HMM with Global Path Constraint in Viterbi Decoding fcr Insolated Word Recognition

전체 경로 제한 조건을 갖는 HMM을 이용한 단독음 인식

Weon Goo Kim*, Dong Soon Ahn*, Dae Hee Youn*
김 원 구*, 안 동 순*, 윤 대 희*

ABSTRACT

Hidden Markov Models (HMM's) with explicit state duration density (HMM/SD) can represent the time-varying characteristics of speech signals more accurately. However, such an advantage is reduced in relatively smooth state duration densities or long bounded duration. To solve this problem, we propose HMM's with global path constraint (HMM/GPC) where the transition between states occur only within prescribed time slots. HMM/GPC explicitly limits state durations and accurately describes the temproal structure of speech simply and efficiently. HMM's formed by combining HMM/GPC with HMM/SD are also presented (HMM/SD+GPC) and performances are compared. HMM/GPC can be implemented with slight modifications to the conventional Viterbi algorithm. HMM/GPC and HMM/SD+GPC not only show superior performance than the conventional HMM and HMM/SD but also require much less computation. In the speaker independent isolated word recognition experiments, the minimum recognition error rate of HMM/GPC(1.6%) is 1.1% lower than the conventional HMM's and the required computation decreased about 57%.

요 약

상태 지속 밀도를 사용하는 Hidden Markov Models(HMM/SD)은 음성 신호의 시간적인 변화를 보다 명확하게 나타낼 수 있다. 그러나 상태 지속 빈도가 완만하거나 제한된 상태가 길면 이러한 장점은 감소된다. 이러한 문제점을 해결하기 위하여, 본 논문에서는 상태간의 천이가 특정한 시간 구간에서만 발생하도록 하는 전체 경로 제한 조건을 갖는 HMM/GPC를 제한한다. HMM/GPC는 상태 지속을 제한하고 음성 신호의 시간적 변화를 단순하고 효과적으로 표현할 수 있다. 또한 HMM/SD와 HMM/GPC를 결합한 새로운 형태의 HMM/SD+GPC를 제안하고 성능을 비교하였다. HMM/GPC는 기존 Viterbi 알고리즘을 약간 수정하여 구현될 수 있다. HMM/GPC와 HMM/SD+GPC는 기존 HMM과 HMM/SD에 비하여 우수한 성능을 보일 뿐만아니라 계산량도 매우 작다. 화자독립 단독음 인식 실험에서, HMM/GPC(1.6%)의 최소 오차는 기존 HMM보다 1.1% 낮았고 계산량도 57% 감소하였다.

*연세대 전자공학과

접수일자 : 1994년 1월 4일

Ⅰ. Introduction

Methods used in speech recognition such as Dynamic Time Warping(DTW)[1] and Hidden Markov Models (HMM's)[2] are similar in that they try to model the temporal structure of speech signals. The characteristics of a speech signal are determined by the shape of the vocal tract as it changes over time and hence by the generated time-varying spectrum. Such changes represent the duration and relative change in the acoustical characteristics (i.e. spectrum, etc.) of the speech signal. As a result, the performance of a speech recognition system largely rests on the accuracy in representing such time-varying traits.

HMM's with state duration(HMM/SD) uses state duration model to represent temporal characteristics more specifically. This shows improved performance over the conventional HMM[4], but requires more memory and computation. Accordingly, nonparametric state duration densities were replaced with parametric forms such as Poisson[3], Gamma[4], Gaussian[5], and Uniform [2] densities. However, the assumed state duration densities could not be guaranteed to provide accurate representations. Furthermore, for a smooth state duration density, distorted matches become possible within the Viteribi algorithm[6]. To overcome such problems, the bounded state duration density (HMM/BSD) with minimum and maximum state duration bounds has been proposed[6]. Nevertheless, this method still produces inaccurate matches for long state durations, and the required computation remains much greater than the conventional Vierbi algorithm.

An effective sub-optimal solution is to estimate HMM parameters and the state duration densities separately, then to use both in the recognition process[6][7]. In this case, the state duration density is obtained directly from the trained HMM parameters and the training data. Even so, this method requires 15 to 20 times the computation of the conventional Viterbi algorithm. Such problems led Rabiner et al[7] to propose using a post processor to include state duration information in the recognition process. This method finds the optimal state sequence by the Viterbi algorithm and backtraking, obtains the duration for each state, and adds the state duration probability to P^* which represents the similarity to the input data. However, this method requires storage of D state duration densities for each word model, where D is the maximum duration of a state. In addition, the final probability is obtained by finding the optimal state sequence by the Viterbi algorithm and backtracking to adjust the state duration density, leading to redundant computation.

To overcome such problems, a globally path constrained HMM's (HMM/GPC) that simply and effectively represents the temporal structure of speech signals is proposed in this paper. During the recognition phase in HMM/GPC, state transitions are restricted to prescribed time slots, leading to fewer mismatches and increased recognition rates. Moreover, the proposed method can be implemented by slight modifications to the Viterbi algorithm and results in less computation than in conventional Viterbi decoding.

Further, the global path constraint is incorporated into the HMM/SD and HMM/BSD for improved performance and reduced computation. In gobally path constrained HMM/SD and HMM /BSD, state duration is restricted by the state duration density and also limited to specified time slots. Therefore, as in HMM/GPC, the global path constraint allows fewer mismatches and greatly reduces computation in the HMM/SD and HMM/BSD.

This paper is presented as follows. The HMM/ GPC with global path constraint is presented in Chapter II. Experimental results are given in Chapter III, and the conclusion is given in Chapter IV.

II. HMM with Global Path Constraint

In the proposed HMM/GPC, two transition time limiting parameters, l_i and u_i, are required in addition to conventional HMM parameters, state transition probabilites a_{ij} and observation probabilies $b_j(O_t)$. The two parameters restrict state transitions to specified time slots during recognition. These parameters are estimated during training to modify the defintion of maxmum likelihood in the conventional HMM during recognition. An HMM is defined as follows :

O : observation sequence

l_i : start time of the normalized state transition time slot of state S_i

u_i : end time of the normalized state transition time slot of state S_i

q : state transition sequences $q = q_1 \cdots q_T$ where q_t is the state at time t

Q : set of all possible state sequences

S : set of states S_i, $i = 1, \cdots, N$

λ : HMM/GPC

The maximum likelihood $P(O|\lambda)$ that model λ generates on observation sequence O in a HMM/GPC is as follows.

$$P(O|\lambda) = \max_{\substack{q_t = S_i,\ l_i \le \frac{t}{T} \le u_i,\\ q \in Q}} P(O \text{ and } q|\lambda) \tag{1}$$

Since the transition time limiting parameters of each state are obtained during training, the above definition is used only in recognition.

To find the HMM/GPC parameters, the conventional Baum-Welch algorithm is first used to obtain a_{ij} and $b_j(O_t)$. The optimal state sequence of the training data is then found by the Viterbi algorithm to estimate the transition time limiting parameters l_i and u_i for each state S_i

For example, if s_{ik} and e_{ik} are the start and end times of state S_i of the optimal state sequence of the kth training data, T_k training data and K is the number of training tokens per word, l_i and u_i are obtained as follows.

$$l_i = \min_{1 \le k \le K} \left(\frac{s_{ik}}{T_k}\right) \tag{2}$$

$$u_i = \min_{1 \le k \le K} \left(\frac{e_{ik}}{T_k}\right) \tag{3}$$

Fig.1 shows the start and end times s_{ik} and e_{ik} of the optimal state sequence q^* of the kth training data in an N-state HMM.

As stated above, the HMM/GPC restricts state transitions to specified time slots so that the number of distorted mathes are decreased. This implies that the HMM/GPC is able to provide a simple and effective representation of the time-varying characteristics of speech signals. Fig.2 shows the region of possible paths of the optimal state sequence in Viterbi decoding for the conventional HMM shown in Fig.3(a) and the HMM/GPC. As shown in the figure, the conventional HMM allows significantly distorted matches. To overcome this problem, the path of the optimal sequence is given a further restriction. An HMM with global path constraint which limits the region of the optimal state sequence is

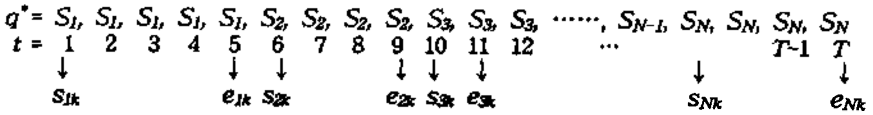
$q^* = S_1, S_1, S_1, S_1, S_1, S_2, S_2, S_2, S_2, S_3, S_3, S_3, \cdots\cdots, S_{N-1}, S_N, S_N, S_N, S_N$
$t = 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ \cdots\ T\text{-}1\ T$
s_{1k}
e_{1k}
s_{2k}
e_{2k}
s_{3k}
e_{3k}
s_{Nk}
e_{Nk}

Fig. 1. Finding the transition time limiting parameters in HMM/GPC (q^* : state transition sequence, t : time, s_{ik}, e_{ik} : transition time limiting parameter at state S_i).

[illegible] a simple and effective representation of the time varying characteristics of a speech signal. [illegible] may be seen as an application of the global path constraint of DTW to Viterbi decoding in HMM's.

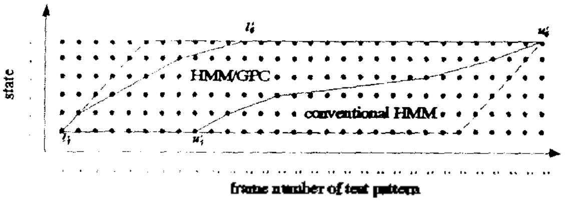

Fig. 2. Possible regions in Viterbi decoding for the conventional HMM and HMM/GPC.

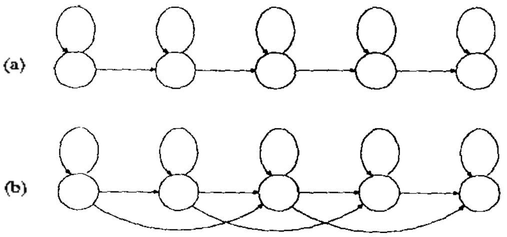

Fig. 3. HMM's with (a) 2 state transitions and (b) 3 state transitions

The HMM/GPC may be implemented as follows by slight modifications in the conventional Viterbi algorithm.

for $t = 1, 2, \cdots, T$

for $j = 1, 2, \cdots, N$

$$\delta_t(j) = \begin{cases} \max\limits_{1 \le i \le N}[\delta_{t-1}(i) + \log a_{ij}] + \log b_j(O_t), & l_j \le \frac{t}{T} \le u_j \\ -\infty, & \text{elsewhere} \end{cases} \quad (4)$$

In addition to requiring only slight modifications, the HMM/GPC also significantly reduces computation. This is because the Viterbi algorithm is only carried out in the region limited by the global path constraint, as shown in Fig.2, and leads to the reduction of mismatches for improved performance.

The biggest difference between the HMM/BSD and the HMM/GPC is that the former places constraints on the duration of each state, while the latter effectively limits the state transitions to prescribed time slots to disallow stray matches.

The global path constraint is also proposed for use in the HMM/SD to improve performance and reduce computation. Various HMM/SD's with different state duration densities were defined. For example, HMM/SD's with Poisson[3], Gammal[4], Gaussian[5] and Uniform[2] state duration densities and the bounded state duration HMM/BSD with minimum and maximum duration constraints were considered. The state duration densities were obtained from the training data by the Viterbi algorithm as in the case of HMM/GPC. If the global path constraint is applied to the HMM/SD (i.e. HMM/SD+GPC), state duration is limited by the state duration density and to a prescribed region.

As in HMM/GPC, the global path constraint in HMM/SD reduces mismatches and reduces computation. The globally path constrained state duration HMM is implemented using the transition time limiting equations of (2) and (3) in addition to the recursive relations of the state duration HMM[6].

Ⅲ. Experimental Results

To evaluate the performance of the proposed method, the HMM of Fig.3 was used for speaker independent isolated word recognition.

3.1 Database

The speech recognition database is composed of 11 Korean digits (*il*(1), *yi*(2), *s'am*(3), *s'a*(4), *oh*(5), *youk*(6), *chil*(7), *p'al*(8), *gu*(9), *kong*(0), *yong*(0) and 3 commands for a total of 14 words.

The training data was obtained from 50 male speakers in their twenties and thirties who pronounced each word twice (14 words × 50 speakers × 2 times = 1400 utterances), and the test data was given by 20 other male speakers in their twenties and thirties who pronounced each word twice (14 words × 20 speakers × 2 times = 560 utterances). Each utterance was recorded on digital audio tape (DAT) using a directed microphone (AT831b) in a relatively quiet environment.

3.2 The Speech Recognition System

A speaker independent isolated word recognition system using HMM's is constructed as follows. The speech signal is passed through a lowpass filter with a cutoff frequency of 4.5 KHz and sampled to 10 KHz at 16 bits. The sampled signal is preemphasized by a filter with a transfer funcion of $1-0.95z^{-1}$ and is partitioned into silence and speech intervals by endpoint detection[8]. The resulting speech signal is blocked into 20ms (200 sample) frames using a Hamming window with 10ms overlap, and 14th order LPC coefficients are obtained through LPC analysis. 14th order LPC cepstrum coefficients to be used in recognition are then found from the LPC coefficients. The feature vectors are vector quantized by a 256-codebook obtained using the LBG algorithm[9] into codeword sequences. The recognition system is based on a discrete observation HMM of 2 types of models as in Fig.3. The Baum-Welch algorithm was used to train the model for each word. In HMM speech recognition, the similarity between the input observation sequence and the model is measured by the Viterbi algorithm, and the model providing the highest probability is chosen to represent the input word. The minimum value of the elements of the observation probability matrix $B=\{b_j(O_t)\}$ was set at $1.0e^{-6}$.

3.3 Comparison of the performances of the conventional HMM and HMM/GPC

Speaker independent isolated word recognition experiments were conducted to compare the performance of the conventional HMM with the proposed HMM/GPC. Fig.'s 4(a) and 4(b) show the speaker independent isolated word recognition error rates of the conventional HMM and HMM/GPC as the number of states using the models of Fig. 3(a) and Fig. 3(b) varies. As shown in the figure, the globally path constrained HMM/GPC showed less error than the conventional HMM regardless of the number of states. These results imply that errors in the conventional HMM are due to erroneous matches. Therefore, reducing mismatches effectively in the HMM/GPC led to improved performance.

Fig. 4(a) compares the performance between the conventional HMM and HMM/GPC according to the number of states where 2 state transitions are possible for each state. For 5 states, the error rate was 5.9% and 3.4% respectively for the conventional HMM and HMM/GPC, showing a maximum improvement of 2.5%. Also, the minimum error was obtained for 9 states, with 2.7% and 1.6%, giving an improvement of 1.1%.

Fig. 4(b) shows the performance of the conventional HMM and HMM/GPC with 3 state transitions as the number of states varies. As can be seen from the figure, the globally path constrained HMM/GPC shows less error than the conventional HMM regardless of the number of states but showed slightly greater error than the results with 2 state transitions as shown in Fig. 4(a). That is, models with 2 state transitions as in Fig. 3(a) showed better performance than the mdoels with 3 state transitions. In the case of 3 state transitions, the minimum error rates of the conventional HMM and HMM/GPC were 3.0% and 2.1% respectively.

By observing the error rate according to the number of states, it can be seen that the error rate is greatly reduced as the number of sates increases for both the conventional HMM and HMM/GPC. This shows that a greater number of states is able to represent the time-varying characteristics of the training data to reduce the number of stray matches in Viterbi decoding. Especially, the conventional HMM is seen to produce greater errors as the number of states becomes small. The HMM/GPC shows improved performance by limiting mismatches using the global path constraint, and showed marked improvement in cases where the number of states is small. This is an important point, since the amount of memory and tranining data as well as compuation required in recognition increases as the number of HMM states increases.

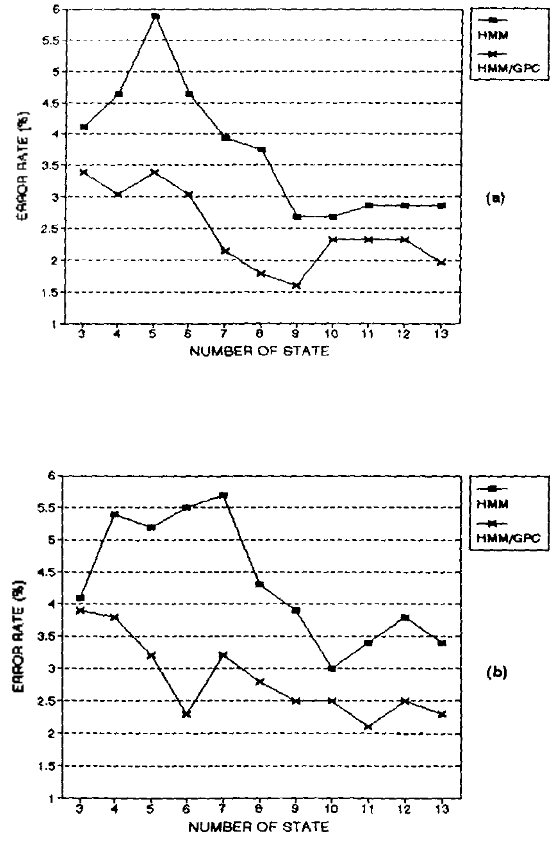

Fig. 4. Speaker independent isolated word recognition error rates for the conventional HMM and HMM/GPC with (a) 2 state transitions and (b) 3 state transitions.

3.4 Comparison of the performances of the conventional HMM, HMM/SD, HMM/SD +GPC and HMM/BSD+GPC

This experiment evaluates the performance of the HMM/SD and HMM/BSD using state durations and . HMM/SD+GPC and HMM/BSD+GPC with the proposed global path constraint. That is, the performances of the conventional HMM/SD with state duration density, the HMM/BSD with minimum and maximum limts in state duration, the HMM/SD+GPC with state duration density and global path constraint and the HMM/BSD+GPC with bounded state duration and global path constraint are compared. The state duration densities used are the Gamma, Poisson, Gaussian, and Uniform functions. The Geometric is the state duration density of conventional HMM. These state duration densities are found from the optimal state sequence of the training data obtained from the trained HMM's and the Viterbi algorithm. The minimum and maximum state duration of the HMM/BSD were also found from the optimal stae sequence.

Table 1 shows the results of the speaker independent isolated word recognition experiments using the HMM of Fig. 3(a) with 9 states. In the table, the case using the uniform state duration HHM/Uniform (2.0%) showed 0.7% improvement over the HMM without state duration (2.7%). Using various state duration densities with minimum and maximum duration (HMM/BSD) also shwoed improved performance over the HMM/SD. Also, using the global path constraint in HMM/SD+GPC and HMM/BSD+GPC also showed improvement over the HMM/SD and HMM/BSD respectively. The best performance was obtatined using the HMM/SD with uniform state duration (SD=Uniform) with the global path constraint in HMM/SD+GPC and HMM/BSD+GPC (1.4%), giving a 1.3% improvement over the HMM without state duration.

Comparing the performance of HMM/SD,

HMM/BSD and HMM/GPC in the table, it can be seen that the proposed HMM/GPC shows better performance than the conventional HMM/SD and HMM/BSD. These results show that the proposed HMM/GPC represents the temporal structure of speech more accurately than the HMM/SD and HMM/BSD, to effectively reduce mismatches in Viterbi decoding.

Table 1. Speaker independent isolated word recognition error rates for N = 9, 2 state transitions

HMM state duration model	HMM/SD	HMM/BSD	HMM/ SD+GPC	HMM/ BSD+GPC
Gamma	2.3	2.0	1.8	1.6
Poisson	2.1	2.0	1.6	1.6
Gaussian	2.5	2.3	2.0	2.0
Uniform	2.0	1.8	1.4	1.1
Geometric	2.7	2.3	1.6	2.0

Another advantage of the HMM/GPC is greatly reduced computation of the conventional Viterbi algorithm. As shown in Fig.2, the region included in the computation of the VIterbi algorithm is much reduced. For example, if V is defined as the number of words, N the number of states and k the number of state transitions per state, for an input pattern of length T, the computation for conventional Viterbi algorithm requires C_{mul} multiplications and C_{log} logarithmic operations[10]. In this paper, k takes values of 2 or 3 according to the number of state transitions.

Meanwhile, the HMM/GPC uses transition time limiting parameters for each word model and requires different amounts of computation for each word, so that the required computation must be determined experimentally. In Table 2, the average computation required to recognize test patterns in speaker independent isolated word recognition experiments by the conventional HMM and HMM/GPC were compared. Here, the number of words was $V = 14$ and average test pattern length was $T = 51.5$ frames. As shown in the table, for the HMM/GPC, the computation required for the conventional Viterbi algorithm dropped to below half in most cases. For $N = 9$ with 2 state transitions as in Fig. 3(a), the required computation decreased about 57%, and for the case with 3 shate transitions as in Fig. 3 (b), computation decreased 62%. The HMM/SD

Table 2. Average computational loads of the conventional HMM and HMM/GPC for all test patterns

number of state transition	$k=2$			$k=3$		
number of state N	$C_{mul} = C_{log}$		reduction ratio(%)	$C_{mul} = C_{log}$		reduction ratio(%)
	HMM	HMM/ GPC		HMM	HMM/ GPC	
3	4326	2436	56.3	6489	3613	55.7
4	5768	2930	50.8	8652	4216	48.7
5	7210	3448	47.8	10815	4782	44.2
6	8652	3949	45.6	12978	5455	42.0
7	10094	4433	43.9	15141	5929	39.2
8	11536	5024	43.6	17304	6623	38.3
9	12978	5572	42.9	19467	7301	37.5
10	14420	6168	42.8	21630	7777	36.0
11	15862	6724	42.4	23793	8337	35.0
12	17304	7180	41.5	25956	8985	34.6
13	18746	7842	41.8	28119	9410	33.3

needs approximately 15 to 20 times the computation of the conventional HMM, which is about 30 to 40 times the computation of the HMM/GPC. Moreover, the HMM/GPC requires less computation in recognition than the HMM/SD and HMM/BSD.

Ⅳ. Conclusion

In this paper, a HMM with global path constraint (HMM/GPC) that effectively represents the time varying characteristics of speech signals is proposed. State transitions were restricted to prescribed time slots in HMM/GPC to reduce stray matches for improved recognition. Also, the proposed method can be implemented by slight modifications to the conventional Viterbi algorithm and greatly reduces the required computation. HMM/GPC only needs two transition time limiting parameters in addition to the conventional HMM parameters, thus incurring alomost no increase in memory.

In addition, by applying the global path constraint to the HMM/SD and HMM/BSD using state duration densities, a method for improving performance and reducing computation was also proposed. In the HMM/SD+GPC where the global path constraint was applied to the HMM/SD, not only is the state duration limited by the state duration density but also has the effect of limiting it to a specified it to a specified time interval. Therefore, as in the HMM/GPC, the global path constraint reduces mismatches in HMM/SD and greatly decreases computation.

Comparing the results of the globally path constrained HMM/GPC and the HMM/SD+GPC with state duration density, it can be seen that the proposed HMM/GPC shows better performance than the conventional HMM/SD and HMM/BSD. This implies that the proposed HMM/GPC can represent the time-varying characteristics of speech signals more accurately than the HMM/SD and HMM/BSD and that distorted matches in Viterbi decoding may be reduced effectively.

REFERENCES

1. H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, No. 1, pp. 43-49, Feb. 1978
2. L. R. Rabiner, "A Tutorial on Hiden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, Vol. 77, No. 2, pp. 257-286, Feb. 1989
3. M. J. Russell and R. K. Moore, "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition," in *Proc. ICASSP*, pp. 5-8, March 1985
4. S. E. Levinson, "Continuous Variable Duration Hidden Markov Models for Speech Analysis," in *Proc ICASSP*, pp. 1241-1244, March 1985
5. K. F. Lee, *Automatic Speech Recognition*, Kluwer Academic Publishers, 1989
6. H. Gu, C. Tseng and L. Lee, "Isolated-Utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations," *IEEE Trans. Signal Processing*, Vol. 39, No. 8, pp. 1743-1743, Aug. 1991
7. L. R. Rabiner, B. H. Juang, S. E. Levinson and M. M. Sondhi, "Some Properties of Continuous Hidden Markov Model Representations," *AT&T Tech J.*, Vol. 64, No. 6, pp. 1251-1270, July-Aug. 1985
8. L. F. Lamel and L. R. Rabiner, "An Improved Endpoint Detector for Isolated Word Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, No. 4, pp. 777-785, Aug. 1981
9. Y. Linde, A. Buzo and R. M. Gray, "An Algorithm for Vector Quantization," *IEEE Trans. Commun.*, Vol. COM-28, No. 1, pp. 84-95, Jan. 1980
10. L. R. Rabiner, B. H. Juang, S. E. Levinson and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent Isolated Word Recognition," *Bell Syst. Tech. J.*, Vol. 62, No. 4, pp. 1075-1105, Apr. 1983

▲Weon Goo Kim

was born in Korea on April 3, 1964. He received the B.S. M.S. and Ph.D degress in electronic engineering form Yonsei University in 1987, 1989 and 1994, respectively. His current research interests are in the area of digital signal processing, speech signal processing and speech recognition.

▲Dong Soon Ahn

is Assistant Professor of computer engineering at Mokpo University.

▲Dae Hee Youn

is Associate Professor of electronic engineering at Yonsei University. For a photograph and biography, see pp. 52 of the December 1990 issue of this journal.