A Study on Speech Recognition using DMS Model

DMS 모델을 이용한 음성 인식에 관한 연구

Tae Ock Ann*, Young Kyu Byun**
안 태 옥*, 변 용 규**

요 약

본 연구는 단어 패턴 중 유사한 특성의 정보에 기초를 둔 DMS(Dynamic Multi-Section) 모델을 제안한다.

이 모델은 각각의 단어를 몇개의 구간(Section)의 시계열로 분할하고, 각각의 구간 모두에 지속 시간 정보와 구간을 대표하는 특징 벡터를 구간의 정보로 등록해 둔 것이다.

단어 패턴에서 모델을 작성하는 절차는 대표 특징 벡터와 지속 시간의 정보를 거리에 따라 반영하면서 단어 패턴과 모델과의 매칭을 반복하여 매칭에 의한 누적 거리가 최소로 되도록 하는 것이다.

제안된 음성 인식 실험을 수행하는 것 이외에도 비교를 위해 DP 방법, HMM 방법 및 MSVQ 방법에 의한 음성 인식 실험을 같은 조건하에서 같은 데이타로 수행하였다.

또한, 제안된 DMS 모델을 이용한 음성인식시에도 DMS/DP 방법에 의한 인식 및 DMS/VQ 방법에 의한 인식 실험을 수행하여 비교하였다.

실험 결과, DP에 의한 인식률은 93.4%이고, HMM에 의한 인식률은 91.6%이며, MSVQ에 의한 인식률은 89.3%이다. 또한, DMS 모델을 이용한 DMS/DP에 의한 인식률은 95.8%이고, DMS/VQ에 의한 인식률은 96.8%이다.

그러므로, DMS 모델을 이용한 DMS/VQ 방법에 의한 인식이 일반적으로 많이 이용되고 있는 DP 방법이나 HMM 방법 및 MSVQ 방법과 비교해 볼 때 인식률도 우수하며, 기억 용량 및 계산량도 감소되어, 본 연구에서 제안하는 DMS 모델의 유용성이 입증되었다.

Abstract

This paper proposes a DMS(Dynamic Multi-Section) model based on the information of the simular features in word pattern.

This model represents each word as a time series of several sections and each section implies duration time information and typical feature vectors.

The procedure to make a model in the word pattern is that typical feature vector and duration time information are reflected in the distance, when matching between word pattern and model is repeated. As the result of it, the accumulated distance by matching is to be minimized.

---

*Department of Computer Engineering, CHONBUK SANUP University of Korea
(전북 산업 대학교 컴퓨터공학과)
**Department of Computer Science, Seoul Polytechnique University of Korea
(서울 산업 대학교 전자계산학과)
접수일자 : 1994년 8월 18일

Besides the proposed speech recognition experiments, for comparison with it, we perform the experiments by DP, HMM and MSVQ method under the same condition and data.

Also, in the experiments based on the proposed DMS model, we perform the experiments of DMS/DP(DP matching by DMS model) amd DMS/VQ(VQ method by DMS model), and compare with each other.

Through the experimental results, recognition rate by DP is 93.4%, by HMM is 91.6% and by MSVQ method is 89.3%. Also, in case of speech recognition using DMS model, recognition rate by DMS/DP is 95.8% and by DMS/VQ method is 96.8%.

Therefore, the recognition rate by DMS/VQ method is superior to those of conventional DP, HMM and MSVQ method. In addition, as the memory space and computational time of DMS model are reduced remarkably, the proposed DMS model proved to be a useful model.

I. Introduction

It is reasonable for small or medium scale vocabulary speech recognition system to use recognition units as word units than subword units like phonemes or syllables. Therefore, this paper carries out the recognition of isolated word of word unit by 146 Korean DDD area names.

For the recognition method of word unit, there are DP pattern matching method[1], MSVQ method[2-5] and HMM(Hidden Markov model)[6-9] method. Among them, DP and HMM, has been known for the recognition methods which produce highly recognition performance. DP, however, requires many memory for templates and has computational load. HMM, a probabilistic method, spends long time for training and needs many training tokens, and these collecting tokens are not easy problem when vocabulary is increased. MSVQ takes advantages of less memory for templates but its performance is not high.

In the pattern matching of word unit, most consonants is spoken shortly than vowels. Especially, plosives or affricates is spoken more shortly. But, one of disadvantage of DP is not considering the importance of utternace spoken relatively short time interval. For instance, stop sound spoken shortly is not considered this deeply in the case of DP or HMM.

DMS model is that each word is devided into some dynamic sections, and typical feature vector and duration time information are reflected in each section.

DMS model proposed in this paper has additional benifits by weight according to the length of phonemes. DP stores all time sequence of feature vectors for templates, but DMS model stores centroid vectors which means segment of acoustically similar features in each section. Centroid is extracted using clustering technique[10-12]. DMS model reduced redundancy feature vectors acoustically similar by dynamic section division and increases performance by using time information.

For making templates of DMS, we divide sections by section division algorithm like to use partially pattern matching method of DP. In testing mode, words are classified by dividing them into appropriate sections, and performing VQ on a section by section basiso Finally we find DMS model that yields the smallest average distortion. Recognition method using DMS model is compared with conventional DP and MSVQ.

For end point detection, ZCR and Energy parameters are used and 12-order LPC cepstrum coefficients were used for feature vectors.

II. Model Generation

Let's consider a word 'Jeju' as an example before making a model. When we speak 'Jeju' as waveform

shown Fig. 1. each phonemes can be divided by four sections.

Each phoneme in a word 'Jeju', contrasts with each section. At that time, it belongs some frames to phoneme 'ㅈ' of section 1. For the feature vectors, we can extract a typical feature vector and ratio of it's time can be used by duration time information. In section 2, 3, and 4, as the method of section 1, we can obtain typical feature vector and duration time information, respectively.

As above, each word is classified into some dynamic sections. Thereafter, we can make the model using typical feature vector and time information that is called DMS Model.

This model is not method that each section is divided into the fixed length like MSVQ, but method that it's section is divided dynamic(variable length).

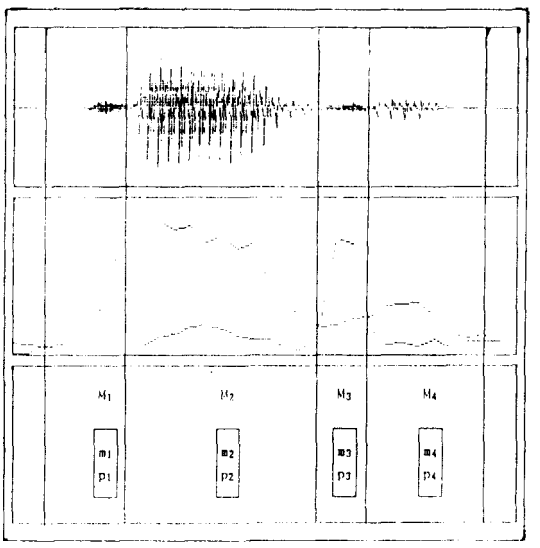


Fig 1. Waveform and word model

1. Section division algorithm

In generation of model M for each word, it is presented by time-sequence of J section and section j ($1 \le j \le J$) contains typical feature vector R(j) and time information P(j).

(Glossary of terms)

i : Frame number of input speech T($1 \le i \le I$)

j : Section number of word Model M($1 \le j \le J$)

$d_v(t_i, m_j)$: Local distance between feature vector t_i of i frame in test pattern and feature vector m_j of j frame in model

$d_s(p(j), i)$: Absolute of difference of final frame number of j section and frame number of test input pattern

D(i, j) : Accumulated distance between i frame and j section

W : Weight for distance of time information

DIS(T, M) : Accumulated distance between input speech T and Model M

e(j) : Final frame of j section in input speech

Now, we define the distance DIS(T, M) between unknown input test pattern $T = t_1, t_2, \ldots, t_i, \ldots, t_I$ (I = Frame number) and each word model M, as follows.

$$DIS(T, M) = \min \{ \sum_{j=1}^{J} [S(j) + P(j)] \} \quad (1)$$

Where, S(j) is a distance between the feature vector of the jth section of training data and typical feature vector of jth section of the word model.

$$S(j) = \sum_{i=m(j-1)+1}^{m(j)} d_v(t_i, m_j) \quad (2)$$

Where, function $d_v(a, b)$ is expressed with the distance betwen a and b. And, P(j) is defined with distarne by time information of jth section.

$$P(j) = W * d_s(p(j), i) \quad (3)$$

Where p(j) is ratio that can be used to divides final number of the jth section by final total frame number. Therefore, we give distance $d_s(p(j), i)$ as follows.

$$d_s(p(j), i) = | p(j) * I - i | \quad (4)$$

And, W is a weight and is given by experiment with changing value. So far, we can know that DIS(T, M) is evaluated with distance by typical feature vectors and time information P of the each section.

In this paper, time information is defined as ratio of final frame number of a certain section per final frame number and typical feature vector is given by clustering method.

When Training data T of I frame was divided into J section, final frame of jth section is represented e(j). Therefore, e(0) = 0 and e(J) = I.

In this paper, DIS(T, M) is computed by the DP algorithm. Conventional DP algorithm is

$$D(i, j) = d_v(t_i, m_j) + min \begin{bmatrix} D(i-1, j) \\ D(i-1, j-1) \end{bmatrix}$$

$$(1 < i \le I,\ 1 < j \le J) \tag{5}$$

and DIS(T, M) is given eq. (6).

$$DIS(T, M) = D(I, J) \tag{6}$$

In the proposed DMS model, eq. (5), (6) include the distance P for time information. Therefore, DIS(i, j) is given as follows.

$$D(i, j) = d_v(t_i, m_j) + min \begin{bmatrix} D(i-1, j) \\ D(i-1, j-1) + P(j-1) \end{bmatrix} \quad (1 < i \le I,\ 1 < j \le J) \tag{7}$$

Eq. (3) and (7) show that distance value of time information is the lowest when difference between frame number of input test pattern and final frame number by time information of j-1 section is the closest.

(Algorithm)

Step 1 : Initialization

$$D(1, 1) = d_v(t_1, m_1) \tag{8}$$

step 2 : Repeat for $2 \le i \le I - J + 1$

$$D(i, 1) = D(i-1, 1) + d_v(t_i, m_1) \tag{9}$$

Step 3 : Repeat for $2 \le j \le J$

$$D(j-1, j) = \infty \tag{10}$$

$$D(i, j) = d_v(t_i, m_j) + min \begin{bmatrix} D(i-1, j) \\ D(i-1, j-1) + d_s(p(j-1), i-1) \end{bmatrix}$$

$$(j \le i \le I - J + j) \tag{11}$$

Step 4 : Repeat for $J \ge j \ge 2$

(Tracking final frame of each section by backtracking)

$$\text{if } \{D(i, j) = D(i-1, j-1) + d_v(t_i, m_j) + W \cdot d_s (p(j-1), i-1))\},\ e(j-1) = i-1 \text{ and } j = j+1 \tag{12}$$

Step 5 : End

$$DIS(T, M) = D(I, J) \tag{13}$$

$$e(J) = I \tag{14}$$

2. Model generation method

It is necessary that words selected from training data are made for model with the best efficiency.

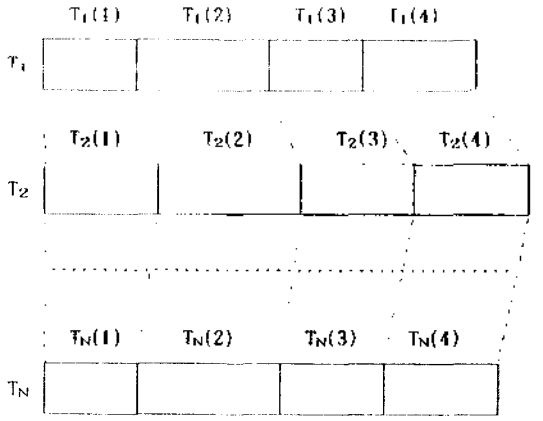
Let's define that word-pattern for training data is T_1, T_2, T_3,....T_n,...,T_N, and distance between word pattern and word model calculated by algorithm is $DIS(T_n, M)$. Then,

$$D_{min} = \sum_{n=1}^{N} DIS(T_n, M) \tag{15}$$

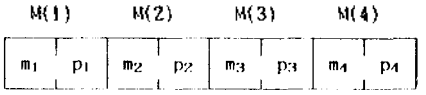
Where, D_{min} generates the word model M to be minimum. In order to so, we need to consider two methods. The first, as it is shown Fig. 2, we assign word pattern in each section. The second, it is method that we make out typical feature vector in pattern of the each section to have been assignment.

We optimize word model in the same way as flowchart of Fig. 3.

First of all, initial word model calculates typical



(a) Training data



(b) DMS model

Fig 2. An example of section assignment of training data.

feature vector in frames which is assigned in each section, in equil-length into the time-axis through training data. Then, duration time information is computed.

Thereafter, we change a boundary line of section by matching between training data and word model using section division algorithm. As an example of the matching, fig. 4 is shown.

Namely, when it becomes $e(0)=0$, $e(1)=a$, $e(2)=6,\ldots,e(J)=I$, section 1 is assigned from 1 to a frame, section 2 is assigned from $a+1$ to b frame,...,etc. Thereafter, word models update by calculation of centroid in feature vectors of new assigned frames. At the same time, time information of word model is registered for the rate that divide final frame number in each section by final total frame number.

After updating by repetition till convergence of word model by algorithm, time information is registered. But, typical feature vector has different

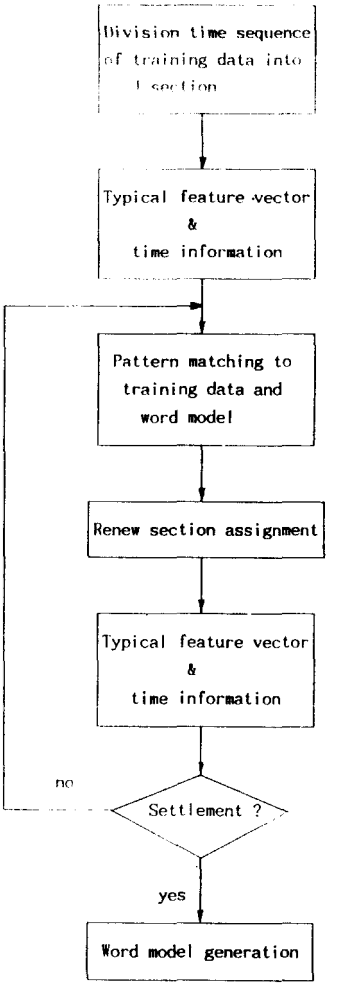


Fig 3. Method of word model generation

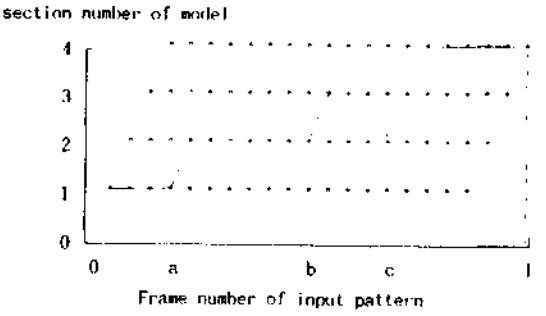


Fig 4. Example of matching path($J=4$).

models according to recognition method. Recognition by DP matching method registered typical feature vector of word model soon after convergence. But, in case of recognition by VQ, we register two typical feature vectors as a model make two cluster centers per section after convergence.

3. Restriction in model generation

1) Local path constraints

Accumulated distance according to optimal path from point(i, j) equal to eq. (7), and local path constraints is shown Fig. 5.

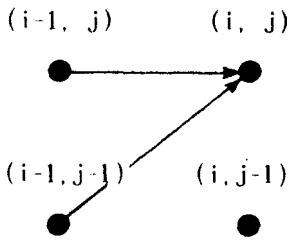


Fig 5. Local path constraint.

2) Global range constraints.

By this constraints, point(i, j) of optimal warping path is constrained and can reduce computation time. In this paper, we constraint global path as Fig. 6.

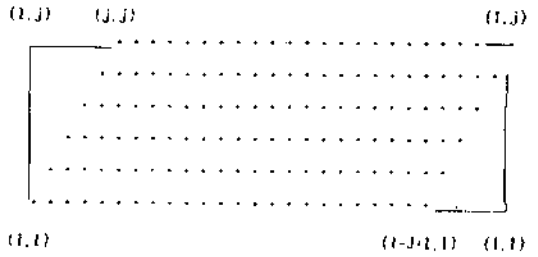


Fig 6. Example of Global range constraint.

Ⅲ. Experiment result

In speech recognition by DMS Model, Korean 146 DDD area names is chosen as the recognition vocabulary, and typical feature vectors and duration time informations of DMS model are made by twice among words spoken three times by three men respectively and we recognize with the remained data which is not used by training data of each speaker.

1. construction for recognition system

Fig. 7 represents the speech recognition system according to proposed DMS model, and all data have sampling frequency as 8KHz, LPF as 3.5KHz, feature parameter as 12th LPC cepstrum coefficient.

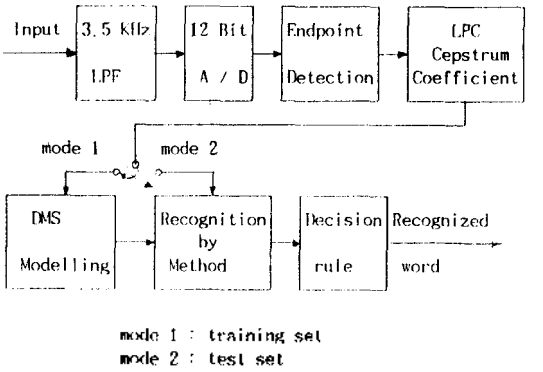


Fig 7. Block diagram of recognition system by DMS model

2. Recognition experiment according to each method

(1) VQ Recognition using DMS model(DMS/VQ)

In the case of word model of 6 section using two codewords in each section, the weight which shows the best recognition rate is 0.6(see table 1). Table 1 and table 2 indicate the best recognition rate when section is 6 and weight is 0.6

Table 1. Recognition result by DMS/VQ of state 6

(Unit : %)

Speaker \ Weight	0.0	0.4	0.5	0.6	0.7	0.8
Speaker A	87.0	97.3	98.6	99.3	98.6	96.5
Speaker B	84.2	95.2	94.5	95.9	97.3	95.9
Speaker C	78.1	92.5	96.6	95.2	93.8	93.2
Total	83.1	95.0	96.6	96.8	96.6	95.2

Table 2. Recognition result by DMS/VQ of weight 0.6

(Unit : %)

Speaker \ Section	4	5	6	7	8
Speaker A	97.9	98.6	99.3	99.3	98.6
Speaker B	93.2	93.8	95.9	96.6	94.5
Speaker C	92.5	92.5	95.2	93.1	92.5
Total	94.5	95.0	96.8	96.3	95.2

(2) DP recognition using DMS model(DMS/DP)

In section 6, the best recognition rate is obtained when weight is 0.3. We fix weight as 0.3 and experiment from section 6 to section 16 by increment of two step. Then, the best recognition rate is obtained when section is 12.

Table 3. Recognition result by DMS/DP of section 12

(Unit : %)

Speaker \ Weight	0.0	0.1	0.2	0.3	0.4	0.5
Speaker A	88.4	97.3	99.3	99.3	97.9	98.6
Speaker B	88.4	94.5	94.5	95.2	94.5	93.2
Speaker C	91.8	93.2	93.1	93.1	91.8	91.1
Total	89.5	95.0	95.6	95.8	94.7	94.3

Table 4. Recognition result by DMS/DP of weight 0.3

(Unit : %)

Speaker \ Section	8	10	12	14	16
Speaker A	98.6	99.3	99.3	98.6	98.6
Speaker B	94.5	93.8	95.2	95.2	95.2
Speaker C	91.1	93.1	93.1	93.1	91.8
Total	94.7	95.4	95.8	95.6	95.2

(3) The recognition by MSVQ

In this experiment, we compared with some MSVQ experiments. These MSVQ codebooks contains two codewords in each section.

Experiments according to section number of MSVQ is shown in table 5.

Table 5. Recognition rate by MSVQ according to section number

(Unit : %)

Section	4 MSVQ	6 MSVQ	8 MSVQ	10 MSVQ	Overlapped 8 MSVQ
Rate	53.4	74.0	81.5	69.2	89.3

(4) The recognition by DTW

In this experiment, we compared with conventional method by DTW. Template generation by conventional DTW in this experiment is selected one reference pattern in each word.

Recognition rate by generating reference pattern is shown in table 6.

Table 6. Recognition rate according to selection of reference pattern

(Unit : %)

Method	Selection of random token	Selection of a token after clustering	Selection by clustering method[11]
Rate	78.5	78.8	95.0

(5) The recognition by HMM

In this experiment, we gave codewords 128 and number of states 8. And, recognition rate by HMM is 91.6%.

3. All-around experiment results

In this paper, we compared recognition rate with DP, MSVQ, HMM and recognition systems using DMS model, and we proved usability in proposed system by comparing memory size and processing speech which demanded in each system.

Therefore, the best recognition rate according

Table 7. Recognition result[7]

(Unit : %)

Method	Recog. rate of DMS model		Recognition rate of MSVQ	Recognition rate of DTW	Recognition rate of HMM
	DMS/VQ	DMS/DP			
Rate	96.8	95.8	89.3	95.0	91.6

Table 8. Requirement of memory and computation time[7]

Method / Classification	DMS model		MSVQ	DTW	HMM
	DMS/VQ	DMS/DP			
Memory unit : number of vectors	23,652	24,528	30,368	75,920	154,672
Computation unit : multiply	304,556	762,120	304,848	1,518,400	206,720
unit : log					140,160

to each method are shown in table 7, and memory size and processing speed which are needed in each recognition method are shown in table 8.

4. Consideration

When section number is 6 and weight is 0.6 [show table 1 and table 2], the best recognition accuracy is given in DMS/VQ method. Korean DDD area names are all consisted of 2-syllable except Eui-Jung-Bu(3-syllable), and each syllable is consisted of maximum 3 phonemes. DMS/VQ is designed for two codewords(feature vectors) and time information per section. Therefore, because VQ using DMS model of section 6 can search features for transition between phonemes as well as features of phonemes, the best recognition accuracy is given.

When section number is 12[show table 3 and table 4], the best recognition accuracy is given in DMS/DP. The reason is like the case of DMS/VQ.

Recognition accuracy of conventional DP matching method[show table 5] is worse than those of methods using DMS mdoels.

A seperate MSVQ codebook[show table 6] is designed for each word in the recognition vacabulary by dividing the words in the codebook's training sequence into equi-length sections. For that reason, MSVQ makes up a cluster among the quite different many features or different cluster in continuous similiar features, because MSVQ is divided equi-length section.

Performance of each system is shown in table 7 and memory and computational requirement of each system is shown in table 8. From table 7 and table 8, DMS/VQ requires the same size memory and processing time as MSVQ and is the best recognition accuracy in three methods.

Ⅳ. Conclusion

In this paper, we performed speech recognition of independent speaker by DMS model, with Korean 146 DDD area names, and compared with recognition experiments by conventional DP, HMM and MSVQ. DMS Model proposed in this paper extracted only a representative information after making a section by feature vector of continuous similar characteristics. At that moment, it is the model which extracts typical feature vector by

clustering method and time information by backtracking method.

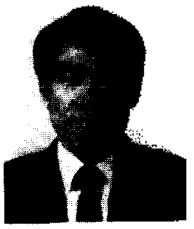
In case of the conventional DP, only a feature vector is used to a distance calculation. But, DMS model is added time information. Therefore, we can use duration time information also when input pattern comes. So in case of DMS/DP it is possible to recognize with some typical feature vectors only. And in case of DMS/VQ, it is possible to divide section of test speech by duration time information.

In case of DMS/VQ proposed in this paper, weight 0.6 and section 6 is the best when we used time imformation as scalar distance value. As an result of comparison experiments, recognition accuracy of conventional DP pattern matching is about 95.0% and recognition accuracy of conventional MSVQ is about 89.3%. In DMS model recognition accuracy of DMS/DP is about 95.8% and recognition accuracy of DMS/VQ is about 96.8%. Therefore, because it extracts typical feature vector of dynamic section under time information, it is proved that it is the best recognition system to use DMS/VQ recognition method proposed in this paper.

References

1. Hiroaki Sakoe and Seibi 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.
2. R. M. Gray, "Vector Quantization," IEEE ASSP Magazine, Vol. 1, pp4-29 Apr. 1984.
3. Y. Linde, A. Buzo, and R. M. Gray "An algorithm of Vector Quantizer Design," IEEE Trans. Comman, Vol. COM-28 pp.84-95, Jan 1980.
4. D.K. Burton, J.E. Shore, J.T. Buck "Isolated-Word Speech Recognition Using Multisection Vector Quantization Codebooks" IEEE Trans. of Acoustics, Speech, Signal Processing Vol ASSP-33. No 4. August 1985.
5. Tae Ock Ann, and Sun Hyub Kim, "An Automatic Speech Recognition of Computer Using Time Sequential Vector Quantization," KITE. Vol. 27, No. 7, pp 157-165, July 1990.
6. L. R. Rabiner, B.H.Juang, "An Introduction to Hidden Markov Models" IEEE ASSP MAGAZINE JAN. 1986.
7. L. R. Rabiner, S. E. Levinson, M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-independent, isolated Word Recognition," Bell System Technical Journal, Vol. 62, No. 4, April 1983.
8. Kai-Fu Lee, Automatic Speech Recognition : The Development of the SPHINX System, Kluwer Academic Publishers, 1989.
9. Tae Ock Ann etc, "Korean Speech Recognition using DHMM," Acoustic Society of Korea, Vol. 10, No.1, pp 52-60, Feb. 1991.
10. J. T. Tou, R. C. Gonzalez, Pattern recognition Principles, Addison-Wesley Publishing Company, Inc, 1974.
11. S. E. Levinson, L. R. Rabiner, A. E. Rosenberg and J. E. Wilpon, "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Techniques for Isolated Word Recognition," IEEE Trans. on ASSP Vol. 27, No. 2, PP. 134-141, APR. 1979.
12. Shilano, K., Kohda, M., "On the LPC Distance Measures for Vowel Recognition in Continuous Utterance," Institute of Electrical and Communication Engineers of Japen, Trans. on D, J63-d, May 1980.

▲ Tae Ock Ann



Tae Ock Ann received the B. E degree in material engineering from University of Wool San, Korea, in 1981, and the M.E., Ph.D. degree in computer engineering from University of Kwangwoon, Seoul, Korea, in 1987 and 1991. He is currently a professor in the CHONBUK SANUP University. His research field is speech recognition, computer vision, AI, neural network, and robotics.

▲ Yong Kyu Byun

Yong Kyu Byun was born in Yangju, Korea, on Nov. 10, 1933. He received the B.E. degree in electrical engineering from University of Yonsei, Seoul, Korea, in 1958, the M. E. degree in electronics engineering from University of Dongkuk, Seoul, Korea, in 1979, and the Ph.D. degree in electronics engineering from University of Kwangwoon, Seoul, Korea, 1991. He is currently a professor in the Seoul Polytechnique University. His research subjects of interest include the digital signal processing, speech recognition and synthesis, computer communication.