Korean Phoneme Recognition by Combining Self-Organizing Feature Map with K-means clustering algorithm

Yong-Ku Jeon, Seong-Kwon Lee, Jin-Woo Yang, Hyung-Jun Lee,
Soon-Hyob Kim

Department of Computer Engineering
Kwang-Woon University
Wolgye-dong, Nowongu
Seoul 139-701
Republic of Korea

ABSTRACT

It is known that SOFM has the property of effectively creating topographically the organized map of various features of input signals. SOFM can effectively be applied to the recognition of Korean phonemes. However, it isn't guaranteed that the network is sufficiently learned in SOFM algorithm. In order to solve this problem, we propose the learning algorithm combined with the conventional K-means clustering algorithm in fine-tuning stage. To evaluate the proposed algorithm, we performed speaker dependent recognition experiment using six phoneme classes. Comparing the performances of the Kohonen's algorithm with a proposed algorithm, we prove that the proposed algorithm is better than the conventional SOFM algorithm.

1. INTRODUCTION

The methods of speech recognition mainly have been based on the pattern matching. As the development of the analog and digital VLSI technology makes a parallel processing possible, the speech recognition technique that utilizes the Artificial Neural Network to duplicates a human brain has been raised. In the large vocabulary speech recognition system, the more vocabulary is increased, the more memory and computational time are increased. To get rid of these problems, we use a sub-word unit such as allophone, phoneme, diphone, and syllable.

In this paper, we apply a pattern matching method to implement a speech recognition system which based on phonemes. When the reference patterns are created by clustering, we use iterative K-means algorithm as the cluster which can solve the limitation of T. Kohonen's SOFM neural network.[7] We combine the SOFM with K-means algorithm to improv the performance of cluster.[3][4] The ultimate goal of an implemented phoneme recognition system is that the system plays a part of the labeler

to transform the speech signal into a quasi-phoneme as phonetic symbol.

2. SPEECH RECOGNITION USING NEURAL NETWORK

Neural network in speech recognition is considered as a black box having only input-output.[2] Neural network approach is easy to produce discriminating function required in pattern classification and to learn without examining all patterns. The neural network used in speech recognition is classified into static structure without feedback elements such as MLP, SOFM, and dynamic structure having backward elements as well as feedforward elements.[8]

3. SOFM NEURAL NETWORK WITH K-MEANS ALGORITHM

An information process of human brain is the unsupervised learning mechanism. SOFM neural network is identical to the conventional clustering method.[5][6] SOFM Neural Network plays a part of detecting a feature that represents a structure in input pattern by the competitive learning and lateral inhibition. The mechanism is forced to the response by localizing a input information into a two-dimensional Neural Network field. The SOFM Neural Network used in this paper has the connected structure of nodes that form a two-dimensional plane topology (Fig.1). In this paper, Euclidean Norm is used for a similarity calculation. If the squared root is removed from the Euclidean distance calculation, the $d^2(I,W)$ means the correlation between the input vector and the weight vector. It is represented by the Squared Error between the input vector and the weight vector. The Euclidean distance is used to modify the weight vectors in the learning.

$$\begin{aligned} d^2(I, W) &= \| I - W \|^2 \\ &= [\ \|I\|^2 + \|W\|^2 - 2\langle I, W\rangle\ ] \\ &= 2(1 - \cos\theta) \end{aligned}$$

The learning of SOFM Neural Network is interpreted as a process that minimizes a mean square error(Fig. 2). But actually decision criterion that terminates the learning process is a learning rate coefficient or the number of iterations.[10] This doesn't guarantee that the network is sufficiently learned. This case appears in Fig. 3. Because a W_2 doesn't be learned to the correct mean of class 2, a pattern X_2 is actually near to W_1, and is classified to the class 1 incorrectly. Therefore it is necessary that the self organized feature map in the rough-tuning is reserved and the weight vector is fine-tuned. In this paper, the local minima problem is tried to be solved by fine-tuning the weight vector using the K-means clustering. The sum of Squared Error of the feature map or the total variance is

defined by

$$V_{SSE} = \sum_{P=1}^{\dot{} } \sum_{I \in S_P} \| I - W_P \|^2$$

The weight vector is controlled by K-means method so that V_{SSE} can reach the local minimum point. K-means method is given as follows

step 1. Allocate the learning pattern set Sp to each node of the learned feature map, and choose the node that has a minimum Euclidean distance for the segmented Sp

step 2. Transform the weight of the selected node into the centroid of the set Sp,

$$W_P(i, j) = \frac{1}{N_P} \sum_{I \in S_P} I$$

here, i, j, is an index of node in M*M feature map,
and $1 \leq i, j \leq M$

step 3. Repeat the step 1, 2 for $1 \leq p \leq k$.

The phoneme recognition system is implemented as Fig. 4.

4. THE RESULTS OF EXPERIMENT AND DISCUSSION

In this paper, Recognition experiment was performed for speaker dependent in six phoneme classes, such as vowels, plosives, fricatives, liquids, nasals,final consonants. In extracting a feature vector for learning and recognition, we constructed 1335 C^iVC^f database to minimize a coarticulation and hand- segmented using the information of duration and valid section for each phoneme. We used the 12 order LPC cepstrum coefficients and normalized them to the value between -1 and 1. The recognition rate is 95.9% for total vowel , 82.6% for consonant, 87.2% for total phoneme classes.(Table 1)

5. CONCLUSION

First, the proposed Neural Network effectively represents good performance in spite of piecewise linear discriminate function. Second, it is necessary to select a proper feature for the Korean speech recognition. Third, it is necessary to examine clearly the feature pattern of the Korean phoneme. Fourth, it is difficult to get a 100% recognition rates in signal level because phoneme classes are overlapped in feature space. Finally, the proposed Neural Network is the network performing autoassociation.

REFERENCE

[1]. R.schlkoff, Pattern recognition, John Wiley & Sons, 1992.

[2]. B.W. wah, "Special Issue on Artificial Neural Networks Guest Editor's Introduction", IEEE Trans. on Computer, Vol. 40, No. 12, Dec. 1991.

[3]. H. SPATH, Cluster Analysis Algorithm, Ellis Horwood, Limited, 1980.

[4]. C. von der Malsberg, "Self-organizing of orientation sensitive calls in the striate cortex", Kybernetik, 14, pp. 85-100, 1973.

[5]. J. Kangs, T.Kohonen, etal., "Variants of Self-Organizing Maps", IJCNN, Vol. 2, pp. 517-522, 1989.

[6]. P. Brauer, "Infrastructure in Kohonen Maps", ICASSP, Vol. 1, pp. 647-650, 1989.

[7]. T. Kohonen, "Self-organized formation of topologically correct feature maps", Biological Cybernetics, 1982.

[8]. Z. Huang, A. Kuh, "A Combined Self-organizing Feature Map and Multilayer Perceptron for Isolated Word Recognition", IEEE Trans. on Signal Processing, Vol. 40, No. 11, Nov. 1992.

[9]. P. Antognetti, V. Milutinovic(Ed.), Neural Networks - Concepts, Applications, and Implementations, Vol. 1, Chap. 3, Chap. 4, Presentice-Hall Advanced Reference Series, 1991.

[10]. E. McDermott, S. Katagiri, "LVQ-based Shift-tolerant Phoneme Recognition", IEEE Trans. on Signal Processing, Vol. 39, No. 6, June 1991.

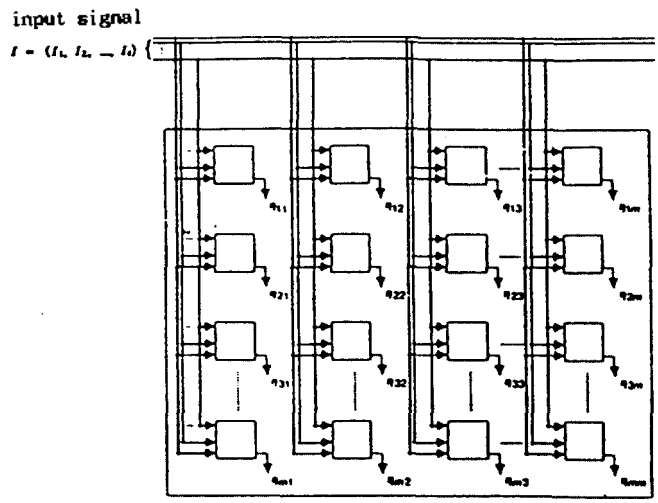

Fig 1. Structure of SOFM Neural Network

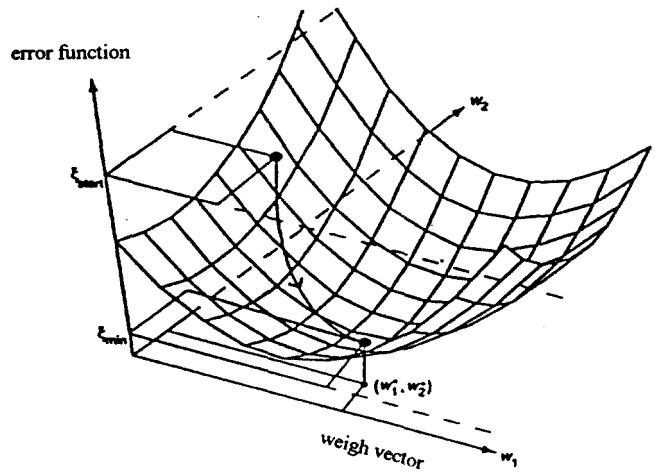

Fig 2. Adjustment of weight vector by Gradient-descent rule

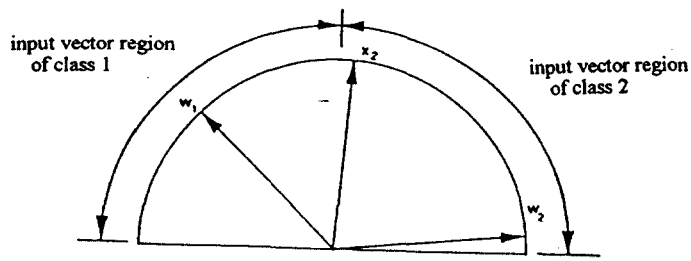

Fig 3. Example of misclassification

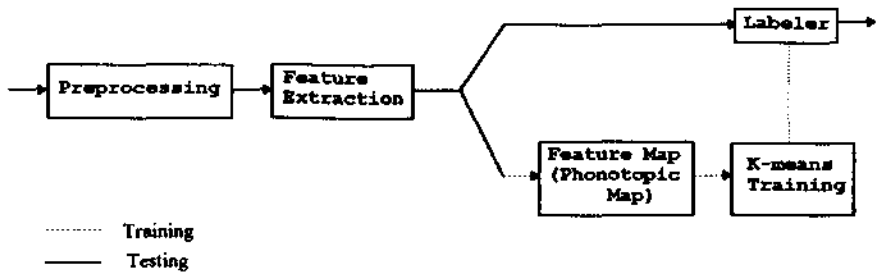

Fig 4. Block diagram of proposed phoneme recognition system

Phoneme	Token		Error	Recognition Rate	
	Learing	Test			
Monopthong					
아	5	20	0	100%	100%
어	5	20	0	100%	
오	5	20	0	100%	
우	5	20	0	100%	
으	5	20	0	100%	
이	5	20	0	100%	
에	5	20	0	100%	
Dipthong					
야	5	20	1	95%	93%
여	5	20	1	95%	
요	5	20	5	75%	
유	5	20	0	100%	
예	5	20	0	100%	
와	5	20	2	90%	
워	5	20	1	95%	
의	5	20	4	80%	
웨	5	20	0	100%	
위	5	20	0	100%	
Total vowel				95.9%	

Phoneme	Token		Error	Recognition Rate	
	Learning	Test			
Plosive					
ㅂ	14	21	3	85.7%	82.7%
ㄷ	14	21	4	81%	
ㄱ	14	21	3	85.7%	
ㅃ	14	56	8	85.7%	
ㄸ	14	56	14	75%	
ㄲ	14	56	9	83.9%	
ㅁ	14	21	2	90.5%	
ㅃ	14	21	4	81%	
ㄲ	14	21	4	81%	

Liquids/nasals/affricatives					
ㅁ	14	21	5	76.2%	79.8%
ㄴ	14	21	6	71.4%	
ㄹ (l/r)	l/14	21	5	76.2%	
	r/14	21	1	95.2%	
ㅅ	14	21	4	81%	82.5%
ㅆ	14	21	5	76.2%	
ㅎ	14	21	2	90.5%	
ㅈ	14	21	2	90.5%	95.2%
ㅉ	14	21	1	95.2%	
ㅊ	14	21	0	100%	

Final consonants					
ㄱ	14	21	4	81%	78.6%
ㄷ	14	21	5	76.2%	
ㅂ	14	21	6	71.4%	
ㄴ	14	21	3	85.7%	
ㄹ	14	21	6	71.4%	
ㅁ	14	21	3	85.7%	
ㅇ	14	19	4	71.4%	

Total	Token		Error	Reconition Rate
	Learning	Test		
	407	989	127	87.2%

Table 1. The recognition result