A Study on Isolated Word Recognition for Implementation of Real-Time Voice Dialing System

실시간 음성 다이얼링 시스템 구현을 위한 단독어 인식에 관한 연구

Hang-Seop Lee*, Jin-Woo Hong*, Gang-Sung Lee*, Soon-Hyob Kim*

이 항 섭*, 홍 진 우*, 이 강 성*, 김 순 협*

Abstract

This paper describes speaker dependent isolated word recognition system for the development of real time voice dialing system. We used DMS model as recognition method beacuse it requires small memory for models and less computation time for matching. So we can get response in 3 seconds after utterance for 50 words vocabulary selected from department names of a university. The performance of the system showed 98 percent recognition rate for 22 sections and for 0.6 time duration weight of DMS model.

요 약

본 논문은 실시간 음성 다이얼링 시스템 구현을 위한 화자종속의 단독어 인식에 대하여 기술하였다. 인식을 위한 모델 작성은 적은 메로리에 계산 시간이 적게 걸리는 DMS 모델을 사용하였다. 인식 대상어는 대학교내의 50개 부서명을 선택하였고, 발성후 3초내에 인식결과를 얻을수 있었다. 시스템은 구간수 22에서 가중치 0.6의 DMS 모델을 표준패턴으로 사용하였을때 98%의 성능을 나타냈다.

Ⅰ. Introduction

One demands more convenient method for dialing. The virtue of development for voice dialing system is recognized because it has more advantages than conventional dialing system. It is guaranteed to be a commercial success if it can be made to work well with low price hardware.[1]

Few institutes have developed workable systems, but all of those required large computation by using dynamic time warping(DTW) or hidden Markov model(HMM).

Although DTW is very effective method for speech recognition, models used by DTW are not efficient because it represents the model as a whole part of speech patterns and it has limitation which can not abstract redundant informations. This unefficient fact of DTW is covered by using HMM.

*Dept. of Computer Eng., KwangWoon University.

본 연구는 1989년 부터 1992년 까지의 한국과학재단 "목적기초연구"의 지원에 의한 것임.

HMM constructs models with the states each of which represents section of similar characteristics. Each state has probabilities of observing symbols. HMM needs many training data. Although the problem of insufficient trainging data can be compansated with several smoothing methods, it shows more unstable results than other method. And computational complexity is high relatively.[5]

Dynamic multisection(DMS) has been proposed to overcom inefficiency of redundant informations and computational complexity of other methods.

The advantages of DMS model are as follows :

1. It requires small memory for references.
2. It requires less computational complexity than DTW.
3. It showed better performance (96.8%) than DTW(93.4%), MSVQ(89.3%) and HMM(91.6%) for 146 DDD area words.

DMS / VQ and DMS / S.S, which are variations of DMS, are used for the recognition methods of voice dialing system. 50-department names in a Kwangwoon university are selected for test vocabulary.

Ⅱ. DMS model

1. Concept[5]

A speech pattern can be considered as a sequence of phonetic symbols which have stable spectral characteristic. A DMS model consists of several sections, dynamically divided from training patters according to spectral characteristics. The boundaries which take partitions of sections are determined dynamically to minimize pattern matching distance between a DMS model and test patterns. A section has two informations, mean vector and time duartion. Mean vector is representative expression of a section which is assumed to have stable spectral characteristic, and time duration is a length of a section. So it has no redundant spectral informations like DTW models and has time duration information which requires simple computation.

2. Section segmentation algorithm

DMS model M of each word consists of section information $M_j(1 \le j \le J)$, and M_j consists of mean feature vector m_j and time duartion p_j.

Initial word model is constructed by dividing all training data into equal length J sections and by calculating centroid vectors with all vectors of same sections and time duration, mean length of each section.

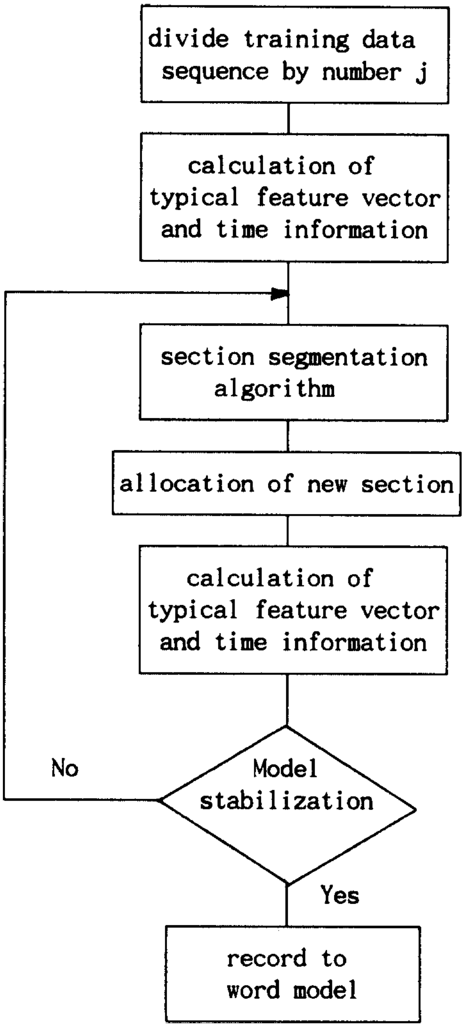

divide training data sequence by number j
calculation of typical feature vector and time information
section segmentation algorithm
allocation of new section
calculation of typical feature vector and time information
Model stabilization
No
Yes
record to word model

Fig 1. The method of word model generation.

After constructing initial model, pattern matching similar to DP matching[6] is performed between each word model Mj and all the training patterns of word j, then section boundaries are determined again by backtracking.

Mean feature vector and time duration of each section are calculated again using the newly determined boundary informations and then update DMS model. Time duration is average of frames for each state. Now DMS model is updated once.

As Fig 1., pattern matching and model reconstruction are repeated until DMS model converges. If accumulated distance of pattern matching between newly constructed model and test patterns is greater than distance of old model, model is converged and processing is stopped.

Local accumulated distance D from start frame up to ith frame of test pattern and jth section of DMS model includes distance P of time duration.

$$D(i, j)=d_v(t_i, m_j)+\min\left[\begin{array}{l} D(i\text{-}1, j) \\ \qquad (1<i\leq I,\ 1<j\leq J) \\ D(i\text{-}1, j\text{-}1)+P(j\text{-}1) \end{array}\right. \quad (1)$$

P(j) is the distance between the time duration of jth section for a word model and of the time to ith frame from start.

$$P(j)=W * ds(e(j),i) \quad (2)$$

$$ds(e(j), i)=|p(j) * I\text{-}i| \quad (3)$$

where, P(j) is the rate expressed in equation (4), sum of end frame position correspond to j sections of the models per length of total frame of test data.

$$p(j)=\sum_{m=1}^{M} e_m(j) \,/\, \sum_{m=1}^{M} I_m \quad (4)$$

And W is the weight constant to make balance with local spectral distance, and it is considerable value affecting to the recognition performance.

Ⅲ. SPEECH RECOGNITION USING DMS MODEL

In this study, we have performed speech recognition experiment using DMS model with following two methods. The first is speech recognition using DMS /VQ, and the second is speech recognition using DMS /SS which is performed by calculating distances from section to section.

3.1 Recognition using DMS/ VQ method

DMS /VQ is the recognition method that divides the test pattern to sections and recognizes by comparing frames of each section and codewords of the reference pattern. Time distance is considered also by calculating difference of time duration of reference pattern from the rate of each section per whole pattern length.

In this study, we have experimented according to the number of codewords per a section. One is DMS /VQ4 that selects two codewords per each section and the other is DMS /VQ3 that selects one codeword per each section.

DMS /VQ4 method performs recognition by accumulating minimum distances from the frame of test pattern to codewords in two contiguous sections of the reference pattern for each frame. We experimented recognition for each model which has sections from 6 to 12.

Whereas, DMS /VQ3 method performs the recognition by comparing distances between three codeword of a reference pattern and a frame of a test pattern and accumulating minimum distances.

The algorithm used for the recognition is shown in Fig. 2

The spoken word is determined by finding a model whose accumulated distance is minimum after matching the test pattern with all of the reference patterns

3.2 Speech Recognition using DMS/ SS method

In DMS /SS(Section to Section), test pattern is divided into sections according to section time duration information of the reference pattern to

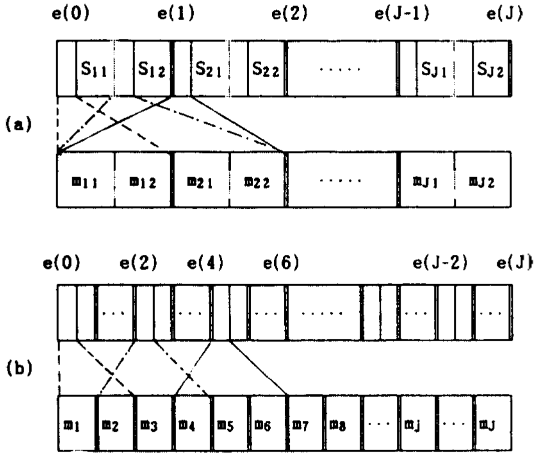


Fig 2. Distance comparison for DMS / VQ.
(a)DMS / VQ 4. (b)DMS / VQ 3.

be matched, and one centroid vector per a section is calculated.

It reduces computation time significantly not because each frame vector of test pattern is matched to codewords of reference patters, but because centroid vector of test pattern is compared with codewords of reference patterns. It is very efficient method if it shows reasonable performance.

We have experimented also in two ways, DMS / SS 4 and DMS / SS 3. DMS / SS 4 compares one codeword of test pattern with four codewords of the reference pattern. DMS / SS 3 compares one codeword of test pattern with three code words of the reference pattern.

DMS / SS 4 method is the same with DMS / VQ 4 method and DMS / SS 3 method is the same with DMS / VQ 3 method in view of generating models and dividing sections of test pattern except comparing vectors. DMS / VQ compares each frame of test pattern while DMS / SS compares each centroid vector of test pattern with reference pattern.

The algorithm of DMS / SS 3 method is shown below.

(DMS / SS 3 recognition algorithm)

Step 1. Initialization

(section division by time information)

repeat $0 \leq j \leq J$

$$e(j) = p(j) * I \qquad (5)$$

Step 2. if i=1

$$DIS(T,M) = \min \begin{bmatrix} d_v(c_1,\ m_i) \\ d_v(c_1,\ m_{i+1}) \end{bmatrix} \qquad (6)$$

Step 3. repeat $2 \leq i \leq I\text{-}1$

$$DIS(T,M) = DIS(T,M) + \min \begin{bmatrix} d_v(c_i, m_{i-1}) \\ d_v(c_i, m_i) \\ d_v(c_i, m_{i+1}) \end{bmatrix} \qquad (7)$$

Step 4. if i=I

$$DIS(T,M) = DIS(T,M) + \min \begin{bmatrix} d_v(c_I,\ m_{I-1}) \\ d_v(c_I,\ m_I) \end{bmatrix} \qquad (8)$$

This algorithm can be represented as follows.

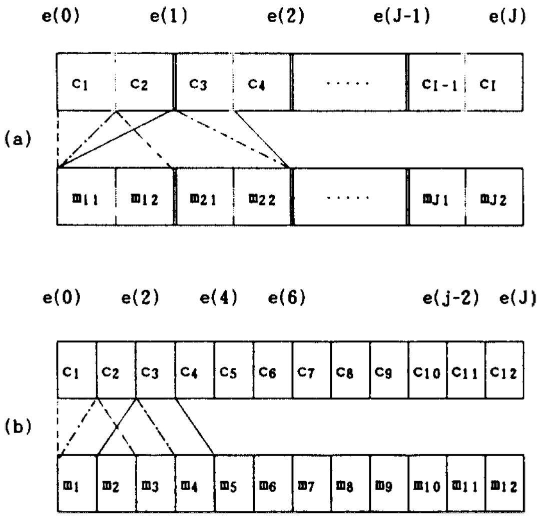


Fig 3. Distance comparison for DMS / SS.
(a)DMS / SS 4 method. (b)DMS / SS 3 method

Ⅳ. Experiments

4.1 Recognition system

1. Vocabulary and feature extraction

50 department names in a university are selected as test vocabulary. Three men participated in recording. Test data are recorded three times per each word and we got total 450 data. 300 data among 450 are used for training and the other 150 data for testing performance.

Analog speech signal is passed 3.4KHz LPF, sampled with 8KHz and converted to 16 bit. After end-point detection, we got 12th LPC cepstrum coefficients[9][10].

2. Organization of recognition system

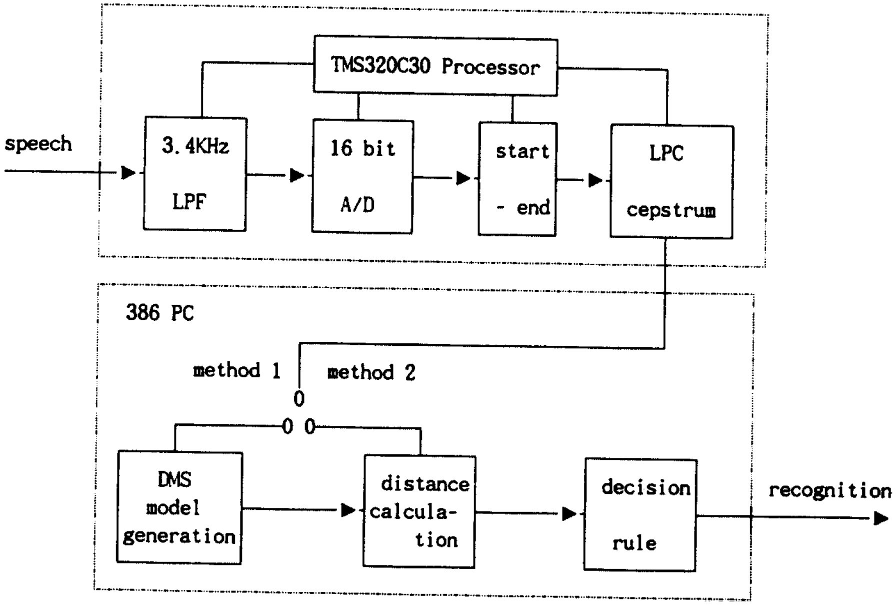
For recognition experiment, we constructed recognition system using TMS320C30 system board which is designed specially for DSP applications. Fig.4 is the block diagram of recognition system.

3. Features of TMS320C30[3][4][11]

TMS320C30 digital signal processor has large on-chip memories, concurrent DMA controller, and instruction cache. It can perform four level pipeline, multiply integer and floating point numbers in single cycle, and executes up to 33 MFLOPS.

4.2 Experiments of DMS/ VQ method

The result of experiment by DMS/VQ is shown in table 1 and 2. Table 1 is the result of experiment for the number of sections from 6 to 12 for DMS/VQ 4 and from 12 to 24 increasing by 2 for DMS/VQ 3. Time duration weight 0.6 is applied after preliminary experiments. DMS/VQ 3 method showed better results than DMS/VQ 4 for all number of sections. The best result was found in section length 16. Table 2 shows the results of experiment for the time duration weight from 0.1 to 1.0 increasing by 0.1 with section length 16. It shows the best recognition result 98.6% at weight 0.6 and section length 16.



method 1 : training data.
method 2 : test data.

Fig 4. Block diagram of recognition system.

Table 1. RECOGNITION RESULT FOR DMS / VQ 4 AND DMS / VQ 3 WITH WEIGHT 0.6.

weight=0.6 (unit : %)

method	section	6	7	8	9	10	11	12
DMS / VQ 4	speaker A	94	96	98	98	98	100	100
	speaker B	94	90	94	96	96	98	94
	speaker C	94	92	94	94	94	94	96
	total	94	92.6	95.3	96	96	97.3	96.6
method	section	12	14	16	18	20	22	24
DMS / VQ 3	speaker A	96	100	100	100	100	100	98
	speaker B	98	98	98	92	94	98	96
	speaker C	94	96	98	94	98	96	98
	total	96	98	98.6	95.3	97.3	98	97.3

Table 2. RECOGNITION RESULT FOR DMS / VQ 3 OF 16 SECTIONS.

section=16 (unit : %)

weight speaker	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
speaker A	96	98	98	100	100	100	100	98	98	96
speaker B	94	98	100	96	98	98	94	98	98	98
speaker C	96	96	94	94	96	98	98	98	98	96
total	95.3	97.3	97.3	96.6	98	98.6	97.3	98	98	96.6

Table 3. RECOGNITION RESULT FOR DMS / VQ 4 AND DMS / VQ 3 WITH WEIGHT 0.6.

weight=0.6 (unit : %)

method	section	6	7	8	9	10	11	12
DMS / SS 4	speaker A	80	88	94	96	98	96	98
	speaker B	74	84	94	92	92	98	96
	speaker C	80	92	90	94	98	98	98
	total	78	88	92.6	94	96	97.3	97.3
method	section	12	14	16	18	20	22	24
DMS / SS 3	speaker A	86	94	94	94	96	98	98
	speaker B	86	84	88	94	94	96	96
	speaker C	92	96	96	96	98	100	100
	total	88	91.3	92.6	94.6	96	98	98

4.3 Experiments of DMS/ SS method

We fixed time duration weight to 0.6 as the result of experiments for DMS / VQ, and experimented changing the number of sections from 12 to 24. For comparison with DMS / VQ, we selected the number of sections equivalant to the number of sections in DMS / VQ experiments.

Table 3. shows that DMS / SS 3 is better than DMS / SS 4 for all number of sections. The best recognition rate is 98% for 22 and 24 sections in DMS / SS 3. For examine the effect of change for the time duration weight we experimented recog-

Table 4. RECOGNITION RESULT FOR DMS / VQ 3 OF 22 SECTIONS.

section=22										(unit : %)
weight speaker	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
speaker A	100	98	96	96	96	98	98	96	96	96
speaker B	96	94	98	96	92	96	98	94	94	96
speaker C	96	98	96	98	98	100	98	96	96	96
total rate	97.3	96.6	96.6	96.6	95.3	98	98	95.3	95.3	96

Table 5. RECOGNITION RESULT FOR DMS / VQ 3 OF 20 SECTIONS.

section=20										(unit : %)
weight speaker	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
speaker A	96	100	96	92	96	96	94	98	96	96
speaker B	94	92	92	94	92	94	92	94	92	94
speaker C	96	98	98	96	98	98	98	96	96	96
total rate	95.3	96.6	95.3	94	95.3	96	94.6	96	94.6	95.3

nition by changing time duration weight from 0.1 to 1.0 for 22 and 20 sections. In case of 22 sections it showed the best recognition rate at weight 0.6 and 0.7. And in case of 20, it showed the better recognition rate at weight 0.6 than 0.7. As the result of experiments, we can find that although time duration weight affects to the recognition rate but more important factor is the number of sections than time duration weight.

V. Implementation of Voice dialing system

5.1 Concept of Voice dialing system[1][2][12]

Voice dialing system adds the speech recognition function to the conventional telephone. A user speaks the desired phone number or the name whom he wants to speak to. Then the system recognizes the voice and dials phone number using modem. It is new and very convinient dialing method for the following reasons.

1.It is not necessary to search or remember the phone numbers.
(It is very convenient for children, the old and the handicapped)

2. It can reduce mistakes caused by pressing buttons.

3. It can be extended to the various telephone services by entering informations using voice.

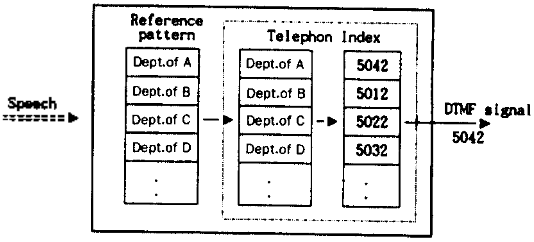


Fig 5. Concept diagram of voice dialing system.

The processing time is important factor to the users. Long processing time for the recognition of voiced word is not allowed for most of users which do not have patience enough. Large vocabulary is not proper to the real time recognition and continuous speech recognition also takes too long time to accept results in real-time because it has many calculations.

Therefore, it is necessary to set some restrictions. We selected 50 words for testing system performance.

5.2 Implementation

As the result of experiments on PC, DMS / VQ method showed better recognition rate than DMS / SS method. For voice dialing system, DMS / SS 3 seemed more reasonable method. DMS / SS 3 showed very small degradation in recognition rate compared to DMS / VQ but computation time is only about half. We selected the number of sections as 22 and time duration weight as 0.6

For implementation of voice dialing system in real-time, we used TMS320C30 system board.

5.3 Configuration of system[13][14]

The system is constructed by PC-386, modem and TMS320C30 board. Reference patters are downloaded from PC to TMS320C30. Input speech signal from microphone is pre-processed on TMS320C30 and changed to the LPC cepstrum coefficients as the feature vectors. Test pattern is compared with reference patterns on TMS320C30 board. The model which has smallest accumulated distance is selected as recognized model. But three candidates are displayed on screen so that the user can select the candidate number if the first candidate is not correct.

Whole organization of the system is represented in Fig.6.

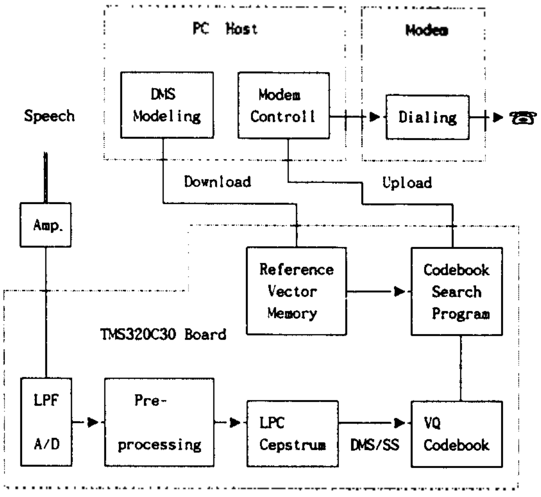


Fig. 6 Voice Dialing System Organization.

Ⅵ. Conclusion

In this paper, we performed a study on the isolated word recognition for implementation of real-time voice dialing system. We have implemented real-time voice dialing system using 386 PC, TMS320C30 system board and modem, and used DMS / SS 3 as recognition algorithm.

In the isolated word recognition experiment, DMS / VQ 3 method showed a 98.6% recognition performance and DMS / SS 3 showed a 98% but we used DMS / SS 3 as recognition algorithm for voice dialing system requiring real-time because it is about twice faster than DMS / VQ method.

We have previously modeled DMS models on PC and downloaded it to TMS320C30 system board and performed all process on it for real-time processing from speech input to recognition and received the result on PC and dialed phonenumber using modem.

Now this system took less than three seconds from end of speech input to recognition. If we utilize enough parallelism of TMS320C30 processor and improve interface with PC, all process will progress within one or two second.

References

1. M. Immendorfer, "Voice Dialer," Electrical Communication, Vol.59, No.3, 1985.
2. A. Fukui, Y. Fujihashi, F. Nakagawa, "SIGNAL PROCESSOR APPLICATION TO VOICE DIALING EQUIPMENT," ICASSP 86, TOKYO 7.8.1.

3. Kun-Shan. L, G.A. Frantz, R. Simar, "The TMS320 Family of Digital Signal Processors," PROCEEDING OF THE IEEE, Vol.75, No.9, pp.1143-1159, September 1989.

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

5. Byun y.k. "A Study on Isolated Word Recognition Using DMS Model," Ph.D. Paper, Kwangwoon nuiv. graduate school, 1991, 2.

6. H. Sakoe, 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.

7. L.R. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, JAN. 1986.

8. 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.

9. L.R. Rabiner, K.C. Pan and F.K. Soong, "On the performance of isolated word speech recognizers using vector quantization and temporal energy contours," AT&T Bell Lab. Tech. J., Vol. 63, pp. 1245-1260, Sep. 1984.

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

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

12. Ikuo Hongo, Hiroshi Kanamura 외, "A Study of the Voice Dialing System," SE79-83.

13. D. Chase, A. Gersho, "REAL-TIME VQ CODEBOOK GENERATION HARDWARE FOR SPEECH PROCESSING," ICASSP 88, Vol 3. pp. 1730-1733, 1988.

14. J.B. Attili, M. Savic, J.P. Campbell,Jr., "A TMS320C20-BASED REAL TIME, TEXT-INDEPENDENT, AUTOMATIC SPEAKER VERIFICATION SYSTEM," ICASSP 88, Vol 1. pp. 599-602, 1988.

▲Hang Seop Lee

was born in Seoul, Korea, on Mar. 15, 1967. He received the B.S., M.S. degree in computer engineering dapartment from the Kwang-Woon University, Seoul, in 1990 and 1992, respectively.

Since February 1992 he has been with the Signal Processing Section in ETRI, as a Junior Researcher.

His research interests include the speech recognition and speech synthesis.

▲Jin Woo Hong

was born in YuJoo, Korea, on April 15, 1959. He received the B.S., M.S. degree in electronics engineering dept. from the Kwang-Woon University, Seoul, in 1982 and 1984, respectively.

Since 1984 he has been with the Transmission System Section and the ISDN Technology Department in ETRI, where he is now a senior member of technical staff. He is currently working toward the Ph.D. degree at Kwang-Woon University from March 1990. His research interests include the ISDN, speech quality assessment, and speech synthesis.

▲Gang Sung Lee

was born is Seoul, Korea, on January 15, 1964. He received the B.S., M.S. degree in Computer engineering department from the Kwang-Woon University, Seoul, in 1986 and 1988, respectively. He is currently working toward the Ph.D. degree at Kwang-Woon University from March 1988. His research interests include the speech recongintion.

▲Soon Hyob Kim

professor, Dept. of Computer Engineering, Kwang-Woon Univ.(Vol.10, 10, No.4, 1991)