CONTINUOUS SPEECH RECOGNITION WITHOUT END-POINT DETECTION

Osamu SEGAWA[†‡], *Kazuya TAKEDA*[†] *and Fumitada ITAKURA*[†]

†Graduate School of Engineering, Nagoya University
Furo-cho Chikusa-ku NAGOYA 464-8603 JAPAN

‡Chubu Electric Power Co., Inc.
Odaka-cho Midori-ku NAGOYA 459-8522 JAPAN

ABSTRACT

A new continuous speech recognition method that does not need the explicit speech end-point detection is proposed. A one-pass decoding algorithm is modified to decode the input speech of infinite length so that, with appropriate non-speech models for silence and ambient noises, continuous speech recognition can be executed without the explicit end-point detection. The basic algorithm is 1) decode a processing block of the predetermined length, 2) traceback and find the boundaries of the processing blocks where the word history in the preceding processing block is merged into one, and 3) restart decoding from the boundary frame with the merged word history. The effectiveness of the method is verified by the two dictating experiments. With consecutive 100 sentences of utterances from a newspaper, the degradation of the recognition accuracy due to the modification of the decoder is about 5% compared with the results when the correct end-point is given. With a 30 minutes dialogue in a moving car, 75 %correct and 69 %accuracy score is obtained.

1. INTRODUCTION

Speech end-point detection is a simple but crucial issue in any type of speech recognition system because errors in the end-point detection cannot be recovered in the later stages of recognition. Therefore, various research efforts have been reported on the speech detection [1][2]. The typical end-point detection method utilizes energy contour and zero crossing counts as the measure for discriminating the speech signal from silence. However, in the real environments where various environmental sounds exist, discriminating speech and non-speech sounds is difficult. One method to improve the robustness of the speech end-point detection is to "recognize" the non-speech segment as well as the speech segments [3][4]. In that method, speech segment is extracted with the margins, i.e., preceding and following non-speech periods, in the pre-processing stage so that decoder is executed in embedded mode. The approach that uses the state probabilities at the particular grammar nodes, typically at the end of the sentences, is also reported [7]. From the decoder's viewpoint, on the other hand, speech end-point detection is indispensable because current speech recognizers assume that the input signal is limited in length. Therefore, in this paper, we will discuss the speech recognizer that can decode an infinite length input signal to overcome the difficulty in speech end-point detection. Since, in the proposed speech recognizer, end-point detection in the conventional systems is replaced with the continuous decoding process of the intervals between utterances by the non-speech acoustic models, explicit end-point detection is no longer needed. Therefore, the speech recognizer can be applied to such systems that continuously monitor the dialogues in the real environment.

In the rest of this paper, we will describe the details of the algorithm in Section 2. The experiments on the newspaper dictation of consecutive 100 sentences will be reported in Section 3. In Section 4 and 5, the experiments on recognizing 30 minutes long dialogue in a driving car will be described. Section 6 summarizes this work.

2. ALGORITHM

In the proposed method, an input speech stream is divided into a processing block (segment) of a predetermined length without discrimination between speech and non-speech (noise or silence). Since the segments are of fixed length, the segments may end in the middle of words (or phoneme). In our method, the decoder estimates the nearest optimal word boundary frame and backtracks to the boundary frame and then starts decoding of the next segment from that point. The algorithm is described below.

First, divide the input speech stream into segments of predetermined length T. From the beginning of segment, perform the frame-synchronous One-Pass-Viterbi beam search with bigram language model. For each frame, the search information (word trellis index[8]) is stored. Word trellis index is a set of survived word-end nodes in beam search,

their likelihood scores and their corresponding start frames.

When the search process has reached to the final frame of a segment N then the procedure given below is performed.

1. For each word-end node survived in the beam search at the last frame, traceback the best pass in the word trellis index and get multiple sentence hypotheses.

2. Find the frame ($t_{boundary}$ in Fig.2) at which all the sentence hypotheses are merged into the best likelihood score pass and estimate the frame as the nearest optimal word boundary. If the optimal boundary frame can be found then output the sequence of words before the fixed last word (W_{last} in Fig.2). Otherwise, output the sentence hypothesis with the best likelihood score from the last frame.

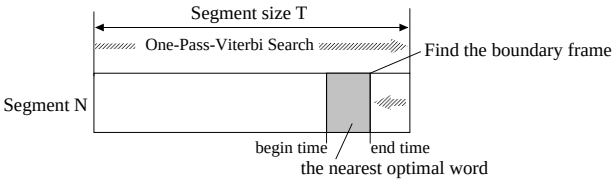

Fig. 1. Search process in a segment.

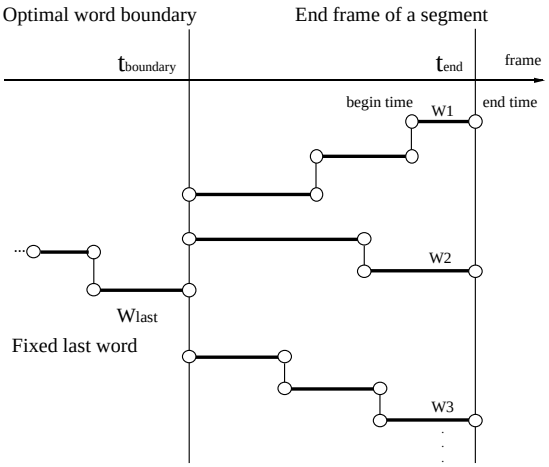

Fig. 2. The nearest optimal word boundary before the end of a segment.

3. At the next segment $N+1$, the beginning time of the fixed last word in segment N is set to the start frame. Then, restart decoding at the frame at Δt before the end of segment N. At the same time, the ID of the fixed last word is propagated to the next segment $N+1$ for using bigram constraint effectively. This step is shown in Fig.3

3. EVALUATION OF BASIC PERFORMANCE

3.1. Experimental Setup

Specifications of the conditions are shown in Table.1.

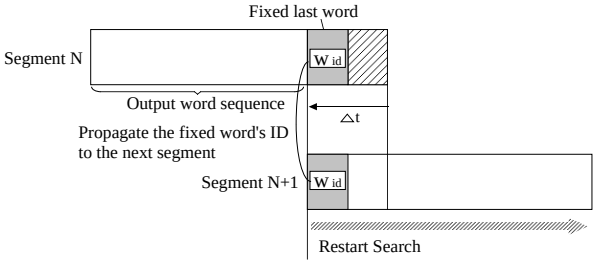

Fig. 3. Search process in the next segment.

Table 1. Experimental conditions in the basic performance evaluation.

Acoustic analysis	Sampled at 16kHz and 16bit Pre-emphasis 0.97 Hamming window Analysis frame length 25ms Frame shift 10ms MFCC (12th)+ΔMFCC (12th) +Δ power
Acoustic model	Phonetic tied-mixture HMM (male speaker model)
Language model	20K vocabulary bigram

As for acoustic model, we have prepared a continuous density context-dependent HMM, that is, phonetic tied-mixture model [9]. The model is trained with the ASJ speech database of phonetically balanced sentences and the JNAS newspaper article utterances, in total 20K sentences by 132 speakers for each gender. As for language model, we have prepared a 20K vocabulary bigram. Training corpus is the JNAS newspaper article texts ('91/1–'97/6) [10].

The evaluation data is a part of JNAS Mainich-newspaper speech corpus. We have selected 100 sentences by 23 male speakers. To prepare the evaluation data, we have connected these sentences into a consecutive utterance. The data length is about 583 seconds (58385 frames).

3.2. Experimental Results

In the experiment, the segment size is varied from 200 to 500 frames. Fig.4 shows recognition results (%accuracy) of the proposed method ("Proposed method" in the graph). For the purpose of comparison, baseline performance of the one-pass decoder (JULIUS[8] one-pass search) is also given ("Baseline" in the graph). In the evaluation of the baseline performance, the input data is the same 100 sentences, but the correct end-point is given. Furthermore, the recognition results of the baseline decoder in which the input data is speech segments of fixed length are given ("Not using proposed method" in the graph). Fig.5 shows the recognition error rate (word substitution, insertion and deletion error).

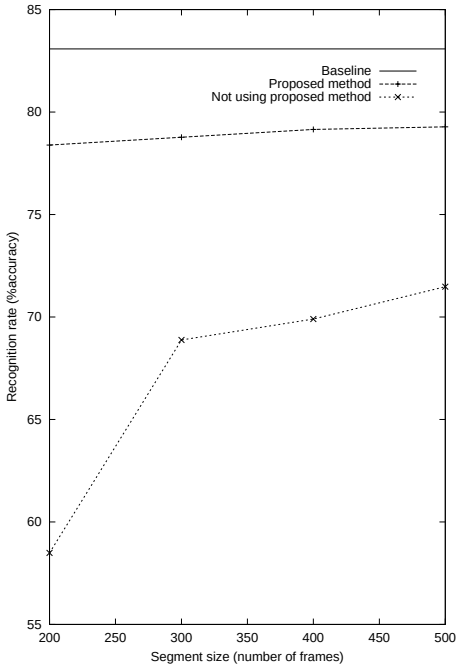

Fig. 4. Evaluation of basic performance with consecutive 100 sentences of utterances from a newspaper.

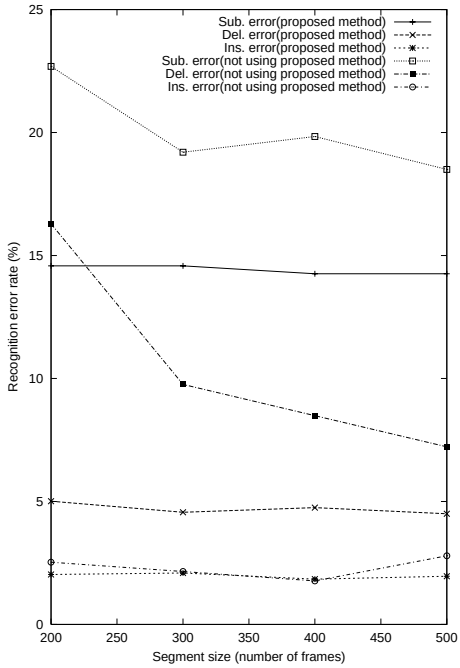

Fig. 5. Recognition error analysis in the basic performance evaluation.

3.3. Discussion

Fig.4 shows that the degradation of the recognition accuracy due to the modification of the decoder is about 5% comparing with the results of baseline (end-point is given). Compared with the result of not using the proposed method, there is an 8 to 20 point increase in %accuracy score. These results show that missing words are repaired appropriately by the proposed method. Especially, in error analysis (Fig.5), word deletion and substitution error are reduced. The latter effect is mainly due to bigram constraint between segments. In respect of segment size, recognition rate does not fall almost until 200 frames. However, less than 100 frames length, recognition rate is decreasing because a search of word boundary would be unstable.

4. DICTATION OF SPOKEN DIALOGUE

For the purpose of evaluating the proposed method in real environment, we have planned a system of spoken dialogue dictation. In building a large scale spoken dialogue corpus, the cost of speech-to-sentence transcription is very high. If this work could be done automatically, the system would be very effective not only for transcription but also for automatic tagging of database.

The next section reports an initial experiment of spoken dialogue dictation. The task domain is car navigation and information retrieval of restaurant and shops, in CIAIR [1] database project [11].

5. EXPERIMENTS

5.1. Experimental Setup

Conditions for the experiment are shown in Table.2. Acoustic analysis feature is the same as Table.1.

Table 2. Experimental conditions of the spoken dialogue dictation.

Acoustic model	Phonetic tied-mixture HMM (gender independent model)
Language model	2K vocabulary class bigram

As for acoustic model, we have prepared a gender independent phonetic tied-mixture HMM for the multiple speaker task. As for language model, we have prepared a class bigram which is estimated by spoken dialogue. Training corpus is a part of CIAIR car spoken dialogue database transcription texts consisting of 3424 sentences (task open). In class bigram estimation, we have used the probability definition as follows

$$P(W_n|W_{n-1}) = P(W_n|C_n)P(C_n|C_{n-1})$$

where C_n is part-of-speech category. Testset perplexity is 29.25, out-of-vocabulary rate is 1.5%.

[1] Center for Integrated Acoustic Information Research, Nagoya univ.

In dictation experiment, we have used a part of CIAIR car spoken dialogue corpus. The evaluation data was recorded with a close-talking microphone and its length is about 30 minutes (181939 frames). Speakers are one male (driver) and one female (playing navigation system role) and the total number of sentences is 230 (male 79, female 151).

5.2. Experimental Results and Discussion

In the experiment, the segment size is set to 400 frames. Table 3 shows the recognition rate (%correct and %accuracy) of each speaker and the total (male+female).

Table 3. Experimental results of spoken dialogue dictation.

Speaker	% Correct	% Accuracy
Male (driver)	69.72	66.42
Female (system)	78.64	70.60
Total	75.44	69.10

The results of this experiment show that it is possible to use the proposed system for continuously monitoring the dialogue in a real environment. It also shows that the system can handle utterances from multiple speakers without any pre-processing of the input speech. In future work, optimization of the acoustic and language models for the task domain and search algorithm with trigram models would be explored to improve the performance of the system.

6. SUMMARY

A new continuous speech recognition method that does not need the explicit speech end-point detection is proposed. A one-pass decoding algorithm is modified to decode the input speech of infinite length so that, with appropriate non-speech models for silence and ambient noises, continuous speech recognition can be executed without the explicit end-point detection. This method is a robust decoding approach because it does not need discrimination between speech and non-speech and does not consider any utterance unit. The effectiveness of the method is verified by the two dictating experiments, with the newspaper consecutive 100 sentential utterances and with a 30-minute dialogue in a moving car.

Acknowledgement: This work has been partially supported by Grant-in-Aid for COE Research (No. 11CE2005).

7. REFERENCES

[1] L.R.Rabiner and M.R.Sambour. “An algorithm for determining the endpoints of isolated utterances”, The Bell System Technical Journal, vol.54, no.2, pp.297-315, Feb 1975.

[2] J.G.Wilpon, L.R.Rabiner and T.Martin. “An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints”, AT&T Bell Laboratories Technical Journal, vol.63, no.3, pp.479-498, Mar 1984.

[3] J.G.Wilpon and L.R.Rabiner. “Application of hidden Markov models to automatic speech endpoint detection”, Computer Speech and Language, vol.2 pp.321-341, 1987.

[4] A.Acero. “Robust HMM-based endpoint detector”, In Proc. European Conference on Speech Communication Technology, pp.1551-1554, 1993.

[5] J.C.Junqua, B.Mak and B.Reaves. “A robust algorithm for word boundary detection in the presense of noise”, IEEE Trans. on Speech and Audio Processing, 2(3) pp.406-412, 1994.

[6] B.Mak, J.C.Junqua and B.Reaves. “A robust speech/non-speech detection alogorithm using time and frequency-based features”, In Proc. ICASSP, pp.269-272, 1992.

[7] K.Takeda, S.Kuroiwa, M.Naito and S.Yamamoto. “Top-down Speech Detection and NBest Meaning Search in a Voice Activated Telephone Extension System”, In Proc. European Conference on Speech Communication Technology, pp.1075-1078, Sep 1995.

[8] A.Lee, T.Kawahara and S.Doshita. “An efficient two-pass search algorithm using word trellis index”, In Proc. ICSLP, pp.1831-1834, 1998.

[9] A.Lee, T.Kawahara, K.Takeda and K.Shikano. “A new phonetic tied-mixture model for efficient decoding”, In Proc. ICASSP, pp.1269-1272, 2000.

[10] K.Itou, M.Yamamoto, K.Takeda, T.Takezawa, T.Matsuoka, T.Kobayashi, T.Utsuro and K.Shikano. “The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus”, In Proc. ICSLP, pp.3261-3264, 1998.

[11] K.Kawaguchi, S.Matsubara, H.Iwa, S.Kajita, K.Takeda, F.Itakura and Y.Inagaki. “Construction of speech corpus in moving car environment”, In Proc. ICSLP, pp.362-365, 2000.