TELEPHONE SPEECH RECOGNITION SYSTEM WITH HIGH NOISE IMMUNITY

Y. Fujihashi and A. Fukui

Integrated Switching Development Division, NEC Corporation
1131 Hinode Abiko, Chiba, 270-11 JAPAN

ABSTRACT

The purpose of this paper is to suggest the use of the word spotting method for isolated word recognition in order to realize a telephone speech recognition system with high noise immunity.

Our recent simulation test reveals that the word spotting method is highly immune to burst noise which causes the decrease of the recognition rate of the conventional system that needed to detect voiced periods by voice power before DP matching.

Based on the algorithm using the word spotting method, we have developed a new telephone speech recognition system. Mounted on one card, it consists of a 16-bit CPU (μPD70216), a DP matching processor (μPD7764) and a speech analysis processor (μPD77C25). In its maximum configuration, this system is capable of recognizing 400 words. In our evaluation test, this system showed a recognition rate of over 98% under noiseless conditions, and over 95% under noisy conditions where human voices and telephone bell sounds were present. The results of this test indicate that this telephone speech recognition system has a very high immunity to noise.

1. INTRODUCTION

With the progress of speech recognition technology, various types of speech recognition systems have been developed for practical use. However, the performance of these systems is easily influenced by the conditions under which they are used. These usage conditions include the characteristics of the individual speaker, voice level and, most of all, the noise condition.

The problem of noise condition stands out with those systems where speech enters through a telephone handset. Because, compared with close-talking microphones, the telephone handset is apt to catch surrounding noise, the noise conditions are often severer, which might be sometimes worsened by noise on transmission lines. Also, once these systems are put into operation, it becomes difficult to adjust the voice level on the telephone set or to reduce the surrounding noise level. In order for this kind of system to work efficiently, it is important to consider countermeasures against noise before putting the system into operation.

In view of these shortcomings of the conventional systems, we discuss in this paper that the word spotting method be applied to isolated word recognition in order to realize highly noise immune telephone speech recognition systems. The following text consists of: Section 2, Study on Surrounding Noise; Section 3, Speech Recognition Algorithm; Section 4, Simulation Test Results; Section 5, System Configuration; and Section 6, Evaluation Test Results.

2. STUDY ON SURROUNDING NOISE

We have developed before a speaker independent speech recognition system and applied it as a voice dialing equipment. [1] This system employed DP matching method, using a multi-template. It showed a recognition rate of 98% for numbers from one to ten under noiseless conditions, while a much lower rate was registered on field.

According to analysis on those incorrectly recognized and rejected input voices, they were accompanied by various kinds of noises. These noises roughly break down into three categories:

(a) Sounds from things around (for example, door slam)
(b) Breathing sounds from the speaker
(c) Voices from people around

These noises were registered in so many different levels and their locations so varied (some were located before and after input voices, some connected to input voices, some overlapped, etc.) that it was difficult to distinguish one noise from others.

Since this speaker-independent speech recognition system detects voiced periods by voice power, if they are accompanied by noise, it fails to detect them correctly. This can often result in incorrect recognition or rejection of these voices and is a primary cause of the decrease of the recognition rate.

3. SPEECH RECOGNITION ALGORITHM

As countermeasures against noise, following various methods have been proposed.

(a) Noise subtraction : Subtraction background noise from input voices
(b) Adaptation of reference : Taking noise influence into account when setting reference patterns
(c) Method of acoustic analysis : Addition of A-characteristics, spectrum normalization, auditory model analysis, etc.

These methods are effective for steady noise, incorrect recognition or rejection of input voices is inevitable if the spoken periods are incorrectly detected because of accompanied burst noise.

In view of the fact that the word spotting method used for continuous speech recognition [2] does not require detecting voiced periods, we considered its application to isolated word recognitions. The word spotting method searches the whole of input voices for particular words, and it is considered an effective method for isolated word recognition of voices accompanied by burst noise.
To realize this method, we used Blockwise DP matching algorithm [4], a variation of Clockwise DP matching algorithm [3]. By using this algorithm, we performed DP matching continuously in real time. Figure 1 shows the DP matching pass allowable range. The DP pass is asymmetrical and the recurrence equation can be expressed as the following.

\[
g(i, j) = \begin{cases} 
g(i, j-1) + d(i, j) + MIN \\
\min(g(i-1, j-1), g(i-1, j), g(i-2, j-1)) 
\end{cases}
\]  

(1)

where \(i\) = input pattern frame and \(j\) = reference pattern frame. \(d(i, j)\) is vector distance, based on Chebyshev distance.

The distance between the reference pattern and the input voice pattern is calculated for each input pattern frame by DP matching as described above, and the calculation data is shown in Figures 2 and 3. The vertical axis is the distance between the reference pattern and the input voice pattern and the horizontal axis is the input pattern frame. Figure 2 shows the correct reference pattern, in which a valley is distinct and the frame with the shortest distance corresponds to the last frame of the spoken period of the input voice pattern. Figure 3 shows the incorrect reference pattern, which has no distinct valley and is nearly flat. So, we used a method to select the recognition candidate according as valley exists or not. However, since the distance for the incorrect pattern also tends to appear small, it is difficult to distinguish valleys from flat parts by using a fixed threshold value.

Therefore, to judge the results of recognition, we used an adaptive threshold determined according to the state of the constant part preceding the valley as well as a threshold which checked continuity of the valley.

On the other hand, recognition rate has a tendency to be affected by fluctuating transmission characteristics of the telephone system. So, we employed the spectrum normalizing method using the least square approximation line [5] in order to prevent a decrease of recognition rate.

4. SIMULATION TEST RESULTS

We conducted a preliminary simulation test to confirm that the word spotting method is an effective countermeasure against noise. Figure 4 shows the block diagram of this simulation test system. Speech was analyzed by the speech analyzer, which was followed by recognition processing by the 32-bit mini-computer. Test data was artificially generated with the computer adding the recorded noise to the original speech. The original speech was recorded through a telephone set in an insulated room. The noise was also recorded through a telephone set. The noise conditions were as follows:

(a) Type ........ Sound from things, breathing sound, office noise.
(b) Location ...... Before speech, after speech (each location was not adjacent; office noise was added all through test)
(c) Level .......... Same as speech level (Not fixed for office noise)

The office noise was recorded through a telephone set in an actual office and human voices and telephone bell sounds included.

The recognition conditions were: speaker-dependent, isolated word and vocabulary of 10 numerals. A group of speakers consisting of five males and five females participated in the test. The same test was conducted on a conventional speech recognition system using the method of detecting voiced periods and performing DP matching. Table 1 compares the test results of these two methods. When the speech was accompanied by noise, the speech recognition rate for the conventional method showed a sharp fall, while that for the word spotting method only fell by 2.54% at worst. This fact confirms that the word spotting method is highly immune to noise.
As the preliminary simulation test confirmed the high noise immunity of the word spotting method, we developed a speaker-dependent telephone speech recognition system, based on the algorithm described in Section 3. Figure 5 shows the hardware configuration of the system. This system consists of a speech analysis processor (SPA) (μPD77C25), a DP matching processor (SRP) (μPD7764) and a 16-bit CPU (V50) (μD70216). The DP matching processor is capable of processing 100 words per chip and the CPU can handle up to four DP matching processors. The CPU judges speech recognition results and controls the entire system.

Table 1 Noise Conditions and Recognition Rate (Preliminary Simulation Test Results)

<table>
<thead>
<tr>
<th>Noise Location</th>
<th>Before speech</th>
<th>After speech</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used Method</td>
<td>Noise Type</td>
<td>Sound from things</td>
<td>Breathing sound</td>
</tr>
<tr>
<td>Conventional</td>
<td>7.15%</td>
<td>8.02%</td>
<td>87.51%</td>
</tr>
<tr>
<td>Word Spotting</td>
<td>98.46%</td>
<td>98.0%</td>
<td>99.38%</td>
</tr>
</tbody>
</table>

Note: Office noise is prerecorded in an actual office and includes human voices and telephone bell sounds.

5. SYSTEM CONFIGURATION

We conducted an evaluation test on the speech recognition system described in Section 5. Recognition conditions were: speaker-dependent, isolated word and vocabulary of 50 words. This test was carried out by four speakers. Three types of noises recorded at an actual office through a telephone set were used. As in the preliminary test, we used the speech data produced by a computer adding the recorded noise to the original speech. The original speech was also recorded in an insulated room through a telephone set. Table 4 shows the test results. As shown in this Table, the recognition rate under the worst noisy condition (noise C) is 95.7%, only 2.8% down from the recognition rate under noiseless condition. These test results indicated that this system has very high noise immunity.

6. EVALUATION TEST RESULTS

We also changed the S/N ratio between speech and added noise by using a computer and examined the relationship between this S/N ratio and speech recognition rate. Figure 6 shows the result of this examination. As evident in this figure, there is a definite correlation between the S/N ratio and recognition rate. Recognition rate suffered a sharp drop when the S/N ratio fell below 14 dB. This suggests that, in order to obtain recognition rate as high as 95% or more, the S/N ratio be over 14 dB for white noise and over 15 dB for office noise.
We also conducted a field test in same condition described above, which involved six speakers (three males and three females) speaking through a telephone set in an actual office. The equivalent noise level per hour of the office is 63 dBA. The test results showed an average of over 95%, the same as those obtained in the evaluation test.

Since some speakers were not familiar with the speech recognition system, they sometimes added unexpected utterance to the speech by saying "Well ..." or "Mr ...". Even in this case, the recognition rate remained as high, proving that the system also works with unexperienced speakers.

![Graph](image1)

Fig. 6 (a) S/N Ratio of Input Voice and Recognition Rate (White Noise)

![Graph](image2)

Fig. 6 (b) S/N Ratio of Input Voice and Recognition Rate (Office Noise)

7. CONCLUSION

We confirmed that the word spotting method based on the Blockwise DP matching method can be effectively applied to isolated word recognition system and that this method increases the recognition rate of telephone speech recognition systems which are mostly used in a relatively noisy place like an office.

By using the word spotting method, we developed a speech recognition system compact enough to mount on one card. The evaluation test on this system reveals that the recognition rate of this system is over 98% under noiseless conditions and over 95% in a relatively noisy office.

REFERENCES