AUTOMATIC DETECTION OF TOPIC BOUNDARIES AND KEYWORDS IN ARBITRARY SPEECH USING INCREMENTAL REFERENCE INTERVAL-FREE CONTINUOUS DP

*Jiro Kiyama** *Yoshiaki Itoh*† *Ryuichi Oka*

Tsukuba Research Center, Real World Computing Partnership
1-6-1 Takezono, Tsukuba-shi, Ibaraki 305 Japan
kiyama@iml.mkhar.sharp.co.jp, itoh@system.kawasaki-steel.co.jp, oka@trc.rwcp.or.jp

ABSTRACT

We propose a new approach for detecting topic boundaries and keywords in arbitrary speech, with neither recognition nor prosodic processing, aiming at quick access to the content of recorded raw speech. This approach is based on the general tendency that frequently-repeated phrases/words in speech are characteristic of topics in discourse, so it uses pairs of phonetically similar segments (PPSSs) of speech to represent topics in speech. This approach has the advantage of being domain and language-independent and robust against variations in the speaker and background noise, as it needs neither a language nor acoustic model in advance. Experiments using simulated dialogues confirmed the good performance of this approach. We also propose Incremental Reference Interval-free Continuous Dynamic Programming (IRIFCDP) as an algorithm for detecting PPSSs in speech for the above method. IRIFCDP can detect PPSSs efficiently in synchronization with the speech, so it is suitable for handling long speech samples.

1. INTRODUCTION

The retrieval, manipulation, and reuse of recorded raw speech is difficult. It takes a lot of time to access a desired part of recorded speech, particularly if the speech is long. If topic boundaries and keywords in arbitrary speech could be automatically detected, it would be possible to skip over irrelevant topics, and construct an overview of the speech content more efficiently, thus enabling quick access to the desired topic.

For this purpose, we propose a new approach based on repetitions in speech. This approach has the advantage of being domain and language-independent and robust against variations in the speaker and background noise, as it needs neither a language nor acoustic model in advance. In investigating the acquisition of useful information from speech without linguistic knowledge, there are methods for segmenting speech according to emphasis using pauses [1] or prosodic information [2], but none have so far used repetition.

We also propose an algorithm called Incremental Reference Interval-free Continuous Dynamic Programming (IRIFCDP) for detecting pairs of phonetically similar segments (PPSSs) in speech for use in the above method. There has been some research on the detection of similar segments in speech [3, 4]. Those techniques, however, are all for detecting segments that are similar to two independent speech samples; they are not intended for use in detecting similar segments within one discourse sample, as needed here. Moreover, these conventional methods are also inappropriate for handling long speech samples, which is necessary in order to implement the above-mentioned functions.

In the rest of this paper, Section 2 explains the formalization and performance evaluation of IRIFCDP and Section 3 describes our approach and methods of detecting keywords and topic boundaries.

2. INCREMENTAL REFERENCE INTERVAL-FREE CONTINUOUS DP (IRIFCDP)

2.1. Algorithm

Consider a sequence of speech feature vectors $X = (x_1, x_2, ..., x_N)$. We perform starting point-free matching of any arbitrary segment $X_i = (x_{\tau_s}, ..., x_{\tau_e})$, $1 \leq \tau_s < \tau_e \leq N$, with X. Here we denote the matching distance between X and X_i as $G(t, \tau_s, \tau_e)$ at frame t. We treat the segment pairs whose matching distances are below a threshold value as PPSSs. As the combination of τ_s and τ_e is enormous, we need an efficient algorithm for calculating $G(t, \tau_s, \tau_e)$.

For easy comprehension, we first explain the matching by Reference Interval-free Continuous DP (RIFCDP) [3] on which IRIFCDP is based. RIFCDP is an efficient matching algorithm between arbitrary section in a reference pattern and arbitrary section in an input pattern. In this case RIFCDP can detect PPSSs by using an input pattern as a reference.

RIFCDP is an extension of Continuous DP (CDP), which is used for spotting words or sentences in spontaneous speech [5]. CDP can efficiently give the optimal path and the accumulated distance $G(t, 1, \tau_e)$ for $1 \leq \tau_e$ by starting point-free matching of arbitrary segment, $x_1, ..., x_{\tau_e}$ with X (Figure

*At present, Sharp Corporation
†At present, Kawasaki Steel Corporation

1). The calculation of $G(t, \tau_s, \tau_e)$ for $1 \leq \tau_s \leq \tau_e$, however, is not shared, so CDP needs a lot of calculation for matching of arbitrary partial segment.

Here we assume that we can use the partial path (solid curve in Figure 1) that corresponds to the section $[\tau_s..\tau_e]$ of the optimal path for the segment $x_1, ..., x_{\tau_e}$ obtained by CDP instead of the optimal path for the segment $x_{\tau_s}, ..., x_{\tau_e}$ (dotted curve in Figure 1). Under this assumption we can approximately obtain $G(t, \tau_s, \tau_e)$ at (t, τ_e) by $G(t, 1, \tau_e) - P(t, \tau_e, \tau_s)$. $P(t, \tau_e, \tau_s)$ is the accumulated distance at τ_s, $1 \leq \tau_s \leq \tau_e$, along the optimal path to the grid (t, τ_e), so $P(t, \tau_e, \tau_e)$ is identical to $G(t, 1, \tau_e)$. RIFCDP gives $G(t, \tau_s, \tau_e)$ by keeping the history of accumulated distance, $P(t, \tau_e, \tau_e)$ for $1 \leq \tau_s \leq \tau_e$ along the optimal path to each grid. We confirmed that this approximation is applicable from experimental results. See Itoh [3] for detail.

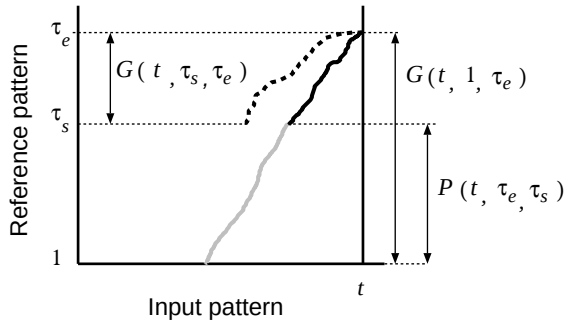

Figure 1: Principle of RIFCDP. The dotted and solid curves respectively indicate the optimal path for $[\tau_s..\tau_e]$ and the partial path of the optimal path for the section $[1..\tau_e]$.

Although the above approach can efficiently calculate $G(t, \tau_s, \tau_e)$, it requires a lot of memory to keep the history for a long input and cannot perform the detection simultaneously with speech input, because the patterns and calculation range of RIFCDP are static.

To overcome this limitation we developed Incremental RIFCDP (IRIFCDP) by extending RIFCDP. The feature of IRIFCDP is a dynamic extension of reference pattern, which is performed by adding the most recent input frame to the end of reference in synchronization with the input frame. This extension enables IRIFCDP to detect PPSSs simultaneously with speech input. The calculation range is extended in proportion to the reference extension. For example, the calculation range becomes $[1..t-D]$ at frame t, where D is a small positive value that is necessary to avoid the matching with segment itself, which is meaningless.

Note that IRIFCDP can also detect PPSSs among two independent input streams by adding the most recent frame from one input stream to the reference and using the another as the input stream.

We also introduce a calculation range restriction into IRIFCDP for efficiency. IRIFCDP can restrict the calculation range to within a certain vicinity $[t-D-W..t-D]$ at frame t, where W means the length of the calculation window (Figure 2). When the range is restricted, the computation amount becomes an order of $O(I*W)$, where I represents the length of input speech, so the amount of processing for one input frame becomes constant regardless of the length of the speech. In this case, the maximum required memory capacity is the constant value of $O(W)$. This enables IRIFCDP to be suitable for long speech in which keywords and topic boundaries detection are required. Although this restriction causes some detection loss for PPSSs whose intervals are long, most PPSSs can be covered with an appropriate W, since important PPSSs appear locally in many cases. The only effect of extending the algorithm of RIFCDP is a modification of the local path calculation on the grid $(t, t-D)$ and $(t, t-D-W)$ at frame t.

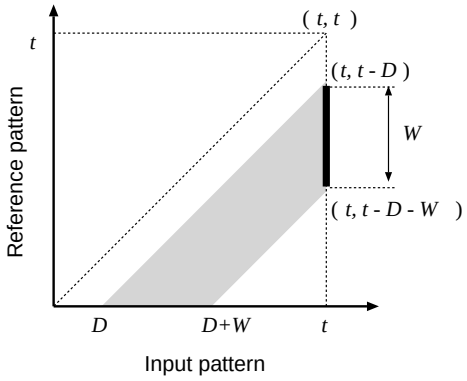

Figure 2: Restriction of calculation range. The thick vertical line and the lightly-hatched region indicate calculation range at frame t and search area respectively.

2.2. Performance evaluation

We investigated the performance of PPSS detection by IRIFCDP. As speech samples, we used a simulated Japanese dialogue spoken by one male speaker (84.4 seconds) [7]. The calculation range was the entire sample. Here, Blurred Vector Field of Spectrum [6] was used as a feature vector in the analysis, with a sampling frequency of 15 kHz, a frame period of 7.5 ms, and a frame length of 16 ms. The minimum and maximum lengths of PPSSs were set to 413 ms and 750 ms respectively. The former was used to eliminate short PPSSs, which are meaningless, and latter to reduce the required computing power.

The detection rate (DR) and the detection quality (DQ) were used as evaluation indices. DR is the time-length proportion of the segments that were actually detected as PPSSs relative to all the correct segments; DQ is the time-length proportion of the correctly detected segments relative to all of the segments that were detected. We want both these indices to have high values. The correct segments consisted of at least four syllables that occurred more than once within the sample.

The best results gave DR of 96.2% and DQ of 4.6%. The

low value for DQ was due to pause segments being detected as PPSSs, since the background noise of the test sample was uniform. However, pauses can easily be eliminated on the basis of segment power.

3. KEYWORD AND TOPIC BOUNDARY DETECTION

3.1. Approach

Frequently-repeated phrases/words in speech tend to be characteristic of topics in discourse. For example, there were some repetitions in the first paragraph of Section 1, in which the words "recorded", "speech", "topic", and "desired" reflect the topic. These segments are very useful in indicating content and in distinguishing topics. This is believed to be true for speech as well, so pairs of phonetically similar segments (PPSSs) of speech can be used to represent topics in speech.

This approach does not employ linguistic knowledge, such as a lexicon, and so it has the advantage of being domain- and language-independent. This also indicates that this method is not limited to speech. Moreover, it needs no prior assumptions about the acoustic properties of the input, so it is robust against variations in the speaker and background noise.

On the other hand, this approach has the inherent shortcoming that the quality of detection is affected by the frequency of appearance of keyword repetition in the speech samples. This approach is also weak at dealing with speech that includes voices from multiple speakers such as conversation, since it is difficult to detect PPSSs among different speakers as long as without speech recognition.

3.2. Keyword Detection

Under the above assumption that PPSSs of speech can be used to represent topics in speech, we can treat extracted PPSSs themselves as keywords of topics.

The procedure of keyword detection is as follows: (1) detect PPSSs by IRIFCDP, (2) remove PPSSs that correspond to pause segments on the basis of segment power, and (3) select segments that correspond to one of each of the PPSSs as keyword segment candidates. If any segments overlap in the input speech, they are merged.

We conducted an experiment to measure the performance of this method. The experimental conditions were the same as for the above experiment. To evaluate keyword detection, we introduced two evaluation indices, the shortening rate and keyword detection rate, which respectively represent the ratio of keyword segment candidates to the original speech and the ratio of keywords candidates to correct keywords. We used repeated content words which consist of at least four syllables in the speech as the correct keywords. In this experiment the ratio of correct keyword segment to the original speech was 11.9%.

Figure 3 shows the result, where P represents the power threshold value for pause removal. This curve was obtained by varying the threshold for whether the segment pair were similar or not. This figure indicates that the content of speech can be roughly grasped from 20% of the original speech.

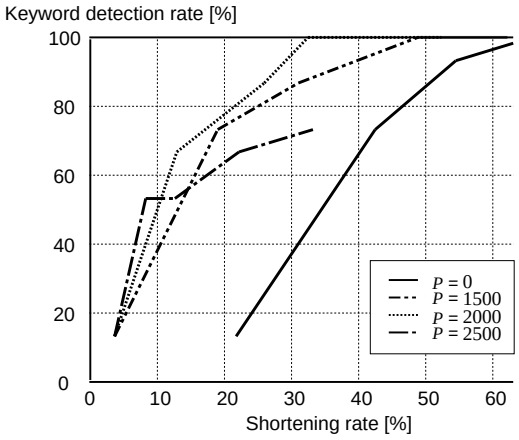

Figure 3: Result of keyword detection.

3.3. Topic Boundary Detection

The basic idea of the detection method is as follows. We consider speech in which two different topics are spoken in sequence. When PPSSs are detected in such speech, two cases can be considered: a) repetition that extends across topics and b) repetition within a topic. In the former case, the repetitions extend across a topic boundary to involve two topics; in the latter, the repetition occurs within a single topic. If repetitions are characteristic of the topic, the second case will occur more frequently than the first. Conversely, time periods that are straddled by few pairs of repeated segments are topic boundaries.

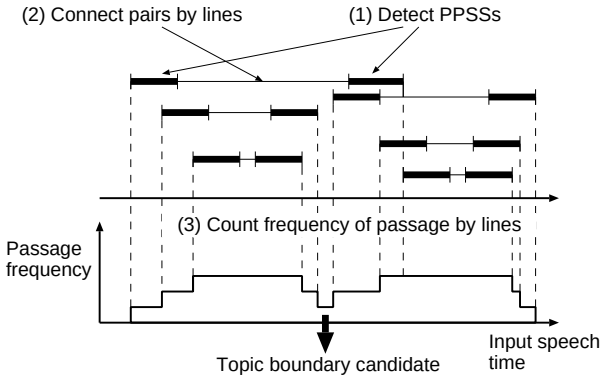

Figure 4: Procedure of automatic topic boundary detection.

The procedure for this method is as follows. (1) First, PPSSs

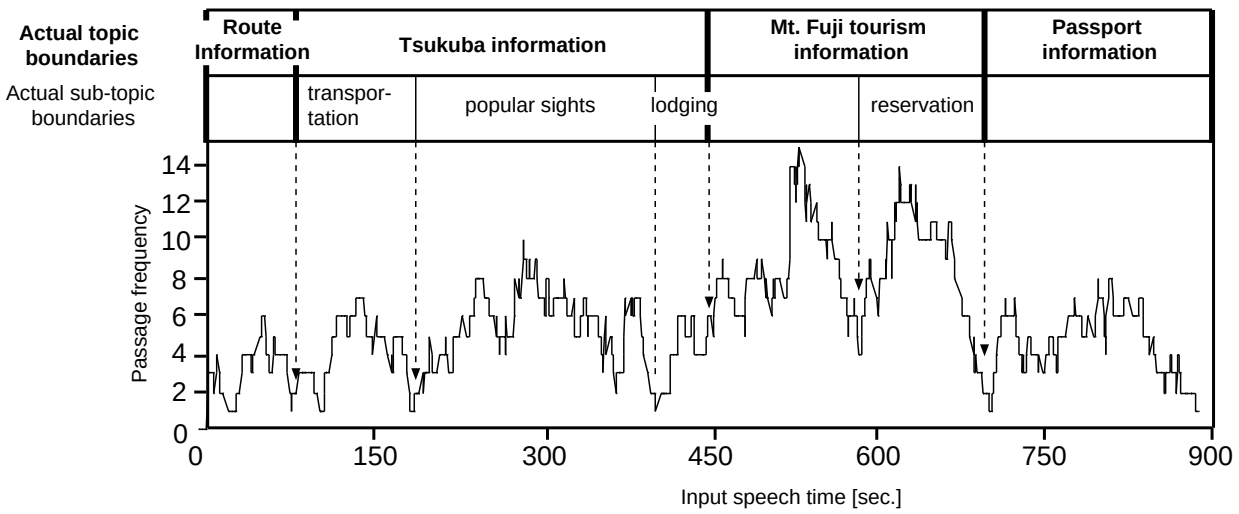

Figure 5: Result of topic boundary detection. Global dips indicate topic boundary candidates.

are detected by IRIFCDP. (2) Then, the respective PPSSs are connected by lines. (3) The number of connecting lines that pass through each time period is determined and the resulting data is plotted in a graph, as shown in Figure 4. Dips in the graph are taken as candidates for topic boundaries. This method is not limited to speech; it is also applicable for text.

To determine the effectiveness of the proposed method, we performed the following experiments. The speech samples we used were constructed by connecting four-topic simulated dialogues uttered by one male speaker [7]. The total length of the speech sample was approximately 900 seconds. The calculation range was set to 60 seconds.

The experimental results are shown in Figure 5. The actual topic boundaries are shown above the figure, which are labeled with the topic names. We can see from the figure that dips in the graph occur near the topic boundaries. Most of the clear dips that appear in places other than the topic boundaries are considered to indicate sub-topic boundaries. For example, the dip that appears near 187 seconds lies within the topic "Tsukuba Information", but it divides that topic into two sub-topics: the part that precedes that dip is a sub-topic concerning "Transportation to Tsukuba"; the part that follows the dip is the sub-topic "Popular Sights in Tsukuba." The other dips also generally correspond to sub-topic boundaries in a similar fashion. Thus, we conclude that the proposed method is promising for detecting the boundaries of topics in spoken discourse.

4. SUMMARY

Our new approach for detecting topic boundaries and keywords, based on the general tendency that repetitions are characteristic of topics, has the advantage of being domain and language-independent and robust against variations in the speaker and background noise. Experimental results showed that this approach is very promising. We also proposed Incremental Reference Interval-free Continuous Dynamic Programming as an algorithm for detecting pairs of phonetically similar segments in speech for the above method. This can detect phonetically similar segments in speech efficiently in synchronization with the speech input, so it is suitable for long speech samples.

ACKNOWLEDGMENT

We would like to thank Dr. Shimada, the director of the Institute of the Real World Computing Partnership for his support of our research.

5. REFERENCES

1. B. Arons, "SpeechSkimmer: Interactively Skimming Recorded Speech", Proc. UIST '93, pp.187-196, 1993.
2. F. R. Chen and M. Withgott, "The Use of Emphasis to Automatically Summarize a Spoken Discourse", Proc. ICASSP-92, pp.I–229–232, 1992.
3. Y. Itoh, et al., "A Proposal for a New Algorithm of Reference Interval-free Continuous DP for Real-time Speech or Text Retrieval", Proc. ICSLP'96, 1996.
4. Y. Hyogo and S. Nakagawa, "Extraction of similar speech patterns by DP-matching in a pair of sentences uttered continuously", IEICE meeting, A-22, 1989 (in Japanese).
5. J. Kiyama, et al., "Spontaneous speech recognition by sentence spotting", Proc. EUROSPEECH '93, pp. 1053–1056, 1993.
6. R. Oka and H. Matumura, "Speaker Independent Word Speech Recognition using the Blurred Orientation Pattern Obtained from the Vector Field of Spectrum", Proc. IJCPR, November, 1988.
7. The Acoustical Society of Japan, "Continuous Speech Corpus for Research", 1992.