Development of a Real-time Voice Dialing System Using Discrete Hidden Markov Models

이산 HMM을 이용한 실시간 음성인식 다이얼링 시스템 개발

(Se Woong Lee*, Seung Ho Choi*, Mi Suk Lee*, Hong Kook Kim*,
Kwang Cheol Oh*, Ki Chul Kim*, Hwang Soo Lee*)
이 세 웅*, 최 승 호*, 이 미 숙*, 김 홍 국*, 오 광 철*, 김 기 철*, 이 황 수*

ABSTRACT

This paper describes development of a real-time voice dialing system which can recognize around one hundred word vocabularies in speaker independent mode. The voice recognition algorithm in this system is implemented on a DSP board with a telephone interface plugged in an IBM PC AT/486. In the DSP board, procedures for feature extraction, vector quantization(VQ), and end-point detection are performed simultaneously in every 10 msec frame interval to satisfy real-time constraints after detecting the word starting point. In addition, we optimize the VQ codebook size and the end-point detection procedure to reduce recognition time and memory requirement. The demonstration system has been displayed in MOBILAB of the Korea Mobile Telecom at the Taejon EXPO'93.

요 약

본 논문에서는 화자독립으로 100 단어를 인식할 수 있는 실시간 음성인식 다이얼링 시스템의 개발에 대하여 기술하였다. 이 시스템에서 음성인식 알고리즘은 전화선 인터페이스를 갖춘 DSP 보드상에 구현되었으며, IBM PC AT/486 상에서 작동된다. DSP 보드에서는 단어의 시작점이 검출된 후에 특징추출, 벡터양자화 그리고 끝점검출 과정이 실시간으로 10 msec의 프레임 구간마다 수행된다. 또한, 본 시스템에서는 인식시간과 기억용량을 줄이기 위해 VQ 코드북의 크기와 끝점검출 과정을 최적화하였다. 본 실시간 음성인식 다이얼링 시스템은 데모 시스템으로 구현되어 대전엑스포'93에서 한국이동통신의 MOBILAB 내에 전시되었다.

I. INTRODUCTION

As speech is one of the most natural and efficient means of communication for most people, there has been much efforts to add speech recognition capability to various machines. In particular, natural and efficient speech interfaces to various information service systems have been keenly required to satisfy the rapid progress of information demand on the way to a highly developed information society. This makes it possible to reduce the service manpower as well as to improve

*한국과학 기술원 정보 및 통신공학과
접수일자 : 1994년 1월 24일

the quality of telecommunication service.

This paper presents a voice dialing system that has been developed to provide various information services at mobile telephone base stations. A demonstration system using desk-top microphone input was developed in advance on a workstation to implement the operation procedure at the switching office with the automatic voice calling capability. We have also developed a real-time voice dialing system on a DSP board plugged in an IBM PC AT/486.

This paper describes mainly the real-time voice dialing system developed on a PC. In Section 2, we present an overview of the developed real-time voice dialing system. Details of the speech analysis and recognition algorithms adapted for real time operation are examined in Section 3. We discuss the performance of the developed voice dialing system in Section 4, and conclusions are made in Section 5.

II. VOICE DIALING SYSTEM OVERVIEW

2.1 Operation Procedure

The voice dialing system has four basic calling modes including, calling by numbers (digits), calling by institution names, calling by person names, and re-dialing, in addition to cancellation and help modes. One can call by numbers, /bunho/, by speaking a seven digit of telephone number digit by digit. Calling by institution names, /kigwan/, is provided by speaking public institution names that can be commonly used to everyone, which include /expo/, /idongtongsin/, /kwahagweon/, /chungwadae/, /sichung/, /bangsongguk/, /kisangcheong/, and /byungwon/ etc.

One can also call by person names, /ireum/, which should be registered in advance. The system can recognize up to 10 person names. The re-dialing mode, /tasi/, allows one to redial the phone number called most recently. The cancellation mode, /chuiso/, allows one to hang up the phone, and the help mode, /annae/, is provided for a usage inquiry of the voice dialing system.

The voice dialing system confirms the phone number after recognizing the number sequence or the name spoken according to each calling mode. If one answers yes, the system dials the confirmed phone number. If one answers no, it prompts a message for him to select a calling mode again. The system allows /ye/, /eung/, /grae/, /yes/ and /O.K./ for an affirmative answer, and /anio/, /ani/, and /no/ for a negative answer.

The operation procedure has been designed to restrict the number of retries for redial when the recognition error occurs or the line is busy. If the system connects to a destination number or it reaches the restricted number of dialing repetition, an end message is issued before completing the execution. Fig. 1 shows an example of the dialing operation in the voice dialing system.

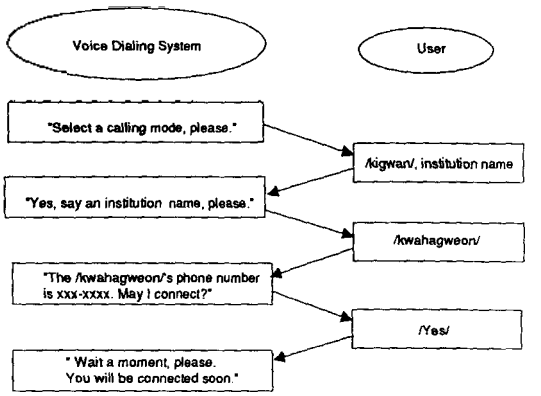

Fig 1. An example of dialing operation.

2.2 Hardware Configuration

H/W configuration of the voice dialing demonstration system is shown in Fig. 2. The Elf DSP Platform is a DSP board plugged in 16 bit AT bus slot, which includes a Texas Instruments' TMS 320C31 floating point digital signal processor and a 16 bit A/D and D/A converter. Voice dialing S/W is run on the DSP to recognize a voice command in real-time [1].

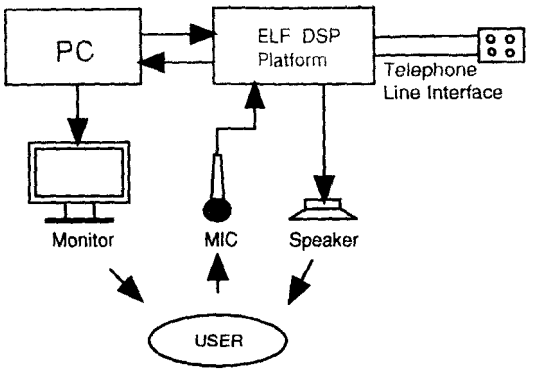

Fig 2. H/W configuration of voice dialing system.

2.3 Software Configuration

A S/W configuration in the demonstration system H/W explained in Section 2.2 is shown in Fig. 3. The analog signal from a headset microphone is converted to discrete numbers by an A/D converter and then end-points of a spoken word are detected. After the end-point detection, the signal that contains only portions of speech is transformed into a series of 16th order melcepstra in the feature extraction procedure [2].

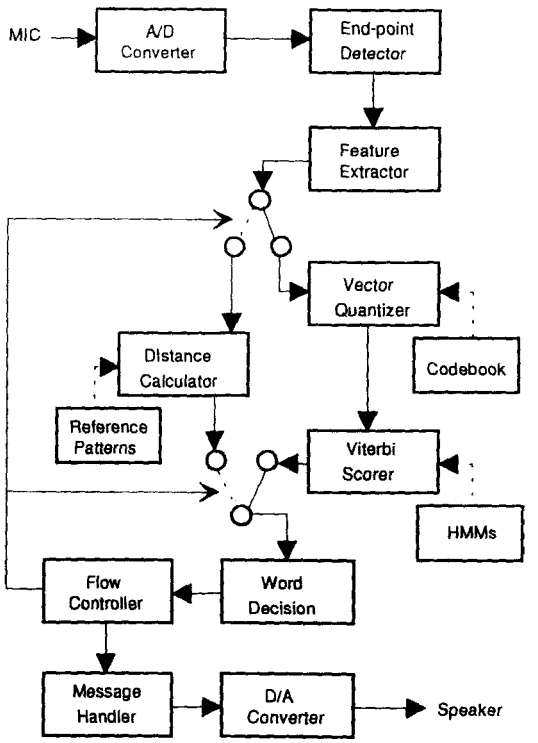

Fig 3. S/W configuration of voice dialing system.

The flow controller in the figure selects the dynamic time warping(DTW) procedure for speaker dependent recognition mode and the hidden Markov model(HMM) procedure for speaker independent recognition mode. In the speaker independent recognition mode, feature vectors are quantized using one of the codebooks according to a selected calling mode. The Viterbi scorer compares the quantized codeword sequence with trained HMM parameters to find a maximum likelihood model [5].

The voice dialing system has five kinds of codebooks and HMMs for four calling modes according to the operation procedure and response word groups (See Table 4.). The flow controller select corresponding codebook and HMMs to the selected operation procedure. In the speaker dependent recognition mode, input test pattern is compared with reference templates that consist of three utterances for each word spoken by the registered user. The reference pattern that has the minimal distance with the input test pattern becomes a recognized result. The result is sent to the flow controller, and it sends a responding message to the message handler for output.

Ⅲ. REAL-TIME SPEECH RECOGNITION ALGORITHM

3.1 Real-time End-point Detection

Input speech signal is processed frame-by-frame in which the length of each frame is 10 msec. Every three frame interval, a short-time average energy $E(i)$ and a zero crossing rate(ZCR) $Z(i)$ of the speech signal are calculated by

$$E(i) = \sum_{n=0}^{n=79} \|s(80i+n)\|, \quad i=0, 1, 2 \tag{1}$$

$$Z(i) = \sum_{n=0}^{n=79} \|sgn(s(80i+n)) - sgn(s(80i+n-1))\|, \quad i=0, 1, 2 \tag{2}$$

$$\text{where } sgn[s(n)] = \begin{cases} 1, & s(n) \geq UM \\ -1, & s(n) \leq LM \end{cases}$$

$$\begin{cases} LM = LM_n = \min(3m_e, 500.0) \\ UEL = UEL_n = 4LEL \\ TZCR = TZCR_n = \min(m_z + 2\sigma_z, 25) \end{cases} \quad (3)$$

$$\begin{cases} LEL = \lambda LM + (1-\lambda) LEL_n \\ UEL = 4LEL \\ TZCR = \lambda TZCR + (1-\lambda) TZCR_n \end{cases} \quad (4)$$

In the equations, LM and UM denote the margin to eliminate the effects of system noise and background noise, and we use the values of 8 and −10, respectively. The threshold values are updated ever five frames by (3) and (4). The upper threshold (UEL), the lower theshold (LEL), and the ZCR threshold ($TZCR$) are calculated by using average energy (m_e), mean (m_z) and standard deviation (σ_z) of ZCR from three frames of speech as shown in (3). The weighting factor for each threshold value in (4), λ, is 0.95^n, where n increases by 1 every five frames.

Pause detection is done by observing 20 frames of input signal. If there is no speech frame during a 20 frame interval, an end-point is detected. Pause is a silence interval between speech frames. As for Korean digits, there is no silence in the utterance. So we do not need pause detection for digit recognition.

We have considered three points in the end-point detection procedure for real-time operation as follows. First, a word starting point is detected without backtracking by calculating the short-time energy of recent three frames including a current frame, which enables us to decide whether the current frame is a tentative starting point or not. Second, the feature vectors are extracted from three frames by the feature extractor, which enables us to perform end-point detection and feature extraction simultaneously with a small input buffer for three frames of speech. Third, the maximum duration of a pause is restricted to 20 frames not to delay the execution of recognition procedure after the end-point detection.

3.2 Optimization of VQ Codebook Size

The performance of discrete HMM based speech recognition system is affected by the size of the VQ codebook. As the size of the codebook increases, it takes more time to encode a feature vector into a VQ codeword and the memory requirement for the codebook increases [7,8]. In addition, if the vocabulary size or feature vector space is not so large, a larger codebook does not always produce a better recognition result.

As the purpose of our voice dialing system is to recognize around one hundred word vocabularies in real-time, we have chosen the codebook size as small as the performance does not degrade. To decide the codebook size, speaker independent digit recognition experiments have been performed with the following speech database :

- recording environment : silent office
- training date : 10 utterances of 5 male and 5 female speakers
- test data : 10 utterances of 5 male and 5 female speakers
- A/D sampling rate : 10 kHz with 12 bit quantization

Table 1 shows the recognition performance of spoken digit for three different VQ codebook sizes. Table 2 lists The required memory size and VQ encoding time for each codebook size are given in Table 2. For the digit recognition experiments, the number of words is 11 and the maximum number of HMM states is 18. From these simulation results, we choose the VQ codebook size of 64 for the voice dialing demonstration system.

Table 1. Codebook size versus digit recognition rate

Number of codewords	Average recognition rate
64	92.4 %
128	91.1 %
256	91.7 %

Table 2. Codebook size versus memory requirements and VQ encoding time

Number of codewords	Codebook size (word)	HMM size (word)	VQ encoding time (msec)
64	1024	16,236	0.17
128	2048	28,908	0.33
256	4096	54,252	0.65

3.3 On-line Speech Analysis

Instead of serially executing feature extraction and recognition algorithms after end-point detection, procedures for end-point detection, feature extraction, and vector quantization are performed simultaneously in every 10 msec frame interval after detecting the word starting point to satisfy a real-time constraint. After A/D conversion with 8 kHz sampling rate, 80 samples of the speech signal constitute a frame for every 10 msec.

In speaker independent recognition mode, feature vectors are extracted from Hamming-windowed speech signal of three frames, i.e., 30 msec. After encoding 16th order melcepstrum into a codeword by VQ, the end-point of a word is decided. This process is repeated every 10 msec. When the end-point is detected, a Viterbi score is calculated for each reference HMM, and then a maximum likelihood model is decided as a recognized word. The processing time for Viterbi scoring depends upon the number of word vocabularies, the number of HMM states and its architecture, and the size of the VQ codebook.

In speaker dependent recognition mode, the distance between a test pattern and each of the reference pattern is calculated through the DTW procedure to decide a reference pattern with minimum distance as a recognized word. The processing time for distance calculation in DTW depends upon the number of word vocabularies, word length, and the number of reference patterns for each word. There are three reference patterns for each word in our system. The processing time and the size of each recognition module are given in Table 3.

Table 3. Processing time and size of each module

Module with a driver	Processing time (msec)	Module size (word)
end point detection	0.2	32,350
feature extraction	1.65	24,900
VQ encoding	0.17	25,126
Viterbi scoring	360	39,197
DTW	285	64,482

Ⅳ. RECOGNITION EXPERIMENTS AND RESULTS

The performance of the real-time voice dialing system is obtained by simulating the same recognition algorithm on a workstation using the same speech data collected from the ELF DSP board on a PC. The speech database is obtained under the following conditions :

- recording environment : usual office noise
- microphone : Icom HS-58 headset
- speech DB contents : 5 word groups of 6 to 11 isolated words
- training data : 17 male and 10 female speakers
- test data : 10 male and 6 female speakers
- number of utterance : 3 times per word (except 5 times per digit)
- A/D sampling rate : 8 kHz with 16 bit quantization

Each word is modeled using a word-based HMM, in which the number of states for each word is proportional to the number of phonemes that constitute a word, and each phoneme consists of three states. Each VQ codebook contains 64 codewords obtained using the modified K-means algorithm [8], and discrete HMM parameters are estimated using the Baum-Welch algorithm. The recognition results for the five groups of word vocabularies are shown in Table 4.

The result of a response word group is the average recognition rate of each word that is recognized as a /yes/ class or a /no/ class, and there are 5 words for the /yes/ class and 3 words for the /no/ class. /young/ and /kong/ in a digit word group, two different utterances for /0/, are also

Table 4. Recognition rate of each word group

Word group	Average recognition rate
command (6 words)	99.3 %
institution name (8 words)	99.0 %
person name (10 words)	98.5 %
response (8 words)	97.1 %
digit (11 words)	88.8 %

classified as the same group. As shown in Table 4, the recognition performance for the command, institution name, person name, and response word groups reveals nearly perfect results except the digit group. The recognition rates for confusable sounds, /il/, /chil/, /kong/, and /ku/ in the digit word group, which are the pronunciation of /1/, /7/, /0/, and /9/, respectively, fall around 80 %.

When we obtain digit sounds in a silent office environment and detect end-points by hand, we can obtain the multi-speaker dependent digit recognition rate of about 99 %, and the speaker independent digit recognition rate of about 92 %. Since Korean digits are a confusable vocabulary set with a single syllable, the end-point detection is critical to the recognition performance. On the other hand, other words that consist of multi-syllable undergo little degradation from an end-point detection error, because it can be classified correctly during the longer classification time. Therefore, the end-point detection algorithm should work more exactly in noisy environments to improve the performance of digit recognition in practical situations.

V. CONCLUSIONS

An implementation of a real-time voice dialing system and its recognition algorithm based on discrete hidden Markov models are described in this paper. The system shows very high recognition performance for the multi-syllable word groups, though there is a room to improve the digit recognition performance. Future work should include robust end-point detection and feature extraction to improve the recognition performance in noisy environments, and to make an interface to the mobile telephone switching office to provide voice dialing service as well as other services including information retrieval by voice commands.

References

1. Atlanta Signal Processors Inc., *ELF DSP Platform Instruction Manual*, pp. 1-9, 1992.
2. S. Imai, "Cepstral analysis on the mel frequency scale," in *Proc ICASSP'83*, Apr. 1983, pp. 93-96.
3. X. D. Huang, Y. Akiri, and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, 1990.
4. H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 26, No. 1, pp. 43-49, Feb. 1978.
5. G. D. Forney, Jr., "The Viterbi algorithm," in *Proc. IEEE*, Mar. 1973, Vol. 61, No. 3, pp. 268-278.
6. S. J. Doh, I. H. Sohn, and M. W. Koo, "A real-time endpoint detection algorithm for speech signal," in *Proc. 5th Korean Signal Processing Conference*, Sep. 1992, Vol. 5, No. 1, pp. 11-14.
7. Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communication*, Vol. 28, No. 1, pp. 84-95, Jan. 1980.
8. J. G. Wilpon and L. R. Rabiner, "A modified K-means clustering algorithms for use in isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 33, No. 3, pp. 587-594, June 1985.