CONTINUOUS DIGIT RECOGNITION FOR A REAL-TIME VOICE DIALING SYSTEM USING DISCRETE HIDDEN MARKOV MODELS

S. H. Choi, H. J. Hong, S. W. Lee*, H. K. Kim**, K. C. Oh, K. C. Kim, and H. S. Lee

Department of Information and Communication Engineering
Korea Advanced Institute of Science and Technology
207-43 Cheongryang-dong, Dongdaemun-ku
Seoul 130-012
Republic of Korea

* R & D Laboratory
Goldstar Telecommunication Co., Ltd.
P. O. Box 532, Anyang, Korea 430-081

** Intellegent Software Laboratory
Samsung Advanced Institute of Technology
P. O. Box 111, Suwon, Korea 440-600

ABSTRACT This paper introduces a interword modeling and a Viterbi search method for continuous speech recognition. We also describe a development of a real-time voice dialing system which can recognize around one hundred words and continuous digits in speaker independent mode. For continuous digit recognition, between-word units have been proposed to provide a more precise representation of word junctures. The best path in HMM is found by the Viterbi search algorithm, from which digit sequences are recognized. The simulation results show that a interword modeling using the context-dependent between-word units provide better recognition rates than a pause modeling using the context-independent pause unit. The voice dialing system is implemented on a DSP board with a telephone interface plugged in an IBM PC AT/486.

1. INTRODUCTION

In this paper, we introduce a word-juncture modeling and a Viterbi search method for continuous digit recognition. Continuous speech recognition requires more complex modeling and search method than isolated one. Additionally, it is difficult to recognize continuous speech because of the ambiguous word boundaries and the coarticulation between words as well as within word.

In this research, we use a word-based HMMs for each digit, and two silence models are for the begining and the ending of continuous digit, respectively. As well, we use context-dependent between-word units to capture the coarticulation between words and to provide a more precise representation of word junctures [1, 2].

There are two search strategies for recognizing continuous speech [3]. One is to segment word sequence as specialized recognition units and then classify words by these units. The drawback of this method is that the recognition performance could be

degraded from inexact segmentations. The stack decoding algorithm and many word-based dynamic programming (DP) search methods such as the two-level DP approach, the level-building algorithm, and the one-stage DP approach follow this idea. The other strategy finds the best search path as well as the word boundaries through the network constructed from the lexicon. In order to find the best path in HMM network, we use the time-synchronous Viterbi search algorithm which is known to be appropriate for real-time operation [4].

This paper is organized as follows. Section 2 presents an overview of the application, the voice dialing system. The word-juncture modeling of Korean continuous digit and the search method to find the best digit sequence are described in section 3 and 4. We investigate the performance of the continuous digit recognition in sectin 5 by recognition experiments, and we conclude this paper in section 6.

2. VOICE DIALING SYSTEM OVERVIEW

We developed a real-time voice dialing system which can recognize around one hundred word vocabularies in speaker independent mode and continuous digit. 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 [5].

Hardware configuration of the voice dialing system is shown in fig. 3. The Elf DSP Platform hardware is a DSP board plugged in 16 bit AT bus slot, which includes a Texas Instruments TMS320C31 floating point digital signal processor and a 16 bit A/D and D/A converter [6].

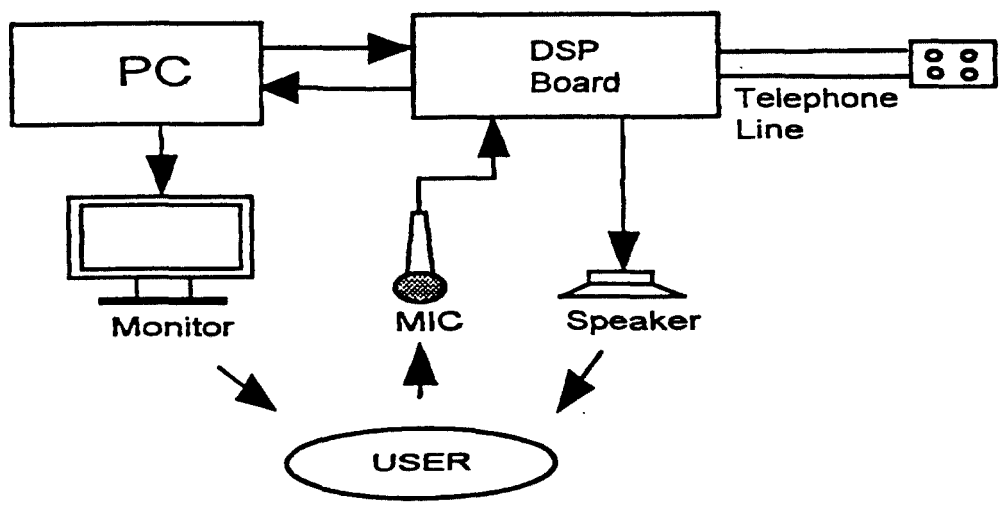


Fig. 3 Hardware configuration of voice dialing system.

The voice recognition algorithm 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.

The recognition rate is about 98.5 % for word groups and 88.9 % for isolated digit in a usual office environment. Also, the system can recognize continuous digits consisted of four digits for an internal office telephone number. In this paper, we investigate a continuous digit recognition method.

3. MODELING AND TRAINING

3.1 Recognition Units for Modeling

For comparison, word junctures are modeled with two kinds of between-word units. One is pause modeling using a pause unit which may be occured between digits. The other is interword modeling using context-dependent between-word units to model coarticulation between digits. We use 56 between-word units to account the combination of 8 final sounds of the preceding word, /ʌ/, /i/, /m/, /a/, /o/, /g/, /u/, /ng/, and 7 initial sounds of the following word, /i/, /s/, /o/, /yu/, /ch/, /p/, /g/, appeared in Korean digit sequence.

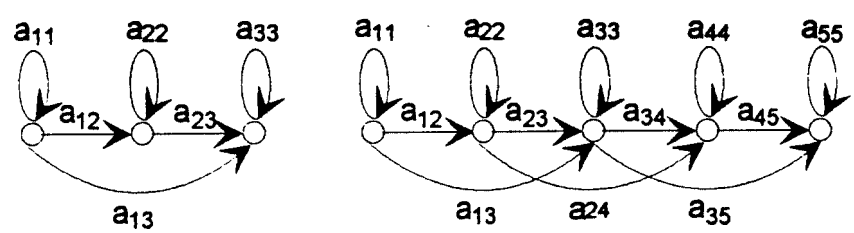
The recognition units defined are as follows :

- 10 whole-word units for each digit
- two beginning and ending silence units
- one pause unit between digits for pause modeling
- 56 between-word units interword modeling

3.2 HMM Modeling

In this paper, mel-scaled cepstrum is used for a speech feature parameter to take into account the auditory characteristics.

We choose discrete hidden Markov Model (HMM) for modeling each speech unit including digit, silence, pause, and between-word units for real-time operation, which requires less training and recognition time than continuous HMM. The continuous feature vectors are mapped into one of the codebook vector that is closest to it in spectral distance and the codeword indices are used for HMM training and recognition. Fig. 1 shows the HMM structure for each speech unit.



(a) (b)

Fig. 1 HMM structures for (a) silence, pause, and between-word units, and (b) digits

3.3 HMM Training

For isolated digit, we construct a HMM consisted of three states, beginning and ending silence HMM, and a five states digit HMM. For continuous digit, we segment each speech unit manually, and each speech unit is used to obtain discrete HMM parameters, transition and observation probabilities. In order to train each HMM, the forward-backward algorithm is used [7]. This iterative training is ended when the increment of the likelihood average is reached at converge threshold value.

4. SEARCH ALGORITHM

We apply the Viterbi search algorithm [8] to find a digit sequence in the HMM recognition. The Viterbi search finds the optimal HMM state sequence associated with the given observation, and also finds the digit sequence in a large HMM representing all possible word sequences. At every time frame all accessible states are updated with probability score, and at the end of the search, the path with the highest score is backtraced.

A finite state network (FSN) for digit sequence is pre-compiled to represent the search space. Fig. 2 represents FSN when one pause or 56 between-word units are used.

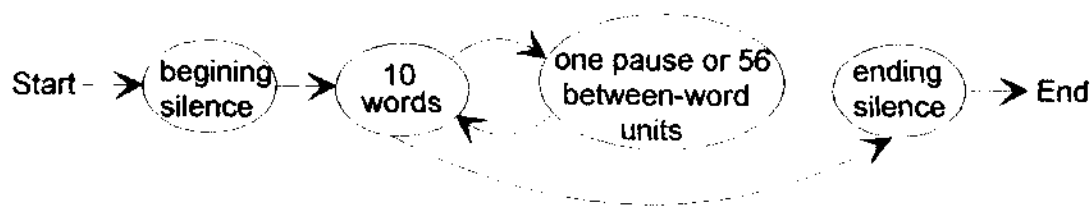


Fig. 2 Finite state network

5. RECOGNITION EXPERIMENTS

The performance of continuous digit recogintion using a interword modeling is evaluated on a workstation with the speech data collected from the Elf DSP board on a PC by simulating the same recognition algorithm running on the Elf DSP Platform.

5.1 Database

The speech database is obtained under the following conditions :

- usual office environment
- Icom HS-58 headset
- 100 continuous digit sequences, each consisted of 4 digits
  (for equal frequency of same word context)
- training data : three male and three female speakers

• test data : two male and two female speakers (different from the training speakers)
• A/D : 8 kHz sampling with 16 bit linear PCM

The end-points are determined on the PC in real time to include about 10 msec silence interval at both ends.

5.2 Continuous Digit Recognition

To evaluate a interword modeling, we compared the interword modeling including the context-dependent 56 between-word units with a pause modeling including only one pause unit between digits. A postprocessing is applied after the Viterbi search to exclude labels obtained from the insertion errors which yield more than 4 digits with smaller speech frames.

5.2.1 Pause Modeling

Table 2. shows the recognition results when one pause unit is used between digits. Sentence recognition rate represents the percentage when all four digits are correctly recognized.

Table 2. Continuous digit recognition result (one pause between digits).

word recognition	83.4 %
sentence recognition	51.8 %

5.2.2 Interword Modeling

In Table 3, the recognition results are listed when 56 between-word units are used between digits.

Table 3. Continuous digit recognition result (56 between-word units).

word recognition	84.8 %
sentence recognition	54.0 %

As seen in Table 2 and Table 3, the interword modeling with context-dependent between-word units gives 1.4 % improvement in word recognition rate over the pause modeling with a pause unit between digits and 2.2 % improvement in sentence recognition rate.

6. CONCLUSIONS

In this paper, we describe a continuous digit recognition method and a development of

real-time voice dialing system. To represent the word-juncture precisely, we use context-dependent between-word units, and applied the Viterbi search algorithm to find the optimal word sequence. Though a slight performance improvement was obtained using the context-dependent interword modeling, the recognition performance was degraded substantially when the speaking rate is very fast and word boundaries are crumbled. For the robust recognition of continuous digit, more flexible modeling and reliable search method is to be investigated.

REFERENCES

1. C. H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg, "Improved Acoustic Modeling for Large Vocabulary Continuous Speech Recognition, " *Computer Speech and Language*, Vol. 6, No. 2, pp. 103-107, April 1992.

2. E. P. Giachin, C. H. Lee, L. R. Rabiner, A. E. Rosenberg and P. Pieraccini, "On the Use of Inter-Word Context-dependent Units for Word Juncture Modeling, " *Computer Speech and Language*, Vol. 6, No. 3, pp. 197-213, July 1992.

3. H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition, " *IEEE Trans. on ASSP*, Vol. 32, No. 2, pp. 263-271, April 1989.

4. K. F. Lee, *Automatic Speech Recognition*, Kluwer Academic Publishers, Norwell, M.A., U. S. A, 1989.

5. S. W. lee, S. H. Choi, M. S. Lee, H. K. Kim, K. C. Oh, K. C. Kim, and H. S. Lee, "Development of a Real-time Voice Dialing System Using Discrete Hidden Markov Models, " *Journal of ASK*, Vol. 13, No. 1E, pp. 89-95, January 1994.

6. Atlanta Signal Processors Inc., *ELF DSP Platform Instruction Manual*, 1992.

7. L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models, " *IEEE ASSP Magazine*, Vol. 3, No. 1, pp. 4-16, January 1986.

8. G. D. Forney, "The Viterbi Algorithm, " *Proc. of IEEE*, Vol. 32, No. 2, pp. 263-271, April 1989.