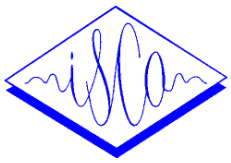

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Extension Number Guidance System

Fumihiro YATO*, Kazuki KATAGISI** and Norio HIGUCHI*

* KDD Kamifukuoka Laboratories
2-1-15 Ohara Kamifukuoka-shi, Saitama 356, Japan

** ATR Interpreting Telephony Research Laboratories
Sanpeidani Inuidani, Seika-cho Soraku-gun, Kyoto 619-02, Japan

Abstract

The authors have developed a speaker independent word recognition system for the purpose of giving extension number guidance to our laboratories. The system consists of a workstation, AD converter, PBX interface and speech synthesizer. The vocabulary consists of the names of persons and laboratories, and its size is about 300. The recognition process is based on three methods: continuous DP matching, multi-templates and SPLIT methods. The system announces the results of recognition and guidance of the extension number to the user by using a Klatt type speech synthesizer.

1. Introduction

In KDD laboratories, speech synthesis and speech recognition are one of the main research topics. This time the authors have developed a extension number guidance system as an application of speech synthesis and speech recognition techniques. Our system is designed as a speaker independent word recognition system, and has an improved user interface by using a rule-based speech synthesis system. In this article, the hardware configuration is first described, next the detail of the recognition process is explained, a new approach for constructing a good user-interface for the word recognition system is proposed, and finally experimental results are presented.

2. Hardware configuration

The task of our recognition system is to provide extension number guidance to our laboratory staffs. There is no restriction on the users who can access the system through the telephone line. The vocabulary consists of the names of personnel and laboratories, and its size is about 300. Our system is designed assuming that the user knows the organization of KDD laboratories.

Figure 1 shows the hardware configuration of our recognition system. In this system, a PBX interface is included to connect and disconnect the subscriber line to the recognition system. The PBX interface also can detect and decode PB signals. The AD converter is used at a 8kHz sampling rate with 16 bits accuracy. A workstation MC5600 and vector accelerator VA1 are used to implement the recognition processes, but the recognition process may take about 30 sec in the worst case. The speech synthesizer is a Klattalk type rule-based speech synthesis system [1], and it is connected to the workstation by RS232C interface. The synthesizer is used for the following purposes.

(1) guide the user
"Please give the name of the lab."
(2) confirm the recognition result
"The system recognizes Yato, is that correct?"
(3) announce the extension number
"The extension number of Mr. Yato is 7382."

Actually these sentences are announced in Japanese, and are constructed by a fixed form sentence plus some key word in the work-station. Because of using a rule-based speech synthesis system we can change the contents of the announcement very flexibly, so it becomes easy to update the system according to personnel changes.

In this system, at first the user calls the system, then the PBX interface detects the ringing tone and interrupts the workstation to start the system program. The user replies to the questions asked by our system in several stages, namely from the level of the names of laboratories to the level of person's names according to the hierarchical structure of the organization of our laboratories as shown in Fig.2. In each stage the number of words handled by the system is less than 20, so the system can recognize the input word with high accuracy. The system then announces the recognition result to the user in order to confirm. The confirmation is performed by using a PB signal. Finally, the extension number is announced to the user.

3. Recognition Process

Figure 3 shows the block diagram of the system process. The input signal is analyzed by a 10th order PARCOR method every 10 msec, and the PARCOR coefficients are then converted to LPC cepstrum coefficients. Detection of the speech period is based on the short time average energy and zero crossing number of each analysis frame. The recognition process is based on continuous DP matching [2], multi-template and SPLIT (strings of phoneme-like templates) methods [3].

3-1. Speech Period Detection

Usually it is difficult to detect the speech period exactly from the input signal through the telephone line, so in our system the interval I_s of candidates for a starting point is found from an utterance, instead of exact start point, by using the short time average energy E_i and zero crossing number Z_i where i is the frame number, and the interval I_e for the end point is also found as shown in Fig.4. The algorithm for obtaining the interval I_s is as following:

(1) Find the first frame S_1 where E_i is greater than E_{t1}. S_1 is the final frame of I_s.
(2) Find the last frame S_2 where E_i is greater than E_{t2} before point S_1.
(3) Search the frames F_k (k=1,K) where Z_i is greater than Z_t from S_2 until frame (S_2 - 25).
(4) If the number K is greater than 2, then F_K is the beginning frame of I_s, otherwise S_2.

The interval I_e is found to apply the above algorithm in the reverse direction, and in the continuous DP matching process, a DP pass must start from some frame included in I_s and finish on some frame included in I_e. Preliminary experiments show that the accuracy of word recognition is improved more than 5% by using this method.

3-2. DP matching

512 phoneme-like templates are produced by clustering a set of LPC cepstrum parameters extracted from the utterances of 296 words of our task spoken by 4 male speakers. The word templates are also produced from the same utterances, and are denoted as strings of numbers of phoneme-like templates. Four templates are therefore used as the reference template of each word.

In our system the continuous DP matching calculation is performed in the following way:

$$g(i,j) = \min \begin{cases} g(i-2,j-1)+2d(i-1,j)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-1,j-2)+2d(i,j-1)+d(i,j) \end{cases} \quad (1)$$

where d(i,j) denotes the distance between the i'th frame of input speech and the j'th frame of the reference template, and g(i,j) denotes the accumulated distance. By using this algorithm, the slope constraint restricts the local slope between 1/2 and 2. If the i'th frame of the input speech which is included in the interval I_e and the final frame J of the r'th reference template are matched, the normalized accumulated distance D(i) is obtained by dividing g(i,J) by the weight factor for the DP path. The smallest D(i) is regarded as the distance D_r between the input word and the r'th reference template. The distance D_r is obtained for every reference template.

3-3. Confirmation Process

As a result of the recognition step, 4 candidates are found based on the distance D_r. According to preliminary experiments, the following process is employed to confirm the recognition result.

If the 4 candidates are the same sort of word, then the system decides that the correct word has been obtained, otherwise the system finds the candidates which satisfy the relation:

$$Dci < 1.2 * Dc1 \qquad (2)$$

where Dci denotes the distance of the i'th candidate. If the 2nd candidate does not satisfy the above condition, then the word sort of the 1st candidate W1 is regarded as the correct answer and our system requests confirmation of the word by using a PB signal. Otherwise, if the 4 candidates include 4 word sorts and the 4th candidate satisfies condition (2), then the system regards the recognition result as ambiguous and requests the user to input the word again. If two or three word sorts are included in the candidates satisfying condition (2), then the system requests the user to select the correct word by using a PB signal. Our system announces the following sentences in each case:

(1) for a unique word sort
*"System recognized word X, if this is correct, please push #, otherwise push *."*
(2) for 2 or 3 word sorts
*"System recognizes word X or Y (or Z), if X is correct then push 1, if Y is correct then push 2, (if Z then push 3,) otherwise push *."*
(3) for an ambiguous result
"System cannot recognize your word, please input it again."

If the user pushes the "*" button, our system returns to input standby. At that time, the word sorts announced by our system are rejected from the list of reference templates.

4. Conclusions

As described above, an improved user interface is applied in our recognition system so it can achieve high accuracy. As an experimental result for 5 male and 2 female speakers, the system gives the correct extension number with more than 99% accuracy. The confirmation sentence includes 2 or 3 word sorts mainly when the reference word set includes words which resemble each other, for example 'keirigakari', 'keibigakari' and 'seibigakari', or 'hashimoto' and 'matumoto'. The ratio of such an occurrence is about 10%.

However even if the system can recognize the input utterance with 100% accuracy, the user will feel that a discrete word recognition system is not comfortable. If we want to improve the user interface of the recognition system still further, the system must have the ability to handle continuous speech. We are therefore planning to introduce phoneme recognition and word spotting methods.

Acknowledgements

The authors wish to thank Dr. K. Ono, Dr. Y. Urano for the encouragement of this research and to express thanks to other members of the AI group of KDD R&D Laboratories for their discussions and cooperation.

Reference

[1] N. Higuchi, S. Yamamoto and T. Shimizu, "A Japanese speech synthesizer based on the production rules," Proc. of 2nd Joint Meeting of ASA and ASJ, J18, Nov. (1988).
[2] S. Hayamizu and R. Oka, "Experimental studies on the connected words recognition using continuous dynamic programming," Trans. IECEJ, J67-D, 6, pp.667-684 (1984).
[3] N. Sugamura and S. Furui, "Isolated word recognition using strings of pseudo-phoneme templates (SPLIT)," J. Acoust. Soc. Japan, (E)5, 4, pp.243-252 (1984).
[4] S.Furui, Digital speech processing, synthesis, and recognition, Marcel Dekker, Inc., New York (1989).

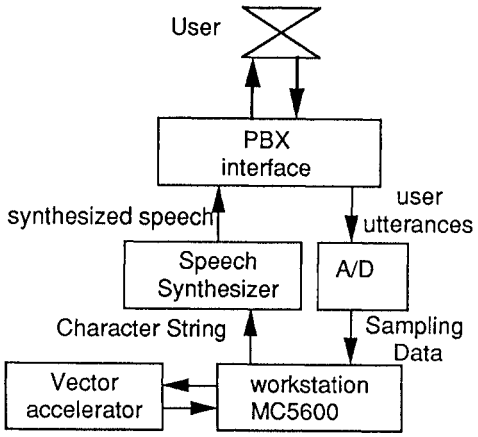

Fig.1 System Configuration

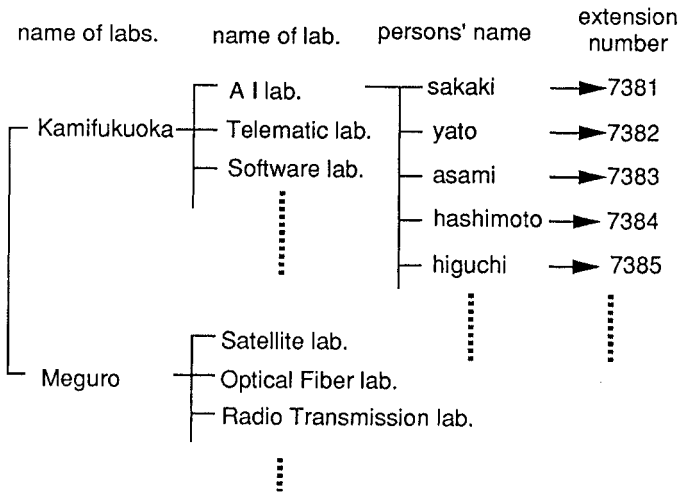

Fig.2 Hierarchical Structure of our Laboratories

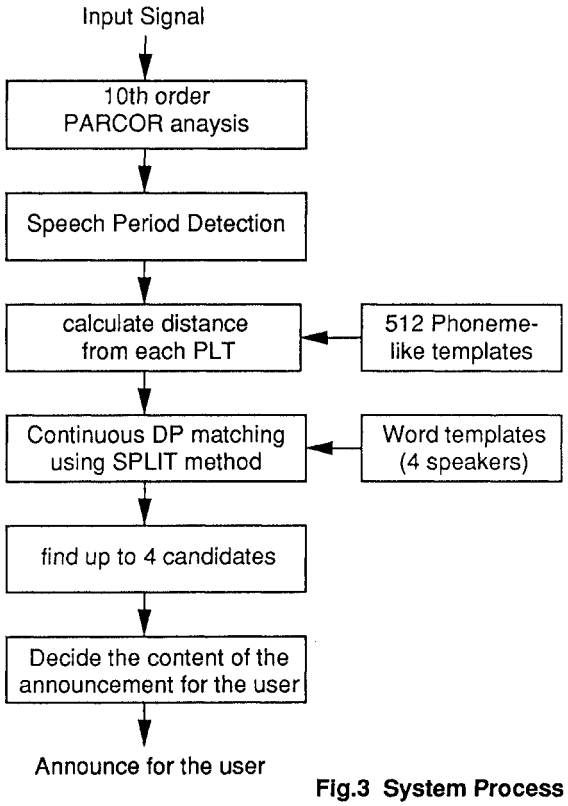

Fig.3 System Process Flow

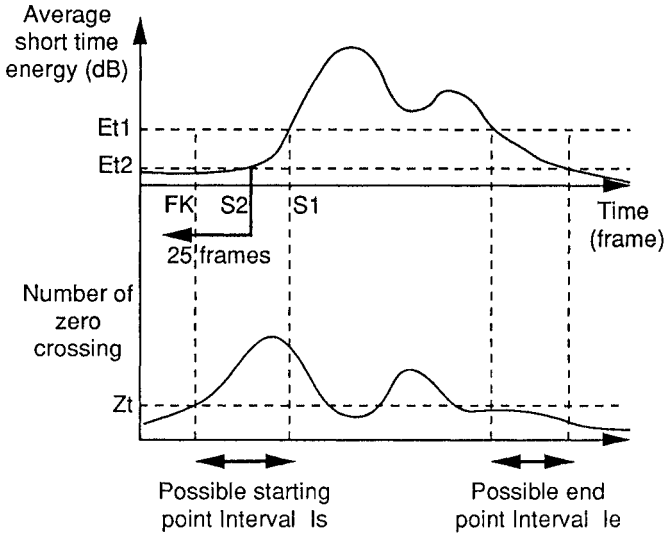

Fig. 4 Speech Detection Process