ROBUST SPEECH DETECTION METHOD FOR SPEECH RECOGNITION SYSTEM FOR TELECOMMUNICATION NETWORKS AND ITS FIELD TRIAL

Seiichi Yamamoto, Masaki Naito and Shingo Kuroiwa
KDD R&D Laboratories
2-1-15 Ohara Kamifukuoka, 356 Saitama, JAPAN
Tel. +81 492 78 7311, FAX +81 492 78 7512, E-mail: yamamoto@lab.kdd.co.jp

ABSTRACT

Input speech to speech recognition systems may be contaminated not only by various ambient noise but also by various irrelevant sounds generated by users such as coughing, tongue clicking, mouth-noises and certain out-of-task utterances. The authors have developed a speech detection method using the likelihood of partial sentences for detecting task utterance in speech contaminated with these irrelevant sounds. This paper describes this new speech detection method and reports on a field trial of speech recognition systems with the proposed speech detection method.

Keywords: *Telephone speech recognition, Speech detection*

1. INTRODUCTION

Various speech recognition systems (SRSs) are being developed these days. The authors developed a prototype of a continuous speech recognition system for a voice-activated telephone extension service. The system has been in operation at the KDD Head Office with about 3000 branch telephones since December 1995, in order to collect speech data from various users[1, 2, 3].

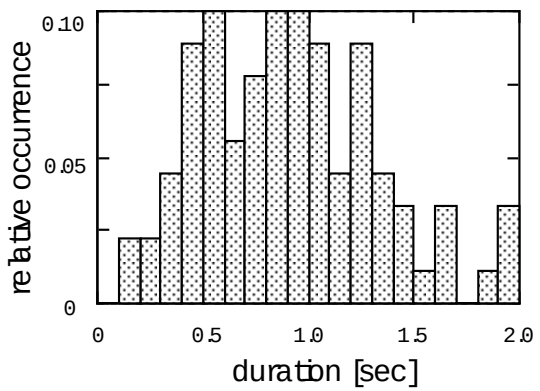

Figure 1: Distribution of Pause Duration Between a Cough and a Task Sentence

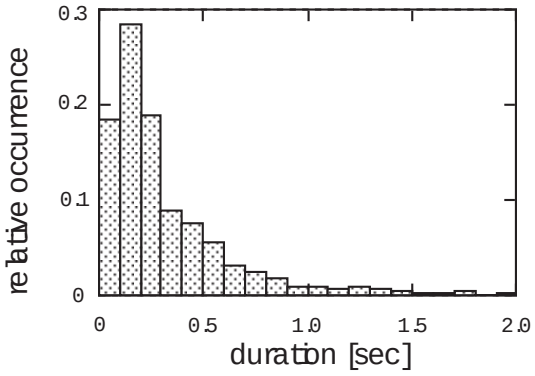

Figure 2: Distribution of Pause Duration Between Phrases

Analysis of speech data collected with the system showed that users behaved in various manners at early stages of the field trial when they were unfamiliar with the system and did not know how the system responds to speech input to the system with out-of-task utterances.

One user behavior which may degrade SRS performance is irrelevant sounds such as coughing, tongue clicking, mouth noises, breathing and certain kinds of out-of-task utterances Some 57 % of misrecognized utterances include such sounds. The analysis results also showed that nearly all such sounds were generated prior to task utterances and a short pause existed between these sounds and the task utterance. Figure 1 shows the distribution of pause duration between a cough and task sentence as a statistical example of a short pause duration. Pauses of various durations however are sometime inserted somewhere between phrases of the task utterance. Figure 2 shows the distribution of pause durations between phrases. Results from these figures show that detection of pause segments, which was commonly used in speech detection methods, may not be suffi-

cient to precisely determine the end of task utterance. The detection of the task utterances is a matter of critical concern for SRS in telecommunication networks, especially since there is always some percentage of users who are unfamiliar with SRS. We therefore propose a speech detection method in order to cope with this problem.

In section 2 of this paper, we describe a new speech detection method. Some experimental results of the new method are given in section 3. Finally, in section 4, we describe an ongoing filed evaluation implemented with SRSs using our proposed method.

2. SPEECH DETECTION METHOD

2.1 End-point Detection

We had already proposed a speech end-point detection method using the likelihood of partial sentence hypothesis to cope with degradation by pauses inserted between phrases[1]. The procedure can be outlined as follows. First of all, the speech recognition process begins as soon as the previous recognition session terminates. Therefore, we do not need to detect the start of the sentence explicitly. Once a recognition session is started, feature extraction, state probability calculation and scoring word sequences are executed in a frame-by-frame manner. Finally, if the best scored word sequence meets the following two conditions, the frame synchronous process is terminated after output of the word sequence as a recognized sentence.

(1) The best sequence can be accepted as a complete sentence by the given grammar.

(2) The final word of the sequence is a /silence/ with duration longer than 400 milliseconds.

The method is robust to changes of speech energy level and pause length. Recognition results showed this method was also effective for detecting endpoints of noisy speech[1, 2].

It is however, still unclear how to detect the start-point of speech contaminated with various ambient noise and irrelevant sounds generated by users.

2.2 Start-point Detection

We improved the speech detection method so as to determine whether an utterance before a pause is a partial sentence or a miscellaneous sound to be disregarded.

The principle of the new method is as follows. The system has two kinds of HMM networks. One is a network defined by task grammar and the other is a network connecting each phoneme model without grammar constraints[4]. The system recognizes input speech frame-synchronously and decides that the present frame is on a segment of a pause when both of the following two conditions are satisfied;

(a) the likelihood of a partial sentence hypothesis (including complete sentence hypothesis) which terminates with a /pause/ or /silence/ has the maximum value in all active cells of the grammar network,

(b) the duration of the stay at the /pause/ or /silence/ is longer than 100 milliseconds.

If the present frame t is on a segment of a pause, the system determines by means of following two conditions whether the utterance before the detected pause is a partial sentence or is an irrelevant sound to be disregarded.

$$\frac{(L(s,t) - L_p(t))}{D(s,t)} \geq Threshold \tag{1}$$

$$0.6 \leq \frac{D(s,t)}{M(s)} \leq 2.0 \tag{2}$$

where, $L(s,t)$ is the likelihood of state s at time t, s is the state of the most likely path at time t, $D(s,t)$ is a duration of speech section which does not include a pause section, $M(s)$ is an expected duration which is calculated from state-transition probabilities of HMMs along the most likely path, $L_p(t)$ is the maximum likelihood of unconstrained phone network, $Threshold$ is the threshold value for a likelihood ratio test, 0.6 and 2.0 are threshold values for duration minimum and maximum respectively. The inequality (1) shows a likelihood ratio condition and the inequality (2) shows a duration condition. When both conditions are satisfied, the partial sentence hypothesis is accepted and the recognition process is continued. On the other hand, if either on these conditions is not satisfied then the system rejects the input before the pause segment and restarts the recognition process from 50 ms before time t.

Figure 3 shows typical speech detection procedure.

3. EVALUATION OF THE SPEECH DETECTION METHOD

In order to evaluate the speech detection method, we installed it a the voice-activated telephone extension system with about 3000 branch phones[1, 2]. The end-point detection was already evaluated in previous papers was used in the following evaluations.

3.1 Preliminary Test for Determining Threshold

To determine the *threshold* for the likelihood ratio condition, we evaluate the speech detection method by inputting 519 individual **coughs** and 200 **task sentences**. Figure 4 shows the false acceptance rate of **coughs** and recognition error rate of **task sentences** as a function of the *threshold*. As shown in

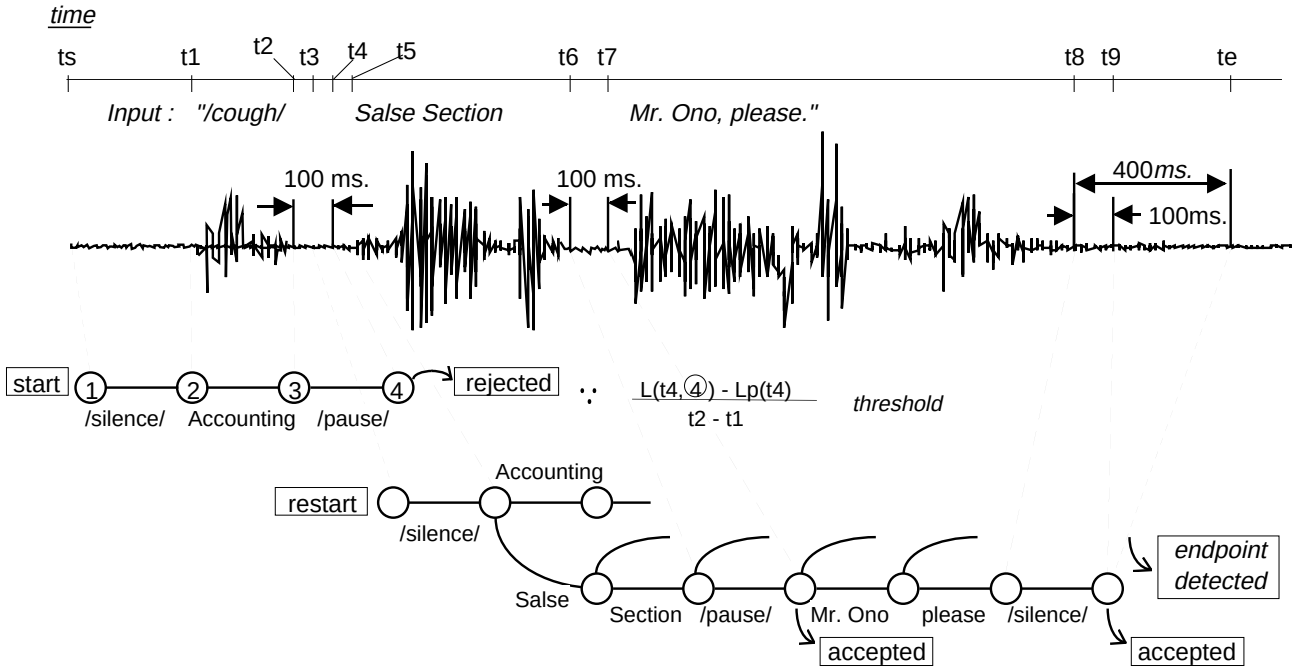

Figure 3: An Example of Speech Detecting Procedure. In this example, partial sentence hypotheses are verified at t_4, t_7 and t_9. At t_4 the best hypothesis is rejected, and the recognition process is then restarted from t_3. Finally, the end-point is detected at t_e, because the best hypothesis become a final grammar state and the final word of sequence is a /silence/ with a duration longer than 400 milliseconds.

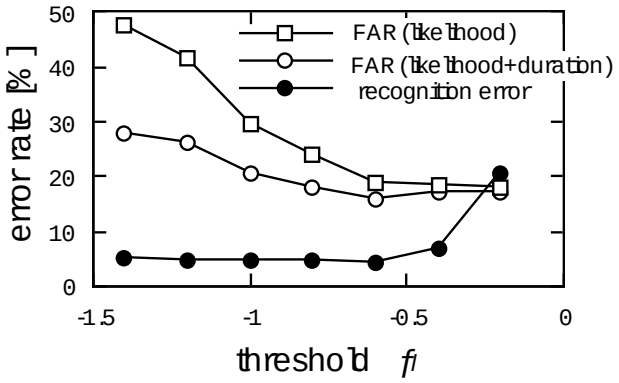

Figure 4: False Acceptance Rate and Recognition Error Rate as a Function of Threshold

the figure, when the *threshold* is *-0.6*, a total of 84 % of coughs were rejected without increasing the recognition error. In the following evaluation we use *-0.6* as the *threshold*.

3.2 Evaluation Test

In order to evaluate the start-point detection method, we drew several irrelevant sounds from data collected in the field trial. These sounds were classified into 13 categories – namely, coughs, inhaling, exhaling, sniffs, mouth clicks, background speech, ringing, pulses as line noise, push tone, sounds of shutting a door, footsteps sounds, paper noise and key click sounds. We constructed 2600 test sentences which included irrelevant sounds prior to the speech by patching these 13 kinds of sounds in prior to the 200 task utterances.

Table 1 shows evaluation results of recognition accuracy. As shown in the table, the proposed start-point detection method reduced errors by 64% using both the likelihood ratio condition and the duration condition.

4. FIELD TRIALS

We first applied the method to a voice-activated telephone extension system in operation at the KDD Head Office with a net of about 3000 branch telephones. Almost all users of the system however, were already familiar with the system and we observed little significant effect from this method. Degradation was not observed among these familiar users of the system who did not utter the irrelevant sounds.

We then applied the method to a new large vocabulary continuous SRS which accepts a user questions about time difference, area codes and country codes for 1500 major areas throughout the world. The

Table 1: Recognition Accuracy for Speech Data with Various Irrelevant Sounds

Kind of Sounds	Rejection Condition and Accuracy Without	Likelihood	+Duration
Without	95.0%	95.0%	95.0%
Coughs	57.5%	80.5%	84.5%
Inhaling	71.5%	77.5%	89.0%
Exhaling	75.0%	83.5%	87.5%
Sniffs	55.0%	75.5%	83.0%
Mouth clicks	85.0%	89.5%	91.5%
Speech	49.0%	81.0%	87.5%
Ringing	49.5%	71.5%	81.0%
Pulses	62.0%	77.5%	83.5%
Tones	53.5%	86.0%	88.0%
Door	73.0%	83.5%	89.5%
Footsteps	52.0%	75.0%	81.5%
Paper noises	52.5%	77.5%	88.0%
Key clicks	58.0%	76.5%	83.5%
Average	61.0%	79.6%	86.0%

Table 2: Recognition Accuracy in the Field Trial

	Without sounds	With sounds
Conventional	96 %	37 %
Proposed	95 %	82 %

recognition scheme (such as grammar-constraints and HMMs) was nearly the same as the voice-activated telephone extension system. The system has been in operation at KDD since May 1997 for trial use by a limited number of users prior to full deployment. In the first week of operation we collected approximately 100 utterances which included irrelevant sounds prior to the task speech. These irrelevant sounds consisted of coughs, out-of-task speech (in particular, "moshi moshi", which is a greeting in Japanese) and fillers which frequently appeared. The system performance in table 2 shows that the recognition accuracy is 95 % for utterances without irrelevant sounds, and 82 % for utterances with irrelevant sounds. In order to confirm the efficiency of the start-point detection method, we applied the conventional method which uses only the end-point detection method, to the same collected data. Results shows that the recognition accuracy was 96 % for utterances without irrelevant sounds and 37 % for utterances with irrelevant sounds. Our proposed start-point detection method therefore reduces errors from irrelevant sounds by 71 %.

4. SUMMARY

In this paper, we proposed a speech detection method using the likelihood of partial sentence hypothesis. This method yields better performance for detection not only at the end-point of speech but also at the start-point of a task utterance when the input speech contains various ambient noises, various sounds generated by users, and pauses between phrases. The detection method was applied to some SRSs and verified to obtain superior accuracy in the field trial. This result may depend on the kind of task and may not be applied to more complicated dialogue systems. This result can however, be adopted to many tasks which are capable of being processed with current speech recognition technology.

Acknowledgment

The authors are grateful to Dr. Murakami, Director of KDD R&D Laboratories for his continuous support of this work. The authors are also grateful to the members of the laboratories for their discussions.

References

[1] K.Takeda, S.Kuroiwa, M.Naito, S.Yamamoto: "Top-Down Speech Detection and N-Best Meaning Search in a Voice Activated Telephone Extension System", EuroSpeech'95, Vol.2, pp. 1075–1078, 1995.

[2] M.Naito, S.Kuroiwa, K.Takeda, S.Yamamoto: "A Real-Time Speech Dialogue System for a Voice Activated Telephone Extension Service", ESCA Workshop on Spoken Dialogue Systems, pp.129–132, 1995

[3] S.Kuroiwa, K.Takeda, M.Naito, N.Inoue and S.Yamamoto: "Error Analysis of Field Trial Results of a Spoken Dialogue System for Telecommunications Applications", IEICE Trans. Inf.&Sys., Vol. E78-D, No.6, pp. 636–641, 1995.6

[4] T.Watanabe and S.Tsukada: "Unknown Utterance Rejection Using Likelihood Normalization Based on Syllable Recognition", IEICE Trans. Vol. D75-D-II, No.12, 1995.12