A Proposal for a New Algorithm of Reference Interval-free Continuous DP for Real-time Speech or Text Retrieval

Yoshiaki ITOH† Jiro KIYAMA‡ Hiroshi KOJIMA§ Susumu SEKI‡nd Ryuichi OKA
Real World Computing Partnership
1-6-1 Takezonno, Tsukuba, Ibaraki, 305 Japan, E-mail: itoh@trc.rwcp.or.jp

ABSTRACT

This paper proposes a new frame-synchronous algorithm for spotting similar intervals by comparing arbitrary intervals in a reference pattern sequence with arbitrary intervals in an input pattern sequence. The algorithm is called Reference Interval-free Continuous DP (RIFCDP) and the experimental results show that RIFCDP is successful in detecting the similar intervals between a reference pattern and an input. We have applied this algorithm to speech retrieval from a speech database and showed the possibility of real-time speech/text retrieval. The proposed algorithm can offer a wide range of applications such as digesting of continuous speech by checking the duplication of input data (same word utterance), and location identification of a mobile robot.

1. INTRODUCTION

This paper proposes a new frame-synchronous algorithm for spotting similar intervals by comparing arbitrary intervals in a reference pattern sequence with arbitrary intervals in an input pattern sequence. The algorithm is called Reference Interval-free Continuous DP (RIFCDP) because it is an extension of Continuous DP (CDP)[1] and because arbitrary intervals in a reference pattern can be regarded as separate reference patterns. This algorithm enables us to use a chain of pattern data (a chain of speech waves) as a reference pattern. In addition, reference patterns, such as keyword patterns, need not be segmented as a chain of time-series data(continuous speech) of arbitrary length is used as a reference pattern.

A method based on DP matching for extracting similar segments in doubly uttered speech has been proposed[2]. In this method, however, spotting-type detection and plural detections of the same segments are difficult.

We have previously proposed a partial-sentence spotting algorithm to determine simultaneously with sentence spotting a phrase unit as a interval of the sentence[3]-[5]. In this algorithm, the adjustment degree between input speech and a previously specified interval in a reference pattern made up of a chain of phrases is determined simultaneously with calculations performed over the entire reference pattern. In the method proposed in this paper, we extend the partial-sentence spotting algorithm to accommodate arbitrary intervals instead of specific intervals.

At each frame in the reference pattern, the algorithm stores as history cumulative distances as far back as desired to a previous frame in the reference pattern, and the starting input frame as well. Then, by assuming each frame of the reference pattern to be the final one, and using this distance history, the adjustment degree between an arbitrary reference-pattern interval and input speech can be determined. In this way, an interval in a reference pattern that has been recognized as having a high adjustment degree with input speech can be detected as the same interval synchronously with frames. The experimental results show that RIFCDP is successful in detecting the same intervals between a reference pattern sequence and an input sequence.

We have applied this new algorithm to speech retrieval from a speech database acting as a reference pattern. We present some experimental results to show the possibility of real-time speech retrieval. In addition, the method is not limited to speech and can be applied to motion images such as location identification of a mobile robot[8]. Moreover, the method can be used to check the duplication of input data (same word utterance) and can be applied to "digesting" of continuous speech data[7]. In these ways, the proposed algorithm can offer a wide range of applications.

2.REFERENCE INTERVAL-FREE CONTINUOUS DP(RIFCDP) ALGORITHM

2.1. The RIFCDP Concept

RIFCDP is an algorithm for spotting similar intervals between reference pattern R and input pattern sequence I synchronously with input frames. The input pattern sequence is therefore assumed to continue infinitely. Here, similar (or the same) interval (Rc,Ic) would lie between the two coordinate points (τ_1, t_1) and (τ_2, t_2), where τ corresponds to a frame number in the reference pattern R and t corresponds to an input frame number.

The RIFCDP algorithm, an extension of the partial- sentence spotting algorithm, can calculate and output an adjustment degree between arbitrary intervals in a reference pattern sequence and arbitrary intervals in an input sequence in

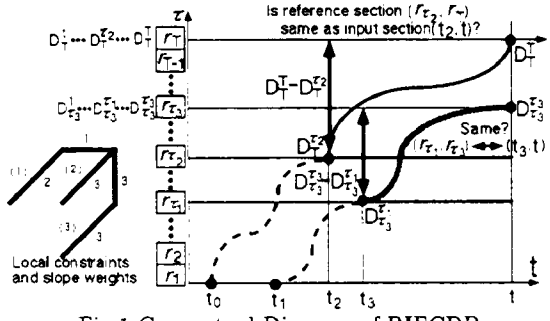

Fig.1 Conceptual Diagram of RIFCDP

†At present, Kawasaki Steel Corporation, ‡ Sharp Corporation, §Nippon Steel Corporation

a spotting-like manner. The concept behind this algorithm is shown in Fig.1. For each frame τ_r in the reference pattern, the cumulative distance at each time point in the path up to that frame as the final one is saved as history. As a result, the interval distance from frame τ_r back any number of previous frames can be calculated by subtracting saved cumulative distances. Then, by normalizing such an interval by a cumulative weight coefficient, the adjustment degree of two intervals can be compared.

Returning for a moment to normal spotting, and denoting the present time as t in Fig.1, the adjustment degree of the entire reference pattern is determined as D_T^T/L_T^T, where D_T is the cumulative distance up to final frame T in the reference pattern and L_T is the cumulative weight coefficient. Given that t_0 is the start time here, then corresponding intervals are specified as $(1,T)$ and (t_0,t). In the same way, the adjustment degree from frame 1 up to any frame τ_3 can be calculated knowing that the distance at the start time is zero, and corresponding intervals $(1,\tau_3)$ and (t_1,t) can be obtained.

In Fig.1, D_x^y denotes the cumulative distance when passing frame y on a path where frame x at present time t is assumed to be the final frame. Likewise, if we let L_x^y and S_x^y denote cumulative weight history and start time history, respectively, the adjustment degree from frame τ_1 to frame τ_3 is expressed as $(D_{\tau_3}^{\tau_3}-D_{\tau_3}^{\tau_1-1})/(L_{\tau_3}^{\tau_3}-L_{\tau_3}^{\tau_1-1})$ and corresponding intervals at that time are determined to be (τ_1,τ_3) and $(S_{\tau_3}^{\tau_1},t)$, where S_{τ_3} is t_3. in the figure. The above methodology enables the adjustment degree of an arbitrary interval in a reference pattern and corresponding intervals to be found. In addition, the adjustment degree of a certain arbitrary interval can be compared with that of another arbitrary interval, such as the one from frame τ_2 to frame T indicated by the other double-sided arrow in the figure.

In this calculation process, it is sufficient to provide cumulative distance history and cumulative weight history to each frame in the reference pattern for CDP calculations. Thus, as is done in CDP, the adjustment degree can be calculated for arbitrary intervals in an infinite input sequence synchronously with frames, and similar intervals can be spotted and extracted. Moreover, by limiting the number of frames to be saved as history based on the minimum and maximum detection intervals that depend on the length to be detected, the efficiency of the calculation process can be raised.

2.2. The RIFCDP Formalization

We show the formalization of the RIFCDP algorithm both for calculating the degree of adjustment of a certain arbitrary interval and for detecting similar intervals using the cumulative-distance history of each frame in the reference pattern.

2.2.1. Calculating Degree of Adjustment

We define the local distance as $d(t,\tau)$ at time t and at frame τ. This time, we use local constraints and asymmetric slope weights, as shown in Fig.1 (bottom left figure). $D(t,\ \tau_2,\ \tau_1)$ and $S(t,\ \tau_2,\ \tau_1)$ denote the cumulative distance history and the start time history respectively, when passing frame τ_1 on a path where frame τ_2 at present time t is assumed to be the final frame. $D(t,\ \tau_2,\ \tau_1)$ is equal to $D_{\tau_2}^{\tau_1}$, as mentioned above. $D(t,\tau_2,0)$ is always defined as 0, as shown in Eq. (1). Initial conditions are shown in Eq. (2).

$$D(-1,\tau_2,0)=D(0,\tau_2,0)=0 \quad (=D(t,\tau_2,0)) \tag{1}$$

$$\begin{cases} D(-1,\tau_2,\tau_1)=D(0,\tau_2,\tau_1)=CONST \\ S(-1,\tau_2,\tau_1)=-1 \qquad (1\le\tau_1\le\tau_2\le T) \end{cases} \tag{2}$$

$CONST$ in Eq. (2) is set to a large value to prevent the adjustment degree from becoming high when a distance is normalize by the longest frames(T). We should define the shortest detection length by using τ_1.

Then we use recurrence equations for calculating D and S frame-synchronously. The variable α(=1,2,3) is one of three start grids in local constraints (Fig.1) in Eq. (4)(5).

When $\tau=1$,

$$\begin{aligned} D(t,1,1) &= 3\cdot d(t,1) \\ S(t,1,1) &= t \end{aligned} \tag{3}$$

When $\tau\ge 2$

$$\begin{cases} P(t,\tau,1)= D(t-2,\tau-1,\tau-1)+2\cdot d(t-1,\tau)+d(t,\tau) \\ P(t,\tau,2)= D(t-1,\tau-1,\tau-1)+3\cdot d(t,\tau) \\ P(t,\tau,3)= D(t-1,\tau-2,\tau-2)+3\cdot d(t,\tau-1)+ \\ \qquad 3\cdot d(t,\tau) \end{cases} \tag{4}$$

Where $\tau=2$

$$P(t,2,3)= D(t,1,1)+3\cdot d(t,2) \tag{5}$$

The best path is selected by the following equation.

$$\alpha^*=arg\min_{\alpha=1,2,3} P(t,\tau,\alpha) \tag{6}$$

Next, the histories of cumulative distances and start times are copied from the selected start grid to the frame τ by Eq.(7) ~ (9) and then the cumulative distance up to τ is renewed by $P(t,\tau,\alpha^*)$ in Eq. (10).

When $\alpha^*=1$,

$$\begin{cases} D(t,\tau,k)=D(t-2,\tau-1,k) \\ S(t,\tau,k)=S(t-2,\tau-1,k) \end{cases} \qquad 1\le k\le\tau-1 \tag{7}$$

When $\alpha^*=2$,

$$\begin{cases} D(t,\tau,k)=D(t-1,\tau-1,k) \\ S(t,\tau,k)=S(t-1,\tau-1,k) \end{cases} \qquad 1\le k\le\tau-1 \tag{8}$$

When $\alpha^*=3$,

$$\begin{cases} D(t,\tau,k)=D(t-1,\tau-2,k) \quad 1\le k\le\tau-2 \\ S(t,\tau,k)=S(t-1,\tau-2,k) \quad (if\ \ \tau\ge 3) \\ D(t,\tau,\tau-1)=P(t,\tau,\alpha^*)-3\cdot d(t,\tau) \end{cases} \tag{9}$$

$$\begin{aligned} D(t,\tau,\tau) &= P(t,\tau,\alpha^*) \\ S(t,\tau,\tau) &= t \end{aligned} \tag{10}$$

We can define the longest detection length as N_{max} and reduce the number of histories to N_{max} at each node and reduce the culculation, memory and copying burden.

2.2.2. Detecting Similar Intervals

In this section, a few methods for detecting similar intervals are described by using degrees of adjustment for certain arbitrary intervals, as mentioned in the previous section. Since similar intervals to be detected should not be too short, let similar intervals to be longer than N_{min}(the shortest detection length).

First, the most similar intervals are determined. Degree of adjustment for the section from $\tau-n$ to τ ($n\ge N_{min}$) at time t is defined as $A(t,\tau,n)$ in Eq.(11).

$$A(t,\tau,n)=\frac{D(t,\tau,\tau)-D(t,\tau,\tau-n-1)}{n} \tag{11}$$

The most similar interval at frame τ is determined by the following equation.

$$n^* = \underset{N_{min} \le n \le \tau - 1}{arg\ min} \quad A(t, \tau, n) \tag{12}$$

Then, the most similar intervals and their degrees of adjustment at time t are determined.

$$(\tau - n^*, \tau) : (S(t, \tau, n^*), t) \tag{13}$$

$$A(t,\ \tau,\ n^*) \tag{14}$$

We describe a few methods for detecting similar intervals between two patterns. We omit the detailed algorithm here.

1. The most similar interval for reference pattern and input pattern is detected by checking all the frames $(N_{min} - T)$ and at each time $(1 - t_{end})$ (We need a certain threshold D_{max} to determine whether the best interval is a similar interval or not).
2. Similar intervals are detected frame-synchronously when the transition of the degree of adjustment in Eq.(14), shows a local minimum.
3. Intervals whose degree of adjustment shows less than the threshold D_{max} at time t and intervals whose degree of adjustment showed less than the threshold D_{max} before t are merged and regarded as similar intervals. This method is also frame-synchronous.

3. EVALUATION EXPERIMENTS

3.1. Evaluation Data

Experiments were performed to evaluate the performance of the RIFCDP algorithm in detecting the same intervals. The object data in these experiments were 30 sentences of conversational data taken from the speech database of the Acoustical Society of Japan, which is spoken by one person. The sampling frequency was 16 kHz, the frame interval was 8 msec, the speech was analyzed with a 20-channel filter bank, and feature quantities used a graduated spectrum field[6].

3.2. Experimental Conditions

It is generally appropriate to use longer units than a phoneme, such as words and phrases, for detecting similar intervals. To determine the length of time that should be extracted, we (1) extracted same intervals of three molas or more, (2) labeled the speech data, and (3) examined the length of time for various numbers of molas. The results of this investigation are shown in Fig.2.

This method assumes a length of one word or longer. A time length of 300 msec is therefore considered to be appropriate as the minimum detection interval from the figure.

Although this method is able to accommodate an infinitely chained input sequence, here we used two arbitrary files taken from a set of 30 speech data files for evaluating the same interval detection performance. We evaluated for all combinations of the above (${}_{30}C_2 = 435$).

3.3. Results and Discussion

If the detected intervals correspond to labeled intervals as the same intervals for both the reference pattern side and the input side, then the detected intervals (time) within the the same intervals are regarded as correct. "Detection Quality"

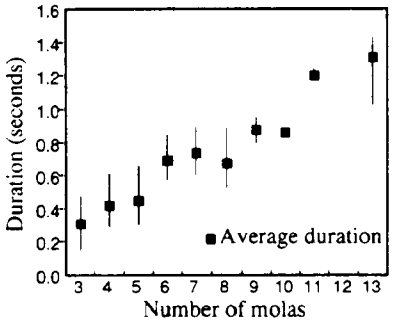

Fig.2 Number of molas & duration time in the same intervals.

as shown in Eq. (1) can be used as an index of the False Alarm(FA) percentage, and "Detection Ratio" as shown in Eq. (2) can be used as a temporal index of the detection rate. Fig.3 uses these indexes with Detection Quality indicated by the horizontal axis and Detection Ratio by the vertical axis.

$$Detection\ Quality = \frac{same\ intervals \cap detected\ intervals}{detected\ intervals} \tag{15}$$

$$Detection\ Ratio = \frac{same\ intervals \cap detected\ intervals}{same\ intervals} \tag{16}$$

In Fig.3, detection performance is examined for five minimum detection intervals: 80, 240, 400, 560 and 720 ms. The results show that the RIFCDP algorithm can detect about 60% of same intervals for a minimum detection interval of 240 or 400 msec, where Detection ratio was about 20%. Detection performance at these minimum detection intervals are the highest. This can be attributed to the fact that same intervals are considered to consist of three or more molas and that 240 or 400 ms agrees with the minimum length to be detected. If the length of the minimum detection interval is made too small, the problem described below will arise, and because pause intervals will become objects of best matching, detection performance drops and FAs increase. Fig.3, on the other hand, show that if the minimum detection interval is made long, the detection performance does not rise much. It is because short similar intervals cannot be detected even if long similar intervals can be detected effectively.

Another results shows that the longer the same intervals (the greater the number of molas) are, the higher the detection performance becomes, similar to word spotting.

As pauses are the same at the signal level, one of major problems in the algorithm is that long pauses will unfortunately be detected as same intervals with a high adjustment degree. While one way of looking at this problem would be to treat such pauses as correct detections, the authors, considering that items which must be extracted are similar intervals in terms of speech, treat pauses as False Alarm(FA) errors. Here, in the process of matching up pauses, we consider that such errors can be reduced by using power to assign a penalty to local distance.

In the Fig.4, C_1 is constant that effects on a penalty to local distance(if C_1 is larger, the penalty becomes larger). In this way, with the minimum detection interval set at 400 ms, a rise in detection performance could be seen as shown in the figure(when C_1 is varied).

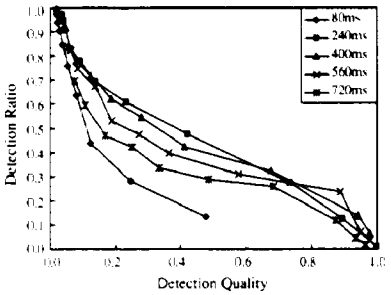

Fig.3 Detection performance of the same intervals in RIFCDP, depending on the minimum detection intervals

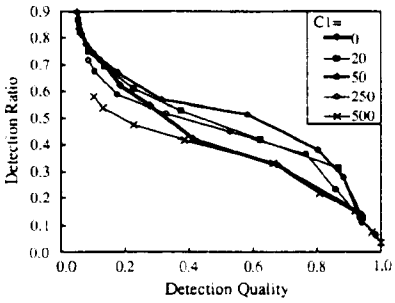

Fig.4 Change in detection performance due to penalties given to pause intervals

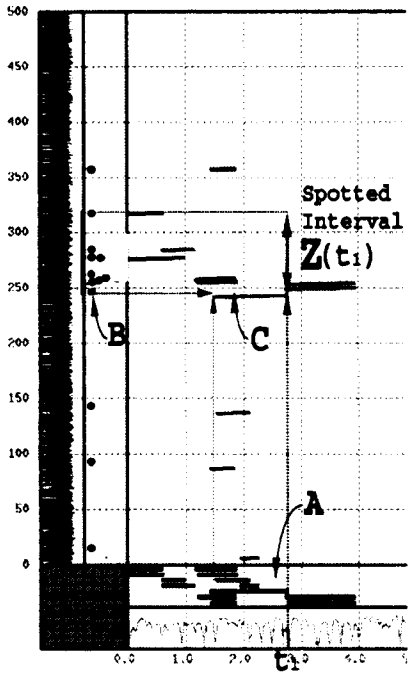

Fig.5 The results of speech retrieval from speech DB

4. REAL-TIME SPEECH/TEXT RETRIEVAL and OTHER APPLICATIONS

We have applied this new algorithm to speech/text retrieval from a database. Since this algorithm is frame-synchronous, by regarding a database as a reference pattern, retrieval is completed at the time when an utterance/input finishes In a retrieval system, the user should input the retrieval request in natural language, and the target intervals that the user wants to extract should be output as soon as the user has uttered/input enough to identify the target interval.

We assume the words and phrases a user utters/inputs for the request appear and converge in the target interval. Then retrieval of the target interval is thought to be possible by identifying the interval used to detect the same words and phrases in the database. The RIFCDP algorithm is able to realize these functions in real-time processing. We tested the possibility of speech/text retrieval from a database. A roughly ten-minute speech database mentioned in the previous chapter was used as the reference pattern. The input was another utterance of the keywords sequence. The interval Z where the RIFCDP output converges is output at time t_1 when the appearance of the output is enough to identify the target.

Assuming a reference pattern of 12.0 sec. and a workstation of about 100 MIPS, the computation time required by RIFCDP is faster than real time. And real-time processing for a roughly two-miniute reference pattern is currently possible by using a parallel machine by using 10 CPUs.

The method is not limited to speech and can be applied to motion images such as location identification for a mobile robot[8]. By preparing motion image data representing known positions of a robot as a reference pattern, the interval in this data which match the time-series data input is searched for, and the present position of the robot can be detected.

Moreover, when an input pattern is added to the reference pattern synchronously with the input, or when the reference pattern is updated, the proposed method can be used to check for input patterns that duplicate intervals already in the reference pattern. The method can also be applied to "digesting" at the signal level[7]. In these ways, the proposed method is applicable to a wide range of applications at the signal level.

5. CONCLUSION

This paper has proposed the Reference Interval-free Continuous DP (RIFCDP) algorithm for spotting similar intervals between arbitrary intervals in a reference pattern and arbitrary intervals in input speech, synchronously with input. This method is an extension of continuous DP, and it makes it possible for arbitrary intervals in a reference pattern to be treated as separate reference patterns. Experiments verified that this method can detect similar intervals.

We have applied this algorithm to speech retrieval from a speech database and showed the possibility of real-time speech/text retrieval. The proposed algorithm can offer a wide range of applications such as digesting of continuous speech and location identification of a mobile robot, etc.

The experiments described here used single speaker. A future evaluation should include object dialog from several speakers and different groups of several speakers.

References

[1] R.Oka,"Phonemic recognition of each frame with vector field feature using continuous dynamic programming",ICASSP(1986).

[2] Y.Hyogo, S.Nakagawa,"Extraction of similar speech patterns by DP-matching in a pair of sentences utterd continuously" ,IEICE Autumn Meeting A-22(1989)(in Japanese)

[3] J.Kiyama, Y.Itoh, R.Oka,"Spontaneous speech recognition by sentence spotting",3-rd Eurospeech, (1993-9)

[4] Y.Itoh, J.Kiyama,R.Oka," "Spotting Partial and Complete Sentences for Spontaneous Speech", ISSD93(1993).

[5] Y.Itoh, J.Kiyama, R.Oka,"Sentence Spotting Applied to Partial Sentences and Unknown Words",ICASSP(1994).

[6] R.Oka, H.Matumura,"Speaker-independent word speech recognition using the blurred orientation pattern obtained from the vector field of spectrum",9-th Int. Joint Conf. on Pattern Recognition(1988)

[7] J.Kiyama,Y.Itoh,R.Oka, "Automatic Detection of Topic Boundaries and Keywords in Arbitrary Speech using Incremental Reference Interval-free Continuous DP" ICSLP'96(1996-10).

[8] H.Kojima,Y.Itoh,R.Oka,"Location Identification of a Mobile Robot by Applying Reference Interval-free Continuous Dynamic Programming to Time-varying Images" International Symposium on Intelligent Robotic Systems, (1995-11).