KOREAN CONSONANT RECOGNITION USING A MODIFIED LVQ2 METHOD

Shozo Makino[†], Yoshiyuki Okimoto[†], Ken'iti Kido[††], Hoi Rin Kim[‡] and Yong Ju Lee[‡]

† Graduate School of Information Sciences, Tohoku University,
2-1-1 Katahira, Aoba-ku, Sendai, 980 Japan.
†† Chiba Institute of Technology, Tsudanuma, 275 Japan.
‡ Electronics and Telecommunications Research Institute,
P. O. BOX, Daeduk Science Town, Daejeon, Korea.

ABSTRACT This paper describes recognition results using the modified Learning Vector Quantization(MLVQ2) method which we proposed previously. At first, we investigated the property of duration of 29 Korean consonants and found that the variances of the duration were extremely big comparing to other languages. We carried out preliminary recognition experiments for three stop consonants P, T and K. From the recognition results, we defined the optimum conditions for the learning. Then we applied the MLVQ2 method to the recognition of Korean consonants. The training was carried out using the phoneme samples in the 611 word vocabulary uttered by 2 male speakers, where each of the speakers uttered two repetitions. The recognition experiment was carried out for the phoneme samples in two repetitions of the 611 word vocabulary uttered by another male speaker. The recognition scores for the twelve plosives were 68.2% for the test samples. The recofnition scores for the 29 Korean consonants were 64.8% for the test samples.

1 INTRODUCTION

The Learning Vector Quantization(LVQ,LVQ2) methods were proposed by Kohonen et al.[1]. They showed that the LVQ2 was superior to the LVQ. However, in the LVQ2 method, the given vector recognized as more than the second rank was not used for the learning. Then we proposed a modified training algorithm[2] to overcome this weakpoint. The modified Learning Vector Quantization(MLVQ2) showed better performance comparing to the LVQ2.

In this paper, we apply the MLVQ2 to 29 Korean consonants. At first we investigated the property of duration of 29 Korean consonants and the optimum learning condition. Then we carry out speaker-independent and multi-speaker-dependent phoneme recognition tasks.

2 PHONEME RECOGNITION USING A MODIFIED LVQ2 METHOD

In the LVQ2 algorithm proposed by T. Kohonen et al.[1], the given training vector x should satisfy the following conditions to modify the reference vector: 1) the nearest class to the given vector is incorrect, 2) the next-nearest class to the given vector is correct, and 3) the training vector falls inside a small, symmetric window defined around the midpoint of the incorrect reference vector and the correct reference vector.

In our preliminary phoneme recognition experiments using the LVQ2 algorithm, we found that the given training vector hardly contributed to the learning when the rank of the given vector was greater than 2. Then we proposed a modified training algorithm for the LVQ2 method[2]. The condition 2) in the LVQ2 algorithm is removed. The modified Learning Vector Quantization algorithm(MLVQ2) is composed of the following five steps:

(1) Reference vectors are chosen using the K-Means clustering method from each class.

(2) A training vector is picked up from the training data set. The nearest reference vector of each class to an input vector is selected.

(3) The rank of the correct class is computed. When the rank of the correct class is n, we assume that the reference vector of the correct class is $\boldsymbol{m}_n$. If n is equal to 1, the processing goes to (2), otherwise the processing proceeds to the next step.

(4) A check is made to see whether or not the input vector falls within a small window, where the window is defined around the midpoint of $\boldsymbol{m}_1$ and $\boldsymbol{m}_n$.

(5) The $n-1$-th and the n-th reference vectors are modified according to the following equation and then the processing proceeds to (2).

MLVQ2:

$$[\boldsymbol{m}_{n-1}]^{t+1} = [\boldsymbol{m}_{n-1} - \alpha(t)(\boldsymbol{x} - \boldsymbol{m}_{n-1})]^t \quad (1)$$

$$[\boldsymbol{m}_n]^{t+1} = [\boldsymbol{m}_n + \alpha(t)(\boldsymbol{x} - \boldsymbol{m}_n)]^t \quad (2)$$

where $\alpha(t)$ is a learning coefficient of function t.

3 PHONEME RECOGNITION SYSTEM

The speech was analyzed by a 29-channel band-pass filter bank. The speech was represented by a sequence of logarithmic spectra with 5-ms frame shift. The phoneme recognition system was similar to the shift-tolerant model proposed by McDermott et al.[3]:

(1) Eight mel-cepstrum coefficients and 8 Δ-mel-cepstrum coefficients were computed for every frame from the logarithmic spectrum. Each reference vector was represented by 240 coefficients (15 frames $\times$ 16 coefficients). Each class was assigned 15 reference vectors chosen by the K-Means clustering method.

(2) A 15-frame window was moved over the given phoneme segment and yields a 240-dimensional input vector every frame.

(3) In the training stage the MLVQ2 algorithm was applied to the input vector as previously described.

(4) In the recognition stage we computed the distance between the input vector of each frame and the nearest reference vector within each class.

(5) From these distances, each class was assigned an activation value a_w as defined by

$$a_w(c,t) = 1 - \frac{d(c,t)}{\sum_i d(i,t)} \tag{3}$$

where d, c and t are distance, class, and frame number, respectively.

(6) The activation value of each frame was summed over the given phoneme segment.

(7) The class with the maximum activation value was regarded as the recognized output.

4 SPEECH SAMPLES FOR KOREAN CONSONANT RECOGNITION

We used the Korean speech corpus made by ETRI. The corpus is composed of 3666 words uttered by three adult males(rjh, pck and hhi), where each of the speakers uttered two repetitions of 611 words. We deal with 29 Korean consonants shown in the table 1. As can be seen from the table 1, the variance of phoneme duration is extremely big comparing to Japanese.

Table 1: Korean consonants used in the recognition experiments

Kind of phoneme	Average duration	Minimum duration	Maximum duration	Kind of phoneme	Average duration	Minimum duration	Maximum duration
m	14.5	5	51	ms	23.0	5	56
n	12.8	1	36	ns	20.6	3	45
ng	19.5	2	67	b	7.8	2	17
bv	12.4	4	22	bs	4.8	1	12
d	8.7	3	49	dv	12.0	2	30
ds	4.8	2	9	g	11.7	3	23
gv	12.3	1	37	gs	4.2	2	9
j	15.8	8	24	jv	13.0	3	28
s	22.4	6	47	P	18.2	1	53
T	23.3	1	60	K	21.1	2	57
C	24.5	5	50	S	31.2	4	54
p	23.6	6	56	t	24.7	3	50
k	27.6	4	56	c	29.0	6	61
h	16.3	2	38	l	18.8	3	58
r	6.7	1	36				

ms, ns, bs, ds, gs: m, n, b, d, g located at syllable-fanal positin.
b, d, g, j: b, d, g, j located at word-initial position.
bv, dv, gv, jv: b, d, g, j located at word-medial position.
b, d, g, j, bv, dv, gv, jv, bs, ds, gs, s: lax
P, T, K, C, S: tense, unaspirated
p, t, k, c: tense, aspirated

Hereafter we carry out two kinds of recognition experiments: speaker-independent and multi-speaker-dependent. In speaker-independent experiments, the reference vectors were obtained from the speech data uttered by rjh and pck, and then were

applied to the speech data uttered by hhi. In multi-speaker-dependent experiments, the reference vectors were obatained from the first repetitions uttered by the three speakers, and then were tested for the second repetitions of the three speakers. "close" means the recognition scores for the training samples. "open" means those for the test samples.

5 PRELIMINARY RECOGNITION EXPERIMENTS

We carried out speaker-independent phoneme recognition experiments for investigating the learning conditions: number of training samples of each phoneme class per iteration(N_s) and number of reference vectors per phoneme class(N_r).

Figure 1 shows the recognition results for P, T and K. When number of iteration was set to 10, recognition scores reached a plateau at 4000 training samples. The number of reference vectors did not affect much on phoneme recognition scores. Hereafter, we use 4000 training samples of each phoneme class per iteration and assign 15 reference vectors to each phoneme class. At each iteration, for each phoneme class, 4000 samples are randomly picked up from the real phoneme samples included in the training samples. Therefore, for the training of three phonemes, 12000 phoneme samples are used at each iteration in the training stage.

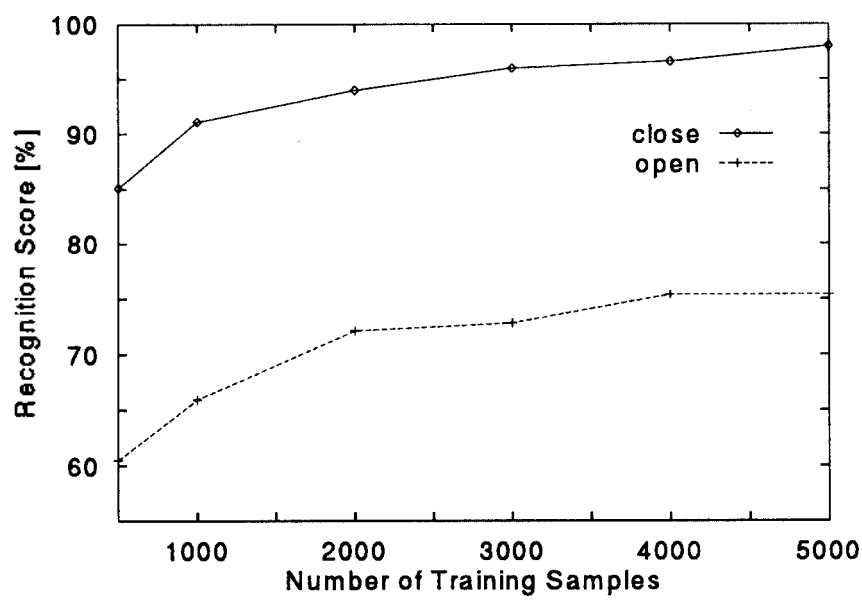

Figure 1: Effect of number of training samples

6 PHONEME RECOGNITION EXPERIMENTS FOR PLOSIVES

Figure 2 shows relation between number of iterations and recognition scores for twelve plosives. The multi-speaker-dependent and speaker-independent recognition scores are shown by "*.dependent" and "*.independent". The learning curve reached

Table 2: Speaker-independent recognition scores for plosives

Input Phonemes	Output phonemes											
	b	d	g	bv	dv	gv	P	T	K	p	t	k
b	61.8	5.3	2.6	2.6			6.6	2.6	3.9	11.8	2.6	
d	17.2	45.3	10.9		1.6		4.7	1.6	9.4	4.7	1.6	3.1
g	3.3	2.2	69.6							3.3		21.7
bv				74.5	5.9	17.6			2.0			
dv				7.4	83.6	5.0	1.6	1.6	0.8			
gv				9.5	8.3	82.1						
P	2.7			4.1	1.4	1.4	83.8	2.7	4.1			
T	1.0		3.0	2.0	1.0	2.0	20.0	49.0	22.0			
K	2.0		3.9		3.9	2.0	8.8	6.9	67.6		2.9	2.0
p	5.4	2.2	1.1		1.1		7.6	1.1		80.4		1.1
t		6.4	7.6				2.1	6.4	12.8	17.0	38.3	9.6
k			5.7						1.4	10.0	1.4	81.4

a plateau and the recognition scores for the test samples were realtively high. Multi-speaker-dependent and speaker-independent phoneme recognition scores were 76.6% and 68.2% for the open experiments. The multi-speaker-dependent recognition scores were about 8% higher than the speaker-independent recognition scores. Table 2 shows a confusion matrix between phoneme classes for the test samples. The phonemes d,T and t gave lower recognition scores. Most of the errors were ocurred between phonemes belonging to the same phoneme groups.

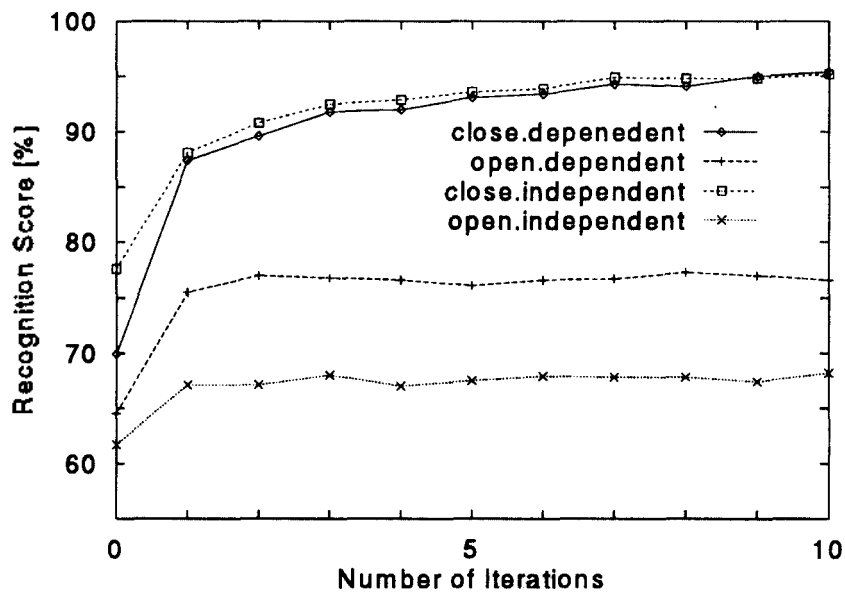

Figure 2: Recognition scores for plosives

Table 3: Recognition scores for 29 Korean consonants

Data set	Phonemes									
	m	ms	n	ns	ng	b	bv	bs	d	dv
Dependenet	64.5	66.5	65.8	54.7	77.7	73.1	76.8	64.9	65.6	84.3
Independent	59.3	30.5	49.0	29.9	70.0	59.2	70.6	54.5	46.9	83.6
Data set	Phonemes									
	ds	g	gv	gs	j	jv	s	P	T	K
Dependent	68.3	77.1	75.0	67.1	82.6	89.7	74.9	36.9	82.0	84.2
Independent	67.1	67.4	73.8	52.8	57.9	85.7	73.0	74.3	57.0	67.6
Data set	Phonemes									
	C	S	p	t	k	c	h	l	r	
Dependent	73.3	98.3	71.7	61.0	78.1	76.6	80.5	96.2	78.1	
Independent	79.4	82.7	77.2	35.1	67.1	72.8	84.9	87.4	68.6	

7 PHONEME RECOGNITION EXPERIMENTS FOR 29 KOREAN CONSONANTS

Table 3 shows recognition scores of 29 Korean consonants for the test samples. The multi-speaker-dependent and speaker-independent recognition scores were 74.1% and 64.8% for the test samples. The phonemes with low recognition scores was P in the multi-speaker-dependent recognition experiment. Many samples of the phoneme P were mis-recognized as T and K. In the speaker-independent recognition experiment, the phonemes with low recognition scores were ms, n, ns, d and t. The tendency of mis-recognition is as follows: ms → ng and ns, n → ng, r, m and h, ns → ng, l, h, n and ms, d → b, g and K, and t → p, k, T and K.

8 CONCLUSION

We applied the MLVQ2 to 29 Korean consonant recognition. The recognition scores for the twelve plosives were realtively good. However, the differences between the phoneme recognition scores for the training samples and those for the test samples are very large comparing to Japanese phoneme recognition. This implies that a large amount of speech data uttered by several hundreds of speakers will be necessary to develop a Korean phoneme recognition method.

References

[1] T. Kohonen, G. Barna and R. Chrisley, *Statistical Pattern Recognition with Neural Networks: Benchmarking Studies,* IEEE Proc. of ICNN, Vol. 1, pp. 61-68(July, 1988)

[2] S. Makino, M. Endo, T. Sone and K. Kido: *Recognition of phonemes in continuous speech using a modified LVQ2 method,* J. Acoust. Soc. Jpn., (E)13,6, pp. 351-360(1992-6)

[3] E. McDermott and S. Katagiri, *Shift-Invariant Phoneme Recognition Using Kohonen's Networks,* Proc. ASSP, pp. 81-84(1989)