A Study on the Isolated word Recognition Using One-Stage DMS/DP for the Implementation of Voice Dialing System

Seong-Kwon Lee, Kang-Seong Lee, Soon-Hyob Kim

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

ABSTRACT

The speech recognition systems using VQ have usually the problem decreasing recognition rate, MSVQ assigning the dissimilar vectors to a segment. In this paper, applying One-stage DMS/DP algorithm to the recognition experiments, we can solve these problems to what degree. Recognition experiment is peformed for Korean DDD area names with DMS model of 20 sections and word unit template. We carried out the experiment in speaker dependent and speaker independent, and get a recognition rates of 97.7 % and 81.7 % respectively.

1. INTRODUCTION

In the experiment, we choose the Korean DDD area names to make a voice dialing system[11]. The VQ(Vector Quantization)[1] quantizing datum into an unit of vector is efficient in computing time, but inefficient in recognition rate. MSVQ(Multi Section VQ)[4] extending VQ divides speech signal into constant length segments and produces codewords in each segment. Because time duration of each syllable in a string is different, MSVQ algorithm assigns dissimilar vectors to a same segment. To solve these problems, the DMS divides the string dynamically to make similar vectors one segment according to the feature. Therefore the short speech feature such as fricatives can be chosen to be a representative feature vector[2]. In this point of view, we call it Dynamic Multi-Section (DMS) model and it makes each model get a time duration information. The reference patterns in our experiments are made by using the DMS model.

2. THE DMS MODELING PROCESS

The DMS model[8] has to divide a word string into some segments dynamically according to the features, and make them have a representative feature vector for each segment and time duration information. Therefore complicated procedures are needed. It consists of two steps. First step is the dynamic segmentation, and Second is the getting representative feature vectors and time duration information. Figure 1. shows the flow chart to make a model.

2.1. DYNAMIC SEGMENTATION ALGORITHM

The jth (1≤j≤J, J is number of DMS model segments) segment of DMS model M for each word has the segment information M(j), M(j) consists of feature vector m_j and duration information p_j. First we segment learning datum equally in the time axis, and gather feature vectors in a segment to get a centroid. This centroid is a representative feature vector in a segment. Considering all segments are same size and dividing the last frame number in each segment by the total number of frames, we can get a time duration information. It is represented as the following.

$$p_j = \sum_{m=1}^{M} e_m(j) \ / \ \sum_{m=1}^{M} I_m \ , \ 1 \leq j \leq J \tag{2-1}$$

where, $e_m\ (j)$ is the last frame number of jth segment, I_m is the number of total frame for learning data, M is the number of total learning datum. As we perform DP-matching algorithm between the learning data and the initial word model, we can change the segment boundary. Getting total accumulated distance by DP-matching and assigning the last frame number in new segment by backtracking, we can get a new centroid in the new segment. This new centroid is regarded as a representative feature vector and model is updated. Simultaneously, the ratio of the number of frames in each segment to total number of frames is registered as a duration information. After updating word model, if the sum of the total accumulated distance by DP matching for each learning data is smaller than the prior distance, then the procedure must be repeated, otherwise, the procedure stops since the model can't be more stable.DP algorithm includes accumulated distance D and distance p_j by time duration information.

$$D(i,\ j) = d_u(t_i,m_j) + \min\left\{ \begin{array}{l} D(i-1,j)\ , \quad (1<i\leq I, 1<j\leq J) \\ D(i-1,j-1) + P_{j-1} \end{array} \right. \tag{2-2}$$

Total distance between learning data and model is obtained by (2-2) and P(j) is the difference distance of time duration information between the ith frame of the learning data and the last frame of jth segment of word model.

$$p_j = W * d_s(e(j),\ i) \tag{2-3}$$

$$d_s(e(j),\ i) = \mid (p_j * I) - i \mid \tag{2-4}$$

Also, W is weighting value for the difference between duration time information and effecting element for an optimal recognition rate.

3. One-Stage DMS/DP RECOGNITION ALGORITHM

3.1. The elementary step of algorithm

We first define the basic concepts before explaining a detail algorithm.
1) Unknown input test pattern consists of i = 1,...,N time frames
2) Reference patterns compared with input word sequences consist of the k words of one or more syllables, that is the collectivity of templates. This includes the representative feature vectors computed in DMS modeling and doesn't include time duration information. From now, we call reference pattern (or template) a DMS template.[14]
3) j = 1,..., J(k) presents the time frame of DMS template k, J(k) presents the length of DMS template.
Our goal is the decision of the best connection of DMS templates, q(1),...,q(R). q(1),...,q(R), what we call "super" reference pattern.

3.2 One-Stage DMS/DP algorithm

Based on the prior concepts, One-Stage DMS/DP algorithm is followed.
Step 1) Initialization

$$D(1, j, k) = \sum_{j=1}^{J} d(1, j, k) \qquad (3-1)$$

Step 2) a) Performing step 2b-2c for i=2,...,N
b) Performing step 2c-2e for k=2,...,K
c) $D(i, 1, k) = d(i, 1, k) + \min\left\{ \begin{array}{l} D(i-1, 1, k), \\ D(i-1, J(k*), k*) \end{array} \right. , k* = 1, \dots, K$

(3-2) d) Performing step 2c for j=2,...,J(k)

e) $$D(i, j, k) = d(i, j, k) + \min\left\{ \begin{array}{l} D(i-1, j, k), \\ D(i-1, j-1, k), \\ D(i, j-1, k) \end{array} \right. \qquad (3-3)$$

Step 3) Using the accumulated distance array D(i,j,k), the best path can be traced reversly from the end frame in DMS template which have the minimum total accumulated distance.

In step 3, the unknown sequences of input words are found by back tracking the decision achieved by minimum operator at each lattice point. For this backtracking procedure, array D(i,j,k) of accumulated distance is to be stored while the recursive related expression is calculated.

4. THE RESULTS OF EXPERIMENT

4.1. The process of making speech data base and DMS Template

Speech data is filtered by a 300Hz-4.5kHz BPF, sampled at 10kHz, converted 16bit data, pre-emphasized, covered a Hamming window, calculated a 10's order autocorrelation, that is LPC coefficient[12] by each 128 samples,and finally calculated LPC cepstrum coefficient. The template is constructed by DMS model. Speech DB is the 146 DDD local names spoken by three male speakers. DMS template is experimented in speaker dependent and independent with three types of model. It is shown in Table 1.

4.2. Implementing a recognition system

The speech recognition system is implemented on IBM PC/486 with TMS320C30 DSP board[10][13]. It is shown in Fig. 3

4.3 Recognition experiment

First we divide each word to 20 sections and make templates to get a good recognition rate. The results are shown in Table 2-6.

5. DISCUSSION

At the case that syllable is applied as a recognition unit, it takes about 2-4 seconds for recognition and the recognition rate is low. At the case of word, it takes about 10-20 seconds with high recognition rate comparatively (about 97.7% in the speaker dependent, 81.7% in the speaker independent). In the experiment, the speech datum spoken in the slow utterances resulted in the recognition rate of 70% below. Therefore, we have to consider the more efficient time normalization and coarticulation problems. For good recognition rate, the number of segments and weights are fixed to 20, 0.6, respectively, and the general experiment results are obtained as the following : 1) Because the computing time is proportion to the number of segments, the tradeoff of the recognition rate to the computing time has to be considered properly.2) It is necessary that the time normalization and coarticulation are studded more seriously.3) Speakers have to utter circumspectly.

6. CONCLUSION

In this paper, we experiment the isolated words recognition in medium scale vocabulary for a real-time voice dialing system. The applied recognition algorithm is the One-Stage DP method that is combined with DMS template method. The recognition is performed with the 146 DDD area names. As a result of the experiment, we get 97.7%, 81.2% recognition rate in speaker dependent and independent, and it takes about 7-14 seconds in a recognition. For a real-time voice dialing system[14], a reference pattern in terms of subword unit must be studied for DMS templates. The algorithm with subword unit, having a effective time normalization, solving a coarticulation effect, must be studied also.

REFERENCE

1. R.M. Gray, "Vector Quantization", IEEE ASSP Magazine, Vol.1, pp. 4-29, Apr. 1984.

2. Soonhyob-Kim, "A study on the analysis and automatic recognition Korean speech",Yonsei Univ. Ph.D, 1982. 12.

3. Jong-Kwan Eun, "The Study of Korean speech recognition system development", KAIST, final report, 1988. 1.

4. D.K. Burton, J.E. Shore and J.T. Buck, "Isolated Word Speech Recognition using Multisection Vector Quantization Codebooks", Vol. ASSP 33, No. 4, Aug. 1985.

5. H. Ney, "The Use of a One-Stage Dynamic Programming Alorithm for Connected Word Recognition", IEEE Trans. on ASSP, Vol. ASSP-32, No. 2, pp. 263 ~ 271, April 1984

6.C.S. Myers and L. R. Rabiner, "A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition", The Bell System Technical Journal, Vol. 60, No. 7, pp. 1389 ~ 1409, September 1981

7. H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans.on Acoustics,Speech and Signal Processing, Vol. ASSP 26, No.1, pp.43-49, Feb. 1978.

8. Y. G. Byun and S. H. Kim "A Study on Isolated word recognition using Dynamic Multisection Model" Ph.d thesis , Kwang Woon Univ.1991.2.

9.M.Immendorfer,"Voice Dialer",Electrical Communication,Vol.59,No3,1985

10. TMS320C30 PC SYSTEM BOARD USER MANUAL, Loughborough Sound Images Ltd. The Technology Centre, England, Ver. 1.0, Jan. 1990.

11. A. Fukui, Y. Fujihashi and F. Nakagawa, "Signal Processor Application to Voice Dialing Equipment", Proceedings ICASSP 86,TOKYO, pp.337~340, 1986

12. J.D. Markel, A.H. Gray, "Linear Prediction of Speech", Spring-Verlag Berline Heidelberg 1976

13. Third Generation TMS320 User's Guide, Texas Instrument, Inc., Houston, 1988.

14. C. Y. Kim and S. H. Kim "A study on connected word recognition for the implementation of a real-time voice dialing system ", Kwang Woon Univ. M.S. thesis. 1992.2

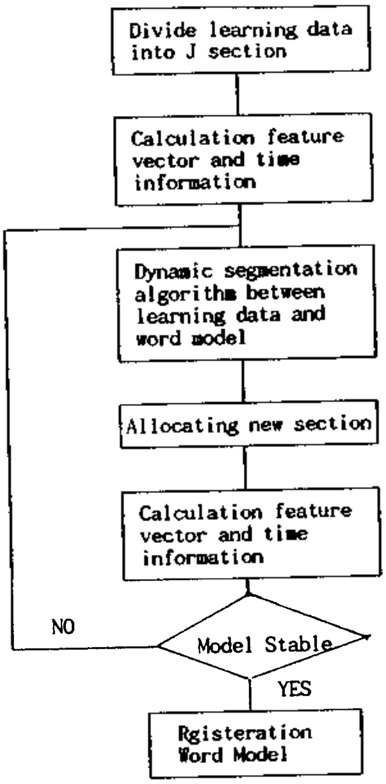
Divide learning data into J section
Calculation feature vector and time information
Dynamic segmentation algorithm between learning data and word model
Allocating new section
Calculation feature vector and time information
NO
Model Stable
YES
Rgisteration Word Model

Fig. 1 The process of word model generation

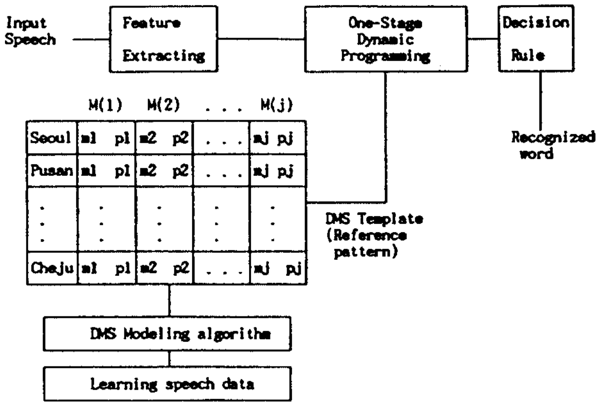
Input Speech
Feature Extracting
One-Stage Dynamic Programming
Decision Rule
Recognized word
M(1) M(2) . . . M(j)
Seoul m1 p1 m2 p2 . . . mj pj
Pusan m1 p1 m2 p2 . . . mj pj
Cheju m1 p1 m2 p2 . . . mj pj
DMS Template (Reference pattern)
DMS Modeling algorithm
Learning speech data

Fig. 3. Block diagram of overall recognition system using One-Stage DMS/DP

Model 1	Two utterances data of A,B,C Speakers
Model 2	Two utterances data of A,B,C,D Speakers
Model 3	Two utterances data of A,B,D Speakers

Table 1. Diagram of model template for speaker

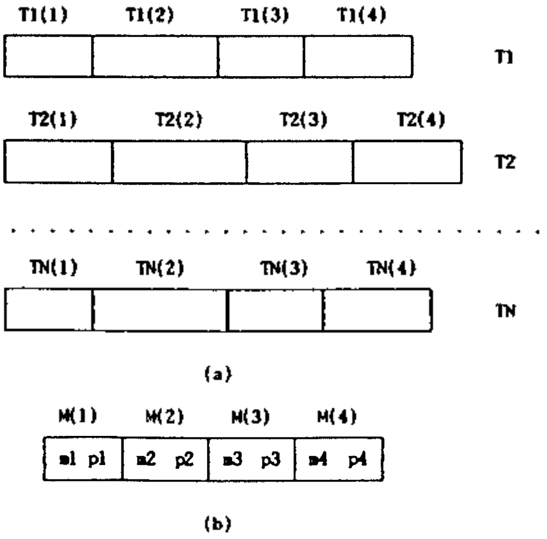

Fig. 2 An example of completed DMS model: (a) training data (b) DMS model

Speaker Independent	Model 1
D-1 Speaker	85.61 %
D-2 Speaker	87.67 %
D-3 Speaker	84.24 %

Table 2. Recognition result with model 1 for Speaker Independent

Speaker Dependent	Model 1
A Speaker	96.23 %
B Speaker	97.26 %
C Speaker	90.75 %

Table 3. Recognition result with model 1 for Speaker Dependent

Speaker Dependent	Model 2
A Speaker	97.26 *
B Speaker	95.20 *
C Speaker	97.26 *
D Speaker	95.54 *

Table 4. Recognition result with model 2 for Speaker Dependent

Speaker Dependent	Model 3
A Speaker	98.28 *
B Speaker	97.60 *
C Speaker	97.26 *

Table 5. Recognition result with model 3 for Speaker Dependent

Speaker Dependent	Model 2	Speaker Independent	Model 3
D-1 Speaker	93.15 *	C-1 Speaker	78.76 *
D-2 Speaker	89.72 *	C-2 Speaker	72.60 *

Table 6. Recognition result with model 2 for Speaker Dependent and with model 3 for speaker independent