Phoneme Classification using the Modified LVQ2 Algorithm

수정된 LVQ2 알고리즘을 이용한 음소분류

Hong-Kook Kim,* Hwang-Soo Lee*
김 홍 국,* 이 황 수*

ABSTRACT

In order to construct a feature map-based phoneme classification system for speech recognition, two procedures are usually required. One is clustering and the other is labeling. In this paper, we first present a phoneme classification method based on the Kohonen's feature map algorithm for clustering and LVQ2 for labeling. Then, to improve the performance of the phoneme classification system, we employ the modified LVQ2(MLVQ2) algorithm instead of LVQ2, which consists of four stages of learning : the selective learning(SL), LVQ2, the perturbed LVQ2, and LVQ2 again. In MLVQ2, improved performance results are obtained by perturbing the weight vectors of the netweork for further training.

In order to evaluate the performance of the proposed phoneme classification algorithm, we first construct six intra-class feature maps for six different phoneme classes by using LVQ2 and MLVQ2. From the phoneme classification tests using these six feature maps, we obtain recognition rates of 60.4% and 65.4% for the LVQ2-based feature maps and the MLVQ2-based feature maps, respectively.

요 약

패턴매칭 기법에 근거한 음성 인식 시스템은 크게 clustering 과정과 labeling 과정으로 구성된다. 본 논문에서는 Kohonen의 feature map 알고리즘과 LVQ2 알고리즘을 각각 clusterer와 labeler로 하는 음소인식 시스템을 구성한다. 구성된 인식시스템의 성능을 향상시키기 위해서 수정된 LVQ2 알고리즘(MLVQ2)을 제안한다. MLVQ2는 selective learning, LVQ2, perturbed LVQ2 그리고 기존의 LVQ2의 4단계 학습과정으로 구성된다.

제안된 음소 인식 알고리즘의 성능을 평가하기 위하여 LVQ2와 MLVQ2를 각각 사용하여 6가지의 한국어 음소군에 대한 feature map을 만든다. 음소인식 실험결과, LVQ2와 MLVQ2를 사용하는 경우 각각 60.4%와 65.4%의 인식률을 얻을 수 있었다.

1. INTRODUCTION

Neural networks, which have the property of massive parallelism and fast adaptation, are characterized by their topologies and internal data representations. An artificial neural network can be trained using two types of training procedures : unsupervised learning and supervised learning [1].

* 한국과학기술원 정보및통신공학과
Department of Information and Communication Engineering, KAIST
접수일자 : 1992. 12. 10.

Among many other algorithms being applicable to speech recognition, the self-organizing feature map algorithm developed by Kohonen showed high performance of separability for input feature vectors[2]. Moreover, in order to obtain improved classification results, he proposed LVQ and LVQ2, another version of LVQ, and reported that the LVQ2-based classifier is more powerful than the back-propagation(BP)-based classifiers[3][4].

In general, a phoneme classifier is composed of a clusterer and a labeler. For example, the VQ-based recognition system can employ the VQ codebooks for each phoneme classes as the clusterer and a distance metric or dynamic time warping(DTW) as the labeler. For the phoneme classification system based on the Kohonen's feature map algorithm, the Kohonen's feature map becomes the clusterer like a vector-quantizer and its neurons are labeled according to the majority votes for a number of different responses[5]. Then, the labeling is done by finding the minimum distance node from input feature vectors and using a majority rule. On the other hand, with some fine adjustments of the weight vectors of the feature map to improve recognition accuracy, the LVQ or LVQ2 can also be used as the labeler. In this case, the traditional K-means clustering algorithm can be used for clustering, too[3].

In this paper, we construct a phoneme classifier using the Kohonen's feature map algorithm to provide the LVQ2 for labeling with better initial states than the K-means clustering algorithm. Neurons of the feature map are labeled without majority voting by using the newly proposed selective learning(SL) algorithm. Then, in order to further improve the classification accuracy for the phoneme classification, we propose and use the modified LVQ2 algorithm(MLVQ2) instead of LVQ2.

II. PROPOSED PHONEME CLASSIFICATION SYSTEMS

A general structure for phoneme classification can be illustrated in Fig. 1. Input speech signal is preprocessed via preemphasis and low pass filtering, and then feature vectors are extracted. A clustering algorithm is applied to these feature vectors to make clusters of similar phoneme classes. These clusters are labeled by using a suitable distance metric and labeled training feature vectors. Most of the phoneme classification systems have the similar structure as shown in Fig. 1. In VQ-based systems, VQ codebooks are constructed for clustering and a distance measure is defined for the labeler. And in LVQ2-based systems, the K-means algorithm is usually used for the clusterer and LVQ2 for the labeler.

Block diagrams of the two proposed classification systems are given in Fig. 2. The first one is composed of the Kohonen's feature map for the clusterer and LVQ2 for the labeler(Fig. 2(a)). Using the Kohonen's feature map instead of the K means algorithm for the clusterer provides the LVQ2 with better initial states for labeling. In this system, after the LVQ2 training is terminated, we may find the room for further improvement. It is expected to improve the recognition accuracy through more training. Therefore, a mo-

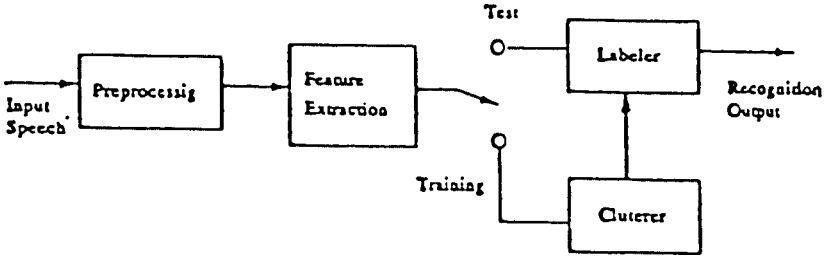

Fig. 1. A general structure of classification system.

dified LVQ2 is proposed for further training of the network and then we construct the MLVQ2-based phoneme classification system as shown in Fig. 2(b).

the feature map. However, the Kohonen's feature map algorithm and LVQ2 are operated on different training modes. That is, the latter operates on a supervised training mode and the former on

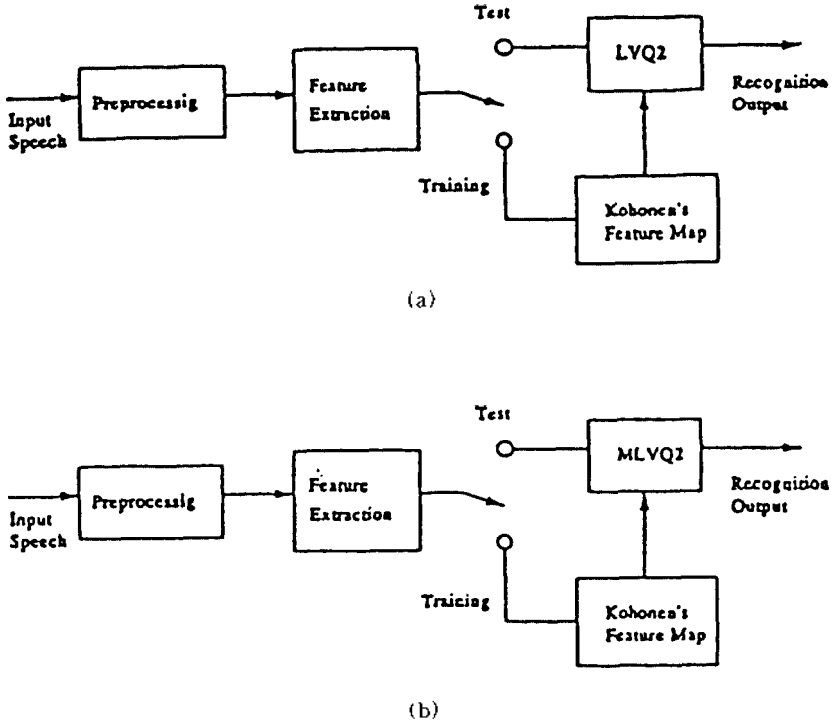

Fig. 2. Proposed phoneme classification systems.
(a)The Kohonen's feature map and LVQ2-based system.
(b)MLVQ2-based system.

Ⅲ. MODIFIED LVQ2 ALGORITHM

The modified LVQ2 algorithm is composed of four stages of learning as shown in Fig. 3. Firstly, training data obtained from the preprocessed input speech signal are clustered using the Kohonen's feature map algorithm employs an unsupervised learning procedure to make a feature map. Therefore, the constructed feature map can be used as a vector quantizer to classify input feature vectors. Our objective in this paper is to design a phoneme classifier by applying LVQ2 to an unsupervised one. Therefore, we must devise a method to link the two algorithms. This method becomes the first stage of MLVQ2, called the selective learning(SL) algorithm.

The SL algorithm transforms an unsupervised feature map into a supervised one. For each output node of the feature map, a proper phoneme label is assigned and the weights of that node become one of the reference vectors of the same phoneme. The SL algorithm is devised with the basis on the iterative adoption of the Bayes' rule as follows.

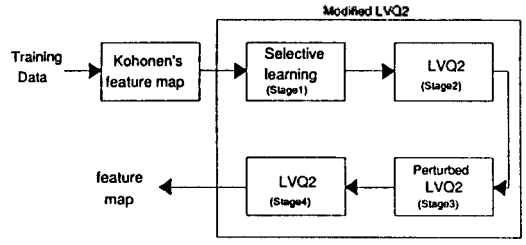

Fig. 3. Training sequence for feature map training.

For a given feature map, the i-th weight vector $m_i(i=1,\cdots,L)$ is assigned to a phoneme $p_k(k=1,\cdots,p)$ if it satisfies the equation

$$Pr(p_k \mid m_i) > Pr(p_l \mid m_i) \text{ for all } k \neq l \quad (1)$$

where $Pr(p_k \mid m_i)$ represents the conditional probability of p_k given m_i. From the Bayes' rule, we obtain

$$Pr(m_i \mid p_k)Pr(p_k) > Pr(m_i \mid p_l)Pr(p_l) \text{ for all } k \neq l \quad (2)$$

where $Pr(p_k)$ is the probability of a phoneme p_k and is obtained by

$$\frac{n(p_k)}{\sum_{k=1}^{p} n(p_k)} \quad (3)$$

where $n(p_k)$ is the number of training data for the phoneme p_k. The SL algorithm assigns a proper phoneme to each output node from (2) and selects the center node of that phoneme from

$$Pr(m_i \mid p_k) > Pr(m_l \mid p_k) \text{ for all } k \quad (4)$$

where i is the center node of phoneme p_k. And then for each iteration, phoneme assignment procedure is iteratively applied with a neighborhood $N(t)$ and a learning gain $\alpha(t)$.

For a given training vector x labeled p_k, if the nearest output node is the i-th node and the i-th node is inside the $N_i(t)$, the weight vector of the i-th node m_i is updated. Otherwise, m_i is not updated. This algorithm is summarized as follows :

1) Phoneme assignment.

$\text{Phone}(m_i(t)) = p_k \ (i=1,\cdots,L)$

iff $Pr(p_k \mid m_i(t)) > Pr(p_l \mid m_i(t))$ for all $l \neq k$.

2) Find the center node of each phoneme.

$\text{Center}(p_k) = i$ for each k

iff $Pr(m_i(t) \mid p_k) > Pr(m_r(t) \mid p_k)$.

3) Find the minimum distance node

$\min \| m_i(t) - x(t) \|$.

4) Update weights.

$m_i(t+1) = m_i(t) + \alpha(t)(x(t) - m_i(t))$ if $i \in N_i(t)$,

$m_i(t+1) = m_i(t)$, otherwise.

$t \leftarrow t+1$.

5) Test the terminating condition.

Terminate if $\frac{D(t+1)-D(t)}{D(t)} < \varepsilon$, or $t \geq T_{max}$.

Otherwise, go to step 1.

As the second stage of MLVQ2, the conventional LVQ2 is used for further training of the feature map. For a given labeled feature vector x, the training in LVQ2 occurs only when the nearest class has the incorrect phoneme label and the next-nearest class has the correct one.

When the LVQ2 training is converged, however, phoneme classification accuracy is not usually high enough. This is because the training is done only under the limited conditions. Therefore, the next stage of MLVQ2 is applied to further improve the classification accuracy. We call this stage of MLVQ2 algorithm the perturbed LVQ2 algorithm. The perturbed LVQ2 algorithm is described as follows. Assume an input vector x(t) is given, and C_i is the nearest class and C_j is the next-nearest class for x(t).

Case1 : $x \in C_j$, then

$m_j(t+1) = m_j(t) - \alpha(t)(x(t) - m_j(t))$,

$m_i(t+1) = m_i(t) + \alpha(t)(x(t) - m_i(t))$.

Case2 : $x \in C_i$, then

$m_i(t+1) = m_i(t) - \alpha(t)(x(t) - m_i(t))$,

$m_j(t+1) = m_j(t) - \alpha(t)(x(t) - m_j(t))$.

Case3 : Otherwise,

$m_k(t+1) = m_k(t)$.

After training the feature map by using this perturbed LVQ2 algorithm, the conventional LVQ2 training is applied again to the perturbed feature map. This is the final stage of MLVQ2.

In the next section, we show phoneme classification experiments to evaluate the performance of each stage of learning and to verify the usefulness of our MLVQ2.

Ⅳ. SIMULATION RESULTS AND DISCUSSION

Fig. 4 shows an overall block diagram of our phoneme classification system. Input speech signal is sampled at 10kHz. A 16-th order LPC analysis is performed on each of the 30ms Hamming windowed speech frames at the frame rate of 10ms. Then we extract the cepstral coefficients from the LPC coefficients as feature vectors[6].

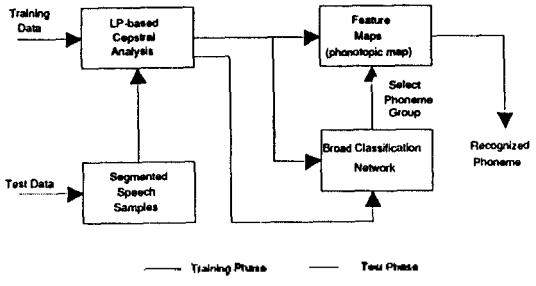

Fig. 4. Block diagram of the proposed phoneme recognition system.

For the first step of our phoneme classification experiment, we consider the task of classifying /b/, /d/, /g/. The tokens in the test and training set are manually segmented into phoneme units from phonetically balanced 100-word spoken by 3 male speakers. There are 220 and 110 tokens in the training set and the test set, respectively.

Using the training data, bdg-map is obtained according to the procedure in Fig. 3. Fig. 5(a) shows the bdg-map obtained using the Kohonen's feature map algorithm and Fig. 5(b) represents the feature map obtained by applying the SL algorithm. Comparing both of the feature maps, the map trained by the SL algorithm is seen to be rearranged such that output nodes of the same

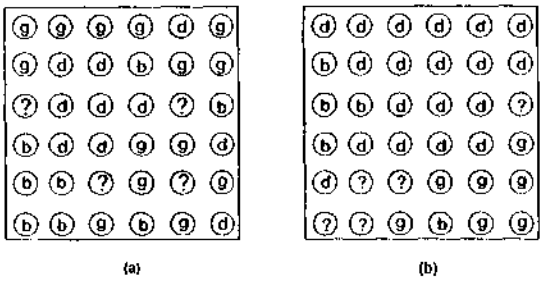

Fig. 5. The lax-maps before and after SL. (a)Before SL (b)After SL

class are located in the neighborhood. The number of updated input vectors for each iteration when applying the second stage of MLVQ2 is shown in Fig. 6(a). Trained iteratively by LVQ2, no more change occurs after 90 iterations, however, the recognition rate, shown in Table 1, is not high enough. For more training, the perturbed LVQ2 algorithm with the maximum iteration of 500 and a linear adaptation rule is ap-

Table 1. Recognition results of the bdg classification for each algorithm.

phoneme	NO.	SL (stage 1)	LVQ2 (stage 2)	Perturbed LVQ2 (stage 3)	LVQ2 (stage 4)
b	24	21(50.0)	17(70.83)	6(25.0)	20(83.33)
d	29	22(75.86)	19(65.52)	12(41.38)	19(65.52)
g	57	27(47.37)	45(78.95)	52(91.23)	47(82.46)
Total	110	61(55.45)	81(73.64)	70(63.64)	86(78.18)

plied to this map. Fig. 6(b) and 6(c) show that the number of updated input vectors after the perturbed LVQ2 is not zero and more training can still be done using LVQ2. Of course, the phoneme recognition result of the perturbed LVQ2 is lower than that of the fourth stage of MLVQ2 because the weights of the map in the third stage are perturbed for the next training. From Table I, we can obtain 4.5% higher recognition rate by using MLVQ2 than that using only LVQ2.

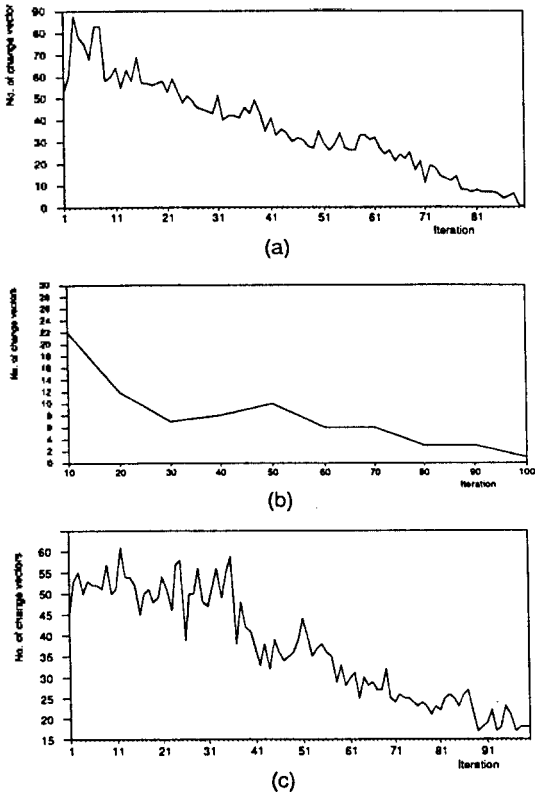

Fig. 6. Distribution of the number of updated input vectors for each iteration.
(a)LVQ2. (b)Modified LVQ2(step1).
(c)Modified LVQ2(step2).

For the next step of our experiment, we simulate the phoneme classification system covering all Korean phonemes. Six feature maps for six different phoneme classes are obtained using proposed MLVQ2. The performance of intra-class recognition is illustrated in the third and forth columns of Table II. We obtain the recognition rate of 60.4% and 65.4% for LVQ2-based and MLVQ2-based feature maps, respectively.

Table II. Korean phoneme recognition results for each algorithm.

phoneme	NO.	SL	only LVQ2	Modified LVQ2
b	24	12	17	20
d	29	22	19	19
g	57	27	45	47
lax	110	55.45%	73.64%	78.18%
pp	5	5	5	5
tt	6	1	1	4
kk	6	5	5	6
glottal	17	64.71%	64.71%	88.24%
p	6	4	6	5
t	5	3	2	4
k	4	2	3	4
asplrated	15	60.00%	73.33%	86.67%
m	21	12	10	14
n	48	10	36	35
l	39	28	36	36
ng	19	10	12	12
llquid-nasal	127	47.24%	74.02%	76.38%
s	11	2	1	1
ss	11	7	5	8
z	21	5	4	4
zz	12	7	6	12
ts	5	0	0	1
h	13	12	12	12
frlvcative	73	45.21%	38.36%	52.05%
I	24	10	9	9
e	9	1	2	2
∂e	15	5	9	8
wi	6	1	4	4
oi	1	1	1	1
i	12	8	8	9
∂	30	8	15	21
a	47	29	42	44
u	21	9	9	13
o	24	14	20	20
j	6	3	6	5
y∂	22	5	2	1
ya	5	2	2	0
yu	6	0	3	2
yo	5	2	1	1
we(w∂e)	11	8	7	5
w∂	5	0	0	0
wa	4	3	0	0
iy	9	0	0	1
vowel	262	41.60%	53.44%	55.73%
total		46.85%	60.43%	65.40%

V. CONCLUSIONS

We present two phoneme classification systems. The one is LVQ2-based system which combines the Kohonen's feature map for clustering and LVQ2 for labeling. The other is MLVQ2-based system in which MLVQ2 is proposed and used instead of LVQ2 to improve classification accuracy. From the results obtained from out computer simulation, the performance of MLVQ2-based system can be used in the speaker-independent continuous speech recognition system to obtain better performance. Further research should be directed to obtain higher recognition accuracy.

▲Hong Kook Kim
Ph.D. student, Department of Information and Communication Engineering KAIST
(11권 4호 참조)

▲Hwang-Soo Lee : Vol.11 No.4

REFERENCES

1. *DARPA Neural Network Study*, AFCEA International Press, Fairfax, VA, 1988.
2. T. Kohonen, "Self-organization and Associative Memory"(2nd Ed.), Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 1988.
3. T. Kohonen, G. Barna and R. Chrisley, "Statistical Pattern Recognition with Neural Network : Benchmarking Studies," Proc.of ICNN, Vol.1, pp.61-68, July, 1988.
4. E. McDermott and S. Katagiri, "Shift-Invariant, Multi-Category Phoneme Recognition using Kohonen's LVQ2," Proc. of ICASSP, pp.81-84, 1988.
5. T. Kohonen, "The Neural Phonetic Typewriter," Computer Mag. pp.11-22, Mar., 1988.
6. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.