimal trajectory.

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

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, DLSMM can be optimaly 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. Computer the accumulated distortion $D(w)$ by DP technique, and find its optimal trajectory $(t, n(t))$ by the backtracking.

step 6. Update weights for each LSM by the following formula along the optimal path $(t, n(t))$.

$$m^w_{n(t),i}(q+1) = \begin{cases} m^w_{n(t),i}(q) + \alpha(q)(x - m^w_{n(t),i}(q)) & \text{if } t \in N_c, \\ m^w_{n(t),i}(q) & \text{otherwise.} \end{cases} \quad (8)$$

The index q indicates a training time.

step 7. Reduce the size of the neighborhood if some 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 predictive neural network models and the proposed model. Instead of the use of the nonlinear predictors, the localized self-organizing maps are adopted in DLSMM. Each LSM can cluster corresponding speech segments in self organizing manner.

III. Experiments

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 seven male speakers, and each speaker uttered each digit word three times. The speech data was sampled at 10 kHz sampling rate, and analyzed with 25.6 ms Hamming window and preemphasis. Then, 16-order LPC analysis was performed, and 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. For training, 50 data from the once utterances of 5 speakers were used, and the other data were used for recognition test.

DLSMM, composed of N_w LSM's, was prepared for each digit word w. The total number of LSM, N_w was determined as fifteen. Every LSM has 16 cells. An initial α was 0.3, and was decreased linearly. Among PNNM's, NPM(Neural Prediction Model) was adopted for comparison. An initial learning coefficient was 0.01, and the number of MLP predictor was the same with that of LSM. Total iteration was 500.

For the data(A) of the speaker who participated in training and those(B) of the speaker who did not participate in training, the recognition rates of PNNM have scored 97% and 88.3%, respect-

ively while those of DLSMM have scored 98% and 88.3%, respectively. So, DLSMM has scored a little high recognition rate with about 40% reduced connections.

Table 1. Recognition results.

Data	PNNM	DLSMM
A	97 %	98 %
B	88.3 %	88.3 %

Ⅳ. Conclusion

In this paper, we proposed a new speech recognition model, DLSMM, and showed its effectiveness by experiments. The proposed model has the similar structure to PNNM, and also has the similar property to DTW. As experimental results, DLSMM has scored a little high recognition rate than PNNM with approximately 40% reduced connections. Consequently, DLSMM is superior to PNNM in speech recognition.

DLSMM based on subword units like demi-syllable, phoneme, or triphone will be studied in the future. DLSMM is easily applicable to the region of connected word and continuous word recognition.

Additionally, discriminative training algorithms will be developed. Nonuniform weighting function on DP plane can be considered. That function can be achieved by the GPD(Generalized Probabilistic Descent) algorithm. Another approach is to derive new discriminative training formulas by the GPD algorithm coupled with MCEF(Minimum Classification Error Formulation). Such researches will be continued.

Reference

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, pp. 109-116, January 1993.
8. L. R. Rabiner, J. G. Wilpon and F. K. Soong, "High performance connected digit recognition using hidden Markov Models," *Proc. ICASSP-88*, pp. 119-122, 1988.

▲나 경 민

1968년 3월 5일생
1990년 2월 : 서울대학교 전자공학과 졸업
1994년 2월 : 서울대학교 대학원 전자공학과 석사과정 졸업(공학석사)
※주관심분야 : 신경회로망, 음성인식, 패턴분류

▲임 재 열
11권4호(92년 8월 31일) 참조

▲안 수 길
11권4호(92년 8월 31일) 참조