NEW NON-SUPERVISED LEARNING METHODS FOR SPEAKER ADAPTATION

Osamu Kakusho and Riichiro Mizoguchi

The Institute of Scientific and Industrial Research,
Osaka University, 8-1 Mihogaoka, Ibaraki, 567 Japan.

ABSTRACT: An inter-related phoneme template system is proposed together with its two nonsupervised learning algorithms. Their efficiency is verified through some computer experiments of word recognition.

1. INTRODUCTION

This paper is concerned with automatic speaker adaptation for speaker independent recognition. A new phoneme template system composed of inter-related phoneme templates is proposed[1] along with two efficient non-supervised learning algorithms. One is based on the selection of the inter-related phoneme templates from a set of templates prepared beforehand. The other is based on the creation of new templates appropriate for each speaker. The former algorithm is performed in "on-line" mode, that is, the selection is made every time a word is uttered. It is useful for rapid adaptation. The latter is performed in "batch" mode, that is, the creation is made after a reasonable amount of words are obtained. Although the adaptation is done one or two days after the first usage, almost complete adaptation can be made in this learning algorithm. The performance of these two non-supervised learning algorithms is verified by computer simulation of a word recognition system.

2. INTER-RELATED PHONEME TEMPLATES

2.1 Construction method

step 1: For each speaker, make augmented feature vectors of the dimensionality 5d by combining every feature vector of the dimensionality d of the frame corresponding to Japanese five vowels.

step 2: Apply k-means method[3] to the augmented vectors and obtain representative vectors of the clusters (one from each cluster).

step 3: Decompose the representative vectors into the original form, each of which is considered as a template of a vowel.

2.2 An example of the inter-related phoneme template

Fig. 1 shows four inter-related templates (pentagons) represented in a two dimensional space composed of the first and second formants frequencies. Speech samples are drawn from the isolated vowels uttered by ten male adults. The vertices of each pentagon are template patterns.

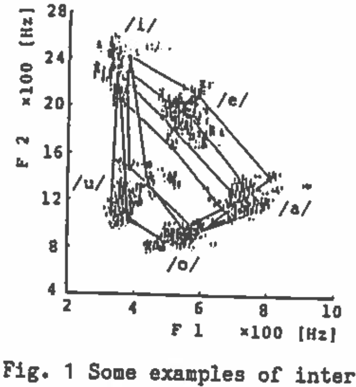

Fig. 1 Some examples of inter-related phoneme templates.

3. NON-SUPERVISED LEARNING METHOD OF ON-LINE TYPE

3.1 Algorithm

Let use-count of a template be defined as a number of input patterns which match best with the template. And let use-count of an inter-related template be defined as a sum of the use-count of the templates contained in it. Then, we have the following learning(selection) algorithm.

step 1: Calculate the use-count of all the templates.

step 2: Select the inter-related template of the maximal use-count.

This algorithm is based on the selection of the templates according to the use-count which are obtained without using the identities of the input patterns. Therefore, it is a non-supervised learning algorithm. The selection can be performed in any time period and in any scheme.

3.2 Evaluation

3.2.1 Vowel recognition

a) Speech samples

Japanese five vowels uttered consecutively like /ieaou/ by 15 male adults were analyzed with LPC method (10kHz sampling, auto-correlation, order 12, and hamming window of length 20ms with shift interval 10ms). Each speaker uttered a sequence of vowels five times. Two of them were used for template construction (600 frames in all, 600 = 15men x 5vowels x 8frames), and the rest of them were used for learning and recognition (675 frames in all, 675 = 15men x 5vowels x 9frames).

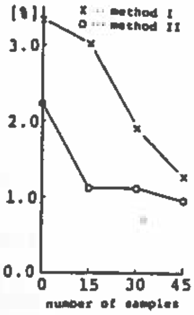

Fig. 2 Learning process(1).

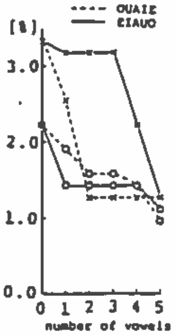

Fig. 3 Learning process(2).

Table 1. Recognition rates.

spkr	bef. L.	aft. L.	
	Rc %	RIR %	Rc %
KI	100	50	98
SN	98	24	100
MA	92	80	96
KO	98	10	98
YA	90	22	98
YU	98	-25	96
YM	98	16	100
AVE.	96.3	25	98.0

b) Recognition method

Recognition is made using the template matching in the 4-dimensional Fischer space constructed based on the samples for template construction.

c) Experiment

As described in a), the speakers used for template construction and recognition are identical, so this is a closed recognition as to the speaker.

Ten inter-related templates were obtained according to the method described in section 2. Fig. 2 depicts the learning process of vowel recognition. The vertical axis shows the error rates(%) and the horizontal one the number of vowels given to the system. The letter "x" denotes the error rate of the conventional templates and the letter "o" that of the proposed template. In order to see the effect of the order of vowels on the learning performance, simulation was done for two kinds of sequences. The error rates corresponding to the sequence /ouaie/ are depicted by broken line and those corresponding to /eiauo/ are depicted by solid one. Selection of the templates is done as follows: Every time when learning of a vowel is done, for the conventional templates, one template corresponding to the vowel is chosen from the templates. For our template, on the other hand, six inter-related templates are chosen after the learning of the first vowel is done. Four, three, two and one inter-related template are chosen after the learning of the second, third, fourth and last vowel is done, respectively. It is seen from Fig. 2 that learning of our templates does not depend on the sequence of vowels, while that of the conventional ones depends largely on the sequence.

Fig. 3 shows the relation between the error rates and the number of learning samples given to the system until it makes the final selection of the template. This figure demonstrates that learning of the inter-related templates is much faster than that of the conventional ones. Some advantages of the non-supervised learning method of the inter-related template are summarized below.

1) Adaptation is fast.
2) Learning process is stable.
3) Learning process is reliable.

3.2.2 Word recognition

Vocabulary of the system is composed of Japanese ten digits 0(/rei/) through 9(/kyu/). Open recognition as to seven male adults was done, where eight inter-related templates were prepared before hand. Table 1 shows the recognition rates for seven speakers. RIR parameter represents the ratio of the improved recognition rate of vowels contained in the digits. This results shows the effectiveness of our non-supervised algorithm.

4. NON-SUPERVISED LEARNING OF BATCH TYPE

The learning method proposed above is based on the selection of a template from a set of them prepared in advance. Therefore, performance of the learning depends on the speakers, that is, adaptation(selection) is done successfully only when at least one template appropriate to the speaker is stored in the system. When no such template is stored, however, much improvement can not be attained by the learning. In order to make the learning more effective, another learning method is proposed in this section.

The learning method creates new templates appropriate to the speakers rather than selection of them. To do this, the algorithm needs a reasonable amount of sample words. Consequently, adaptation to a speaker is made one or two days after his first use of the system. This is why the algorithm is said to be of batch type.

4.1 Algorithm

The block diagram of the total system is shown in Fig. 4, in which a block surrounded by broken line corresponds to the proposed learning algorithm.

4.1.1 Clustering of input words

A clustering method [2] is applied to respective sets of words uttered by a speaker. Since the clustering algorithm requires only a distance matrix as input data, it is easily executed.

4.1.2 Identification of the categories of the clusters

Clusters obtained above are labeled according to the majority rule using the labels given by the system itself. There are two alternatives of the treatment of the minorities in the rest of the operations:

A) Reject them and
B) Relabel them to the category of the majority.

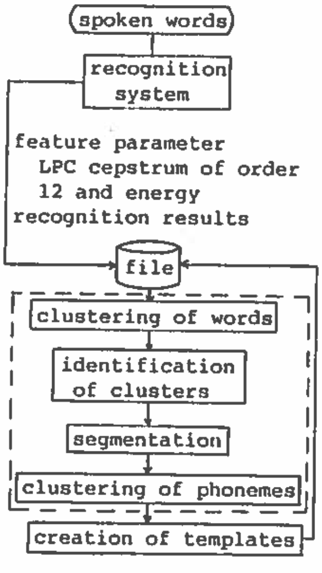

Fig. 4 Block diagram of the total system.

In the case of A), the new templates are created by using only words supposed to be recognized successfully by the system. In the case of B), on the other hand, such words that are supposed to be mis-recognized are also used for creating new templates.

4.1.3 Segmentation

Segmentation of every word is done by using the energy and label associated with it.

4.1.4 Creation of the inter-related templates

According to the operation thus far, a set of frames labeled one of the Japanese vowels are obtained. New templates are created from these frames by clustering procedures shown below.

step 1: K-means method of number of clusters 8 is applied to the whole set of frames.

step 2: For each vowel, count the number of frames belonging to respective clusters.

step 3: For each vowel, select major groups until they cover 80% of population of the vowel. And consider them as the templates of the corresponding vowel.

step 4: Register the template as an inter-related template.

Table 2 Recognition rates(%).

spkr	T-1	T-2	T-3
A	98.0	94.0	100
B	91.8	97.9	91.8
C	98.0	100	100
D	93.9	98.0	100
E	90.0	92.0	98.0
F	98.0	98.0	98.0
G	68.8	91.3	93.5
H	76.0	84.0	94.0
AVE	89.4	94.4	96.9

4.2 Evaluation

Performance evaluation of the proposed learning method was done in word recognition of 32 words. Phoneme templates were obtained from the words uttered by 31 male adults. Fifteen times utterances of ten words in the vocabulary made by eight speakers other than the 31 speakers were used for the evaluation. Ten utterances were used for initial recognition and collection of data for the non-supervised learning. The final recognition was done by using the rest of 5 utterances.

Table 3 Improvement of vowel recognition(%).

speaker	1->2	1->3
F	3.7	29.6
G	54.4	57.9

We have three sets of templates:

T-1: 31 templates before learning
T-2: Template created using category identification A) in step 2.
T-3: Template created using category identification B) in step 2.

The results are shown in Tables 2 and 3. Table 2 shows the recognition rates of the respective speakers and Table 3 shows the values of RIR. It is seen from both tables that the batch type non-supervised learning algorithm attains much improvement especially for the speakers having low initial recognition rates. Furthermore, performance of T-3 is slightly better than that of T-2. This is because T-3 is constructed from the mis-recognized words as well as recognized ones.

5. CONCLUDING REMARKS

Inter-related phoneme templates have been proposed together with two types of non-supervised learning algorithms. The results of the computer experiment has demonstrated the efficiency of them and shown the possibilities of this application to the real world situations.

(References)

[1] Mizoguchi, R., et al.:"Word recognition system for unspecified people based on inter-related phoneme templates", Trans. of the IECE Japan, Vol. J67-A, 6, pp. 572-579, 1984.
[2] Mizoguchi, R., et al.: "A nonparametric algorithm for detecting clusters using hierarchical structure", IEEE Trans., PAMI-2, 4, pp. 292-300, 1980.
[3] Anderberg, M.R.:"Cluster analysis for applications", Academic Press, 1973.