東北大学機関リポジトリ
TOUR
Tohoku University Repository

Continuous speech recognition with modified learning vector quantization algorithm and two-level DP-matching

著者	牧野　正三
journal or publication title	IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92
volume	1992
number	1
page range	597-600
year	1992
URL	http://hdl.handle.net/10097/46641

doi: 10.1109/ICASSP.1992.225838

CONTINUOUS SPEECH RECOGNITION WITH MODIFIED LEARNING VECTOR QUANTIZATION ALGORITHM AND TWO-LEVEL DP-MATCHING

Shozo Makino, Mitsuru Endo, Toshio Sone and Ken'iti Kido

Research Center for Applied Information Sciences, Tohoku University, Sendai, 980 Japan.

ABSTRACT

This paper proposes a new phoneme recognition method based on the Learning Vector Quantization(LVQ2) algorithm. We propose three kinds of modified training algorithms for the LVQ2 algorithm. In the recognition stage, the likelihood matrix is computed using the reference vectors and then the optimum phoneme sequence is computed from the matrix using the 2-level DP-matching with duration constraints. The recognition score of phonemes in isolated spoken words was 89.1% for the test set. The phoneme recognition scores obtained by the modified LVQ2 algorithms were higher than that obtained by the original LVQ2 algorithm. We applied this method to a multi-speaker-dependent phoneme recognition task for continuous speech uttered Bunsetsu-by-Bunsetsu. The phoneme recognition score was 85.5% for the test speech samples in continuous speech.

1.0 INTRODUCTION

The neural network approach is one of the very promising approaches to a phoneme recognition task [1,2]. The Learning Vector Quantization (LVQ, LVQ2) algorithms proposed by Kohonen et al.[3] are classified as one of the neural network approaches. They showed that the LVQ2 algorithm was superior to the LVQ algorithm. However, in the LVQ2 algorithm, two reference vectors are modified at the same time if the first nearest class to an input vector is incorrect and the second nearest class to the input vector is correct. That is, if the given vector is recognized as the third rank, the modification is not occurred. McDermott et al.[4] developed a shift-tolerant phoneme recognition system based on the LVQ2 algorithm. Iwamida et al.[5] developed a LVQ-HMM phoneme recognition system. In this system, speech is at first transformed to a vector-code sequence using a code-book made by the LVQ2 algorithm and then the discrete type of HMM is applied to the vector-code sequence. However, those two systems did not make a phoneme sequence hypothesis but discriminated a phoneme from a given phoneme group for a given segment.

As described above, there still remain several problems in constructing a phoneme recognition system using the LVQ2 algorithm as follows:

(1) No training algorithm if the rank of the given vector is greater than the second rank, and

(2) No segmentation and recognition method for continuous speech.

In this paper, at first, we will propose three kinds of modified Learning Vector Quantization algorithms (MLVQ2). Next, we will investigate the optimum dimension to represent the reference vectors. Finally we will construct a phoneme recognition system by integrating the 2-level DP-Matching[6]. The system produces a phoneme sequence hypothesis by taking into account phoneme duration constraints.

2.0 MODIFIED LEARNING VECTOR QUANTIZATION ALGORITHMS(MLVQ2)

The Learning Vector Quantization algorithm constructs non-linear boundaries for classification problem using a training algorithm. In the LVQ2 algorithm, the reference vector is modified when a given training vector x satisfies the following three conditions: 1)the nearest class to the given vector must be incorrect, 2) the next-nearest class to the given vector must be correct and 3)the training vector must fall 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 if the rank of the given vector was greater than the second rank.

We propose three modified training algorithms for the LVQ2 algorithm. In the modified LVQ2(MLVQ2) algorithms, p reference vectors are modified at the same time if the correct class is within the N-th rank where N is set to some constant. The modified LVQ2 algorithms consists of the following 6 steps. In step 1, reference vectors are chosen using the K-Means clustering method from each class. In step 2, the nearest reference vector of each class to an input vector is selected. In step 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 m_n. In step 4, n is checked to see whether or not n falls in the range of $2 \leq n \leq N$. In step 5, the 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 m_1 and m_n. In step 6, the i-th reference vector is modified according to one of the following three versions of the modified LVQ2 algorithm.

0-7803-0532-9/92 $3.00 © 1992 IEEE

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on April 12,2010 at 02:52:02 UTC from IEEE Xplore. Restrictions apply.

MLVQ2.a[7,8] step 6:

$$[m_i]^{t+1} = [m_i - \alpha_n(t)(x - m_i)]^t \qquad (1)$$

$$(i=1, \cdots, n-1)$$

$$[m_n]^{t+1} = [m_n + \alpha_n(t)(x - m_n)]^t \qquad (2)$$

MLVQ2.b step 6:

$$[m_{n-1}]^{t+1} = [m_{n-1} - \alpha_1(t)(x - m_{n-1})]^t \qquad (3)$$

$$[m_n]^{t+1} = [m_n + \alpha_1(t)(x - m_n)]^t \qquad (4)$$

MLVQ2.c step 6:

$$[m_i]^{t+1} = [m_i - \frac{\alpha_1(t)}{n-1}(x - m_i)]^t \qquad (5)$$

$$(i=1, \cdots, n-1)$$

$$[m_n]^{t+1} = [m_n + \alpha_1(t)(x - m_n)]^t \qquad (6)$$

where,$\alpha_n(t)(0<\alpha_n(t)\ll 1)$

$$\alpha_n(t) = \alpha_0(1 - \frac{t}{T})^n \qquad (7)$$

T = No. of iterations × No. of samples. $\alpha_0 = 0.02$

In the MLVQ2.a, if the correct training vector is recognized as the n-th rank, the top $n-1$ reference vectors are moved away by α_n and the n-th reference vector is moved nearer by α_n. In the MLVQ2.b, the $n-1$-th reference vector is moved away by α_1 and the n-th reference vector is moved nearer by α_1. In the MLVQ2.c,the top $n-1$ reference vectors are moved away by $\alpha_1/(n-1)$ and the n-th reference vector is moved nearer by α_1.

3.0 COMPARISON AMONG THE VARIOUS LVQ ALGORITHMS

Recognition experiments were carried out for comparing the various LVQ algorithms. The recognition experiments were carried out for given phoneme segments: the beginning and final frames of each input phoneme are given. The trainings were carried out for phoneme samples in the 212 word vocabulary uttered by 7 male and 8 female speakers. The recognition experiments of 30 phonemes were carried out for phoneme samples in the 212 word vocabulary uttered by another 3 male and 2 female speakers.

Speech is analyzed by a 29 channel band-pass filter bank. The speech is represented by a sequence of logarithmic spectra with 10-ms frame shift.

The phoneme recognition system for the comparison is similar to the shift-tolerant model proposed by McDermott et al.[4]:

(1) 8 mel-cepstrum coefficients and 8 Δ mel-cepstrum coefficients are computed for every frame from the logarithmic spectrum. The value of each coefficient is normalized by the maximum magnitude of each coefficient. Each reference vector is represented by 112 coefficients(7 frames × 16 coefficients). Each class was assigned 15 reference vectors chosen by the K-Means clustering method.

(2) A 7-frame window is moved over the given phoneme segment and yields a 112(16x7) dimensional input vector every frame.

(3) In the training stage the various LVQ2 algorithms are applied to the input vector as described above.

(4) In the recognition stage we compute distances between the input vector and the nearest reference vector within each class.

(5) From this distance measure, each class was assigned an activation value a_w as follows:

$$a_w(c,t) = 1 - d(c,t) \;/\; \sum_i d(i,t) \qquad (8)$$

where d, c and t are distance, class and time, respectively.

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

(7) A class with the maximum activation value is regarded as a recognized output.

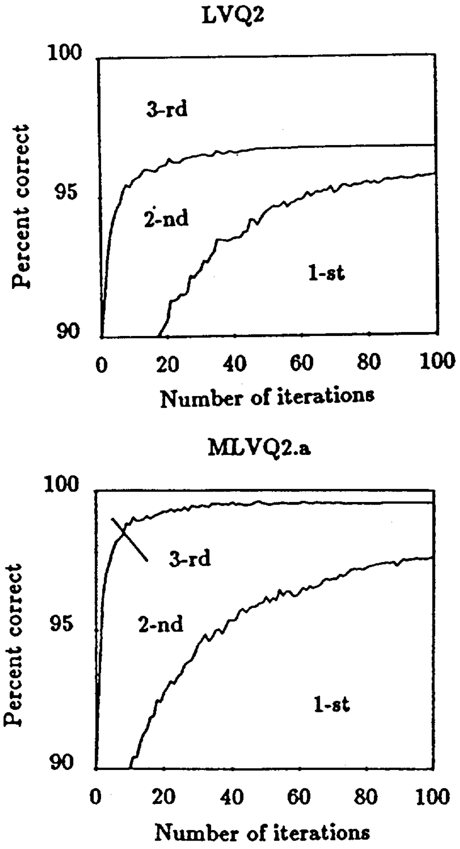

Figure 1 Relation between percent correct and number of iterations

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on April 12,2010 at 02:52:02 UTC from IEEE Xplore. Restrictions apply.

Figure 1 shows the relation between percent correct in discrimination among /b/, /d/ and /g/ consonants, and number of iterations. The modified LVQ2 algorithms gave higher recognition scores compared to the LVQ2 algorithm because the modification was carried out even if the rank of the given vector was greater than the second rank. Furthermore, the top-2 recognition scores obtained by the MLVQ2 algorithms were higher than those obtained by the LVQ2 algorithm.

Table 1 shows the phoneme recognition scores of 30 phonemes obtained by the various algorithms. The recognition scores obtained by the modified LVQ2 algorithms are higher by about 5% than those obtained by the LVQ2 algorithm. However, the MLVQ2.a gave a lower recognition score compared to the LVQ2 when the number of iteration increased and N was set to 30. In the MLVQ2.a algorithm, there remains a problem to define the value of N. The recognition scores obtained by the MLVQ2.b and MLVQ2.c increased as the number of iteration and N increased. Accordingly we will use the MLVQ2.b for phoneme recognition hereafter.

Table 1 Comparison among various LVQ algorithms

No. of iteration = 10					
Algorithm	LVQ2	MLVQ2.a		MLVQ2.b	MLVQ2.c
N	(2)	10	30	(30)	(30)
Training set	82.4%	88.0	87.5	86.7	87.0
Test set	78.4	83.5	83.3	83.2	83.1

No. of iteration = 50					
Algorithm	LVQ2	MLVQ2.a		MLVQ2.b	MLVQ2.c
N	(2)	5	30	(30)	(30)
Training set	86.4%	91.7	84.4	92.1	92.3
Test set	81.3	85.1	80.6	85.5	85.9

4.0 INVESTIGATION ON OPTIMUM DIMENSION OF REFERENCE VECTORS

In the experiments described above, we used the 112 dimensional vector computed from the 7-frame window, where each frame is represented by the 8 mel-cepstrum coefficients and the 8 Δ mel-cepstrum coefficients computed by the regression analysis over 5 frames. In this section, we will investigate the optimum dimension for representing a reference vector. We will investigate the following dimension:

(1) Number(N_c) of the mel-cepstrum coefficients computed in every frame

(2) Number(N_d) of the Δ mel-cepstrum coefficients computed in every frame

(3) Number(N_s) of frames of the time span for computing Δ mel-cepstrum

(4) Number(N_w) of frames of the time window of the reference vector

We examined those variables described above by phoneme recognition experiments for /b/, /d/ and /g/ samples in the 212 word vocabulary uttered by 3 male and 2 female speakers, where the reference vector of each phoneme was constructed using speech samples uttered by another 7 male and 8 female speakers.

At first, N_w is set to 7 and then the optimum N_c and N_d are examined for N_s=3,5,7. The combination of 8 mel-cepstrum coefficients and 8 Δ mel-cepstrum coefficients or 16 Δ mel-cepstrum coefficients gave the best recognition scores for the test set. Next, the optimum N_s is examined. The two kinds of parameters showed the best recognition scores at N_s=5. There is no significant difference between the two kinds of parameters in the recognition scores. We will use 8 mel-cepstrum coefficients and 8 Δ mel-cepstrum coefficients obtained from 5-frame span hereafter. Under the conditions previously-described, we investigated on the optimum N_w. The recognition scores did not show a sharp peak. We will use 7 frames for N_w because the recognition score reached a plateau for the training set and relatively higher recognition score was obtained for the test set.

5.0 RECOGNITION OF PHONEMES IN SPOKEN WORDS

The recognition scores described in the previous section were obtained for the given segments. It is necessary to carry out segmentation of speech for used as an acoustic processor in continuous speech recognition system. It is desirable to carry out simultaneously recognition and segmentation of phonemes. In this paper, we will use the 2-level DP-matching for recognition and segmentation of phonemes in continuous speech. The phoneme recognition system is as follows:

(1) A 7-frame window is moved over an input speech and yields a 112(16x7) dimensional input vector every frame.

(2) The distances between the input vector and the nearest reference vector within each class are computed every frame.

(3) An optimum hypothesis for a phoneme sequence is made using the 2-level DP-matching by taking into account phoneme duration constraints.

The following four kinds of constraints are examined for integrating to the 2-level DP-matching.

(a) Minimum and maximum duration constraints of phoneme independent of the context.

(b) (a)+phoneme connection constraints between successive two phonemes.

(c) Minimum duration constraints of phoneme dependent on the preceding phoneme, where the maximum duration constraints of phoneme is defined independent of the context.

(d) Minimum duration constraints of phoneme dependent on the preceding phoneme, where no constraints are used for the maximum duration constraints of phoneme.

The recognition experiments were carried out for evaluating the effectiveness of the duration constraints mentioned-above. The experimental conditions were the same as those described in the section 3.

Table 2 shows the recognition scores for the various constraints. As can been seen from the table 2, the minimum duration constraints of phoneme dependent on the preceding phoneme are very effective,

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on April 12,2010 at 02:52:02 UTC from IEEE Xplore. Restrictions apply.

on the contrary, the maximum duration constraints of phoneme are not necessary.

Table 2 Recognition scores for four kinds of duration constraints

Condition	Training set		
	Recognition score	Insertion score	Deletion score
(a)	96.2	38.2	0.2
(b)	95.8	13.2	0.3
(c)	95.4	3.7	0.6
(d)	95.3	3.7	0.7

Condition	Test set		
	Recognition score	Insertion score	Deletion score
(a)	91.1	57.6	0.3
(b)	89.7	23.9	0.8
(c)	89.1	7.3	1.1
(d)	89.1	7.3	1.2

Next we investigated the effectiveness of the following methods, where d_c is a Euclid distance of a phoneme class.

(a) Method using the square of the Euclid distance and the 2-level DP-matching

$$d2_c = d_c^2 \tag{9}$$

(b) Method using the following activation value and the 2-level DP-matching

$$a_c = 1-d_c / \sum d_i \tag{10}$$

(c) Method using the activation value and the DP-matching for selecting an optimum phoneme sequence[8,9]

(d) Method using the logarithmic activation value and the 2-level DP-matching

Table 3 shows recognition scores for the test set using the various kinds of methods. By comparing the method (b) to the method (c), the 2-level DP-matching is superior to the DP for selecting an optimum phoneme sequence because the 2-level DP-matching uses the information concerning to the phonemes with the rank $\geq$2. All distances or activation values gave similar performances.

Table 3 Recognition scores for four kinds of methods

Method	Recognition score	Insertion score	Deletion score
(a)	89.2	7.5	1.2
(b)	89.1	7.3	1.1
(c)	86.4	8.8	2.2
(d)	89.1	7.3	1.1

6.0 RECOGNITION OF PHONEMES IN CONTINUOUS SPEECH

Recognition experiments were carried out for continuous speech uttered Bunsetsu-by-Bunsetsu. Each of two adult male uttered 148 sentences. The sentence speech were analyzed in the same fashion as described in the section 3. The additional training was carried out for phoneme samples in 70 sentences uttered by the two male speakers, where the reference vectors of each phoneme obtained from the spoken words were used as the initial values. The recognition experiment of 30 phonemes was carried out for phoneme samples in the remaining 226 sentences uttered by the same two speakers. The recognition scores reached a plateau at ten iterations. The recognition score, insertion score and deletion score were 85.5%, 6.9% and 4.0%, respectively.

7.0 CONCLUSION

We proposed three modified LVQ2 algorithms and showed their superiority to the original LVQ2 algorithm. We also showed that the 2-level DP-matching using the Euclid distance obtained by the MLVQ2.b gave the best performance. The minimum duration of phoneme dependent on the preceding phoneme is the most effective constraint to achieve a high recognition score. In order to apply reference vectors obtained from spoken words to continuous speech, ten iterations in the additional training are sufficient for the adaptation.

REFERENCES

[1] S. Makino, T. Kawabata and K. Kido, "Recognition of Consonant Based on The Perceptron Model", Proc. of ICASSP-83, pp. 738-741(April, 1983)

[2] S. Moriai, S. Makino and K. Kido, "Phoneme Recognition in Continuous Speech Using Phoneme Discriminant Filters", Proc. of ICASSP-86, pp. 2251-2254(April, 1986)

[3] 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)

[4] E. McDermott and S. Katagiri, "Shift-Invariant Phoneme Recognition Using Kohonen Networks", Proc. ASJ meeting, pp. 217-218(October, 1988)

[5] H. Iwamida, S. Katagiri, E. McDermott and Y. Tohkura, "A Hybrid Speech Recognition System Using HMMs with An LVQ-Trained Codebook", Proc. of ICASSP-90, pp. 489-492(April, 1990)

[6] H. Sakoe:"Two-Level DP-Matching -A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition-", IEEE Trans. Acoust. Speech and Signal Process., ASSP-27, 6, pp. 588-595(1979)

[7] M. Endo, S.Makino and K. Kido, "Phoneme Recognition Using A LVQ2 Method", Trans. IEICEJ, SP89-50 (September, 1989) (in Japanese)

[8] S. Makino, A. Ito, M. Endo and K. Kido,"A Japanese Dictation System Based on Phoneme Recognition and A Dependency Grammar", Proc. of ICASSP-91, (May, 1991)

[9] S. Makino, S. Moriai and K. Kido, "A Method for Selecting An Optimum Phoneme Sequence Using A Posteriori Probabilities of Phonemes", Journal of ASA supplement No. 1, PPP5 (November, 1988)

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on April 12,2010 at 02:52:02 UTC from IEEE Xplore. Restrictions apply.