A MATCHING ALGORITHM BETWEEN ARBITRARY SECTIONS OF TWO SPEECH DATA SETS FOR SPEECH RETRIEVAL

Yoshiaki Itoh

Iwate Prefectural University

ABSTRACT

This paper proposes a new matching algorithm to retrieve speech information from a speech database by speech query that allows continuous input. The algorithm is called Shift Continuous DP (CDP). Shift CDP extracts similar sections between two speech data sets. Two speech data sets are considered as reference patterns that are regarded as a speech database and input speech respectively. Shift CDP applies CDP to a constant length of unit reference patterns and provides a fast match between arbitrary sections in the reference pattern and the input speech. The algorithm allows endless input and real-time responses for the input speech query. Experiments were conducted for conversational speech and the results showed Shift CDP was successful in detecting similar sections between arbitrary sections of the reference speech and arbitrary sections of the input speech. This method can be applied to all kinds of time sequence data such as moving images.

1. INTRODUCTION

Information retrieval for speech or moving images is required in today's multi-media environment and will be needed more given the increase of such data on the Internet. Other needs will also occur, such as digesting and indexing speech or moving images. Information retrieval of speech or moving images needs a fast matching algorithm to extract the desired data sections from such databases.

We have proposed an algorithm to detect similar sections between two time sequence data sets, called Reference Interval Free Continuous DP (RIFCDP) [1]. This paper proposes a new fast and simple algorithm for this purpose. The algorithm is called Shift Continuous DP, which quickly spots between arbitrary sections of the database and arbitrary sections of the query input. The algorithms allow endless input and real-time responses to the query in the algorithm. This algorithm can be utilized with many types of applications, such as speech information retrieval, speech summarization [2], and location detection for robots, which is a moving image application [3].

Recently, much research on information retrieval of speech and moving images has been carried out. Most of it concerns fast matching between a database and the constant length of key data that is given before retrieval starts [4, 5]. Shift CDP is an algorithm that allows endless query input because any section of input can be regarded as a key for retrieval. It is difficult to perform detailed matching for speech data with methods using histograms [6] as the relation to the time axis is especially important in speech.

This algorithm can also be applied to "digesting" at the signal level [2]. When the reference pattern is updated synchronously with input, the proposed method can be used to check for input patterns that duplicate the same words.

2. SHIFT CONTINOUS DP ALGORITHM

The reference pattern R, that is considered as a database, and input pattern sequence I are expressed by Eq. (1) below, where $R\tau$ and I_t both indicate a member of a feature parameter series at the frame τ and time t respectively.

$$R = \{ R_1, \cdots, R\tau, \cdots, R_N \}$$
$$I = \{ I_1, \cdots, I_t, \cdots, I_\infty \} \quad (1)$$

Shift CDP is an algorithm that spots similar sections between reference pattern R and the input pattern sequence I synchronously with input frames. The input pattern sequence is assumed to continue infinitely, as indicated in the above equation. Here, a similar section (Rs, Is) would lie between the two coordinate points (τ_1, t_1) and (τ_2, t_2).

2.1. The concept of Shift Continuous DP (Shift CDP)

In the algorithm to solve the above problem, all the matching should be done between (τ_1, t_1) and (τ_2, t_{now}) at each input, where $1 \le \tau_1 \le \tau_2 \le N$, $1 \le t_1 \le t_{now}$. Let the minimum and maximum length for Rs be N_{min} and N_{max} respectively, so $N_{min} \le \tau_2 - \tau_1 \le N_{max}$. These constraints give the desired length to be detected and also reduce the calculation burden. To perform optimal matching for the length from N_{min} to N_{max} at frame τ_2, this frame is assumed to be the end frame of CDP and CDP is performed for the N_{max}-N_{min} pattern whose mean length is $(N_{min}+N_{max})/2$. Thus, CDP has to be done about $(N_{max}-N_{min}) \times N$ times for the reference pattern and its calculation burden is expected to be heavy even if such constraints for matching length are given.

Shift CDP can reduce the calculation burden described above. The concept behind this algorithm is shown in Fig. 1. First, unit reference patterns are taken from reference pattern R. A unit reference pattern (URP), has a constant frame length of N_{CDP}. The first URP is composed of frames from the first frame to the N_{CDP}-th frame in R. The starting frame of the second URP is shifted by N_{shift} frames and the

second URP is composed of the same number of N_{CDP} frames, from the N_{shift} +1-th frame. In the same way, the k-th URP is composed of N_{CDP} frames from the $k \times N_{Shift}$+1-th frame. The last URP is composed of N_{CDP} frames from the last frame of R toward the head of R. The number of URPs becomes N_{CDP}=[N/ N_{Shift}] +1 where [] indicates the integer that does not exceed the value.

For each URP, CDP is performed. It is not necessary to normalize each cumulative distance at the end frame of a URP because the length of all the URPs is same. As described above, Shift CDP is a very simple and flat algorithm that just performs CDP for each URP and integrates the results.

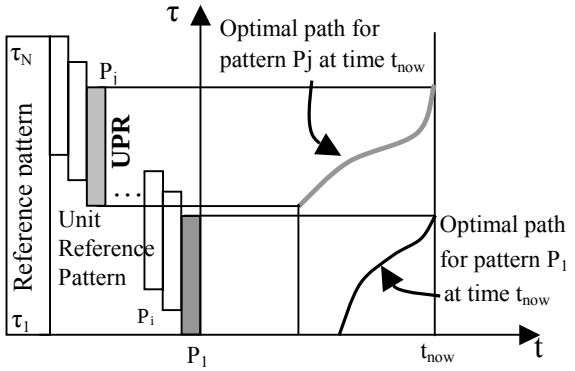

Fig. 1 The concept of the Shift CDP algorithm

2.2. Formalization of Shift CDP

This section formalizes the Shift CDP algorithm. We let the vertical axis represent the reference pattern frame $\tau(1 \le \tau \le N)$, and the horizontal axis as input time t. The local distance between the input t and frame τ is denoted as $D_t(\tau)$. Here, asymmetric local restrictions are used as the DP path, as shown in Fig.2.

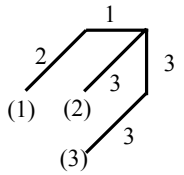

Fig.2 Asymmetry restrictions and slope weight

Let $G_t(i, j)$, $G_{t-1}(i, j)$ and $G_{t-2}(i, j)$ denote the cumulative distance up to frame j in the i-th URP at input time t, t-1 and t-2 respectively. Input time t is the current time and t-1 is the previous time. In the same way, $S_t(i, j)$, $S_{t-1}(i, j)$ and $S_{t-2}(i, j)$ denote the starting time. Let $\tau_S(i)$ and $\tau_E(i)$ be the frames of R that correspond to the starting and ending frames of the i-th URP. These locations should be calculated before input. To reduce the total calculation, $D3_t(\tau)$ and $D2_{t-1}(\tau)$ are calculated and saved apart from $D_t(\tau)$ according to the DP path restrictions, as shown in Fig. 2. $D3_t(\tau)$ and $D2_{t-1}(\tau)$ denote three times for $D_t(\tau)$ and two times for $D_{t-1}(\tau)$ respectively shown in Eq. (4) and Eq. (12). Initial conditions are given as:

$$\begin{cases} G_t(i, j) = G_{t-1}(i, j) = G_{t-2}(i, j) = \infty \\ S_t(i, j) = S_{t-1}(i, j) = S_{t-2}(i, j) = -1 \\ \qquad (1 \le i \le N_{PAT},\ 1 \le j \le N_{CDP}) \end{cases} \quad (2)$$

$$D2_{t-1}(\tau) = \infty \ (1 \le \tau \le N) \quad (3)$$

The following are recurrence formulas that can be calculated synchronously with input. The local distances between each frame of R and current input t are calculated beforehand. These prior calculations suppress the multiplications below.

$$\begin{cases} D_t(\tau) = d(\tau, t) \\ D3_t(\tau) = 3 \times D_t(\tau) \quad (1 \le \tau \le N) \end{cases} \quad (4)$$

LOOP i $(1 \le i \le N_{PAT})$: for each URP i,
LOOP j $(1 \le j \le N_{CDP})$: for each frame j of URP i,

$$\text{at } j=1, \quad \begin{cases} G_t(i, 1) = D3_t(\tau_S(i)) \\ S_t(i, 1) = t \end{cases} \quad (5)$$

$$\text{at } j \ge 2, \quad \begin{cases} P(1) = G_{t-2}(i, j-1) + D2_{t-1}(\tau_S(i) + j-1) \\ \qquad + D_t(\tau_S(i) + j-1) \\ P(2) = G_{t-1}(i, j-1) + D3_{t-1}(\tau_S(i) + j-1) \\ P(3) = G_{t-1}(i, j-2) + D3_t(\tau_S(i) + j-2) \\ \quad + D3_t(\tau_S(i) + j-1) \\ \text{but at } \tau=2, \\ \qquad P(2) = D3_t(\tau_S(i)) + D3_t(\tau_S(i)+1) \end{cases} \quad (6)$$

Here, the three terms of P in Eq. (6) represent the three start points of the path restrictions shown in Fig. 2. An optimal path is determined according to the following equation.

$$\alpha^* = arg\ min_{(\alpha=1, 2, 3)} P(\alpha) \quad (7)$$

The cumulative distance and the starting point are updated by Eq. (8) using $P(\alpha^*)$.

$$G_t(i, j) = P(\alpha^*) \quad (8)$$

$$S_t(i, j) = \begin{cases} S_{t-2}(i, j-1) & (\alpha^*=1) \\ S_t(i, j-1) & (\alpha^*=2) \\ S_{t-1}(i, j-2) & (\alpha^*=3) \end{cases} \quad (9)$$

$$\text{but at } j=2 \text{ and } \alpha^*=3, \ S_t(i, 2) = t$$

End LOOP j, if $j=N_{CDP}$. : end the CDP for the i-th URP.
End LOOP i, if $i=N_{PAT}$.: end the process of the current time t

The adjustment degree A(t, i) of the i-th URP at time t is given by the cumulative distance at the last frame without normalization by the length of reference patterns because the length of all URPs is equal.

$$A(t, i) = G_t(i, N_{CDP}) \quad (10)$$

After determining similar sections at time t, the cumulative distances, starting points and local distances are updated below. This procedure is just renewing the index of arrangements in the actual program that produce no calculation burden.

$$\begin{cases} G_{t-2}(i, j) = G_{t-1}(i, j) \\ G_{t-1}(i, j) = G_t(i, j) \quad (1 \le i \le N_{PAT}) \\ S_{t-2}(i, j) = S_{t-1}(i, j) \quad (1 \le j \le N_{CDP}) \\ S_{t-2}(i, j) = S_{t-1}(i, j) \end{cases} \quad (11)$$

$$D2_{t-1}(\tau) = 2 \times D_t(\tau) \quad (1 \le \tau \le N) \quad (12)$$

There are ways to determine similar sections according to the application purpose [1]. For example,

1. Detect the most similar section,
2. Detect any similar sections,

in the reference and the given input. In this paper, a threshold value is set and all the sections are detected when the adjustment degree exceeds the threshold value because there might be plural similar sections in the reference pattern or speech DB for input query.

3. EVALUATION EXPERIMENTS

3.1. Evaluation data and conditions

Experiments were performed to evaluate the performance of the Shift CDP algorithm in detecting similar sections. The object data in these experiments were 30 sentences of conversational data taken from the speech database of the Acoustical Society of Japan. This dialog was about route guidance and spoken by a single person. The sampling frequency was 16 kHz, the frame interval was 8 msec. A 36-dimensional graduated spectrum field was used for feature parameters whose Euclid distance determined the local distances.

Here, we considered similar sections that should be detected in speech as words and phrases. We defined the similar sections as sections that become three moras or more in the text of the phoneme sequence and labeled the speech data. The length of time for these similar sections was 300 msec to 1.5sec. Although this method can accommodate a continuous input sequence, here we used two speech data taken from a set of 30 data for the reference pattern and the input pattern. Among 435 combinations ($_{30}C_2$), 85 combinations had similar sections.

3.2. Results and Discussion

First, to evaluate detection performance when varying the threshold value at the time of detection judging, we performed an experiment at various lengths of a URP (N_{CDP}=5, 10, 20, 30, 40, 50, 70, 90, 110, 130, 150 frames) where the number of shifts is one frame(N_{Shift}=1). For the measurement of the detection performance, "Detection Quality" and "Detection Ratio" are introduced as shown in Eq. (13) and (14). Detection Quality can be regarded as a type of False Alarm (FA) rate used in word spotting.

$$\textbf{\textit{Detection Quality}} = \frac{(\textit{similar sections} \cap \textit{detected sections})}{(\textit{detected sections})} \quad (13)$$

$$\textbf{\textit{Detection Ratio}} = \frac{(\textit{similar sections} \cap \textit{detected sections})}{(\textit{similar sections})} \quad (14)$$

The results are shown in Fig. 3 with Detection Quality indicated on the horizontal axis and Detection Ratio on the vertical axis. The results show that the Shift CDP algorithm can detect about 60% of similar sections for a URP 400 msec long (50 frames) and Detection Quality of about 25%.

The upper graph of Fig. 3 shows that the detection performance is obviously declining when the URPs are composed of 5 or 10 frames. There are two main reasons for this. The first reason is that correct detection occurred such as for phonemes that were shorter than expected. The second reason is that the short URP resulted in miss-detections, with shorter keywords causing many FAs in word spotting. In this way, the performance declines when the length of URPs is set too short. The lower graph shows the case of longer URPs. The performance declines when the length of a URP exceeds 90 frames. This is because it is difficult to extract similar sections that are shorter than the URP length. The best detection performance is obtained at a URP of 400ms (50 frames). This is thought to be related to the length of the shortest similar section of 300ms.

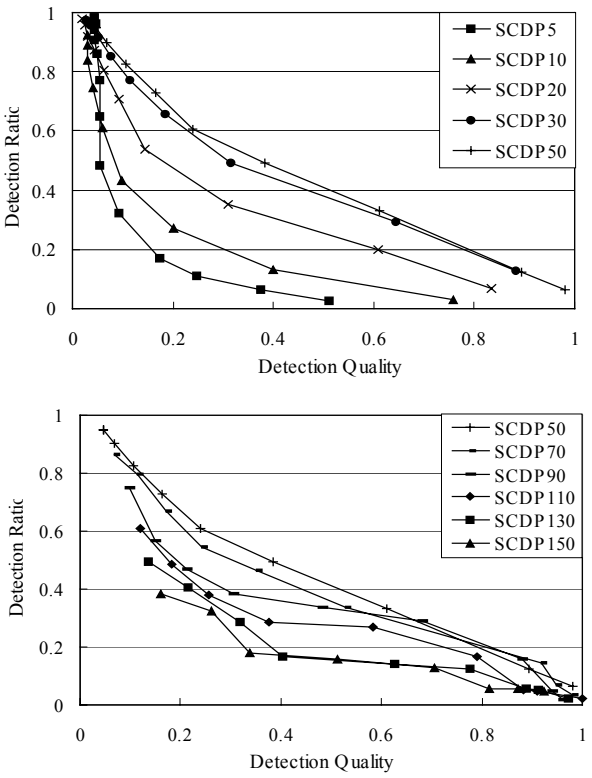

Fig. 3 Detection performance according to the length of a unit reference pattern (URP) in Shift CDP

Another experiment was performed at various shifting frames (N_{Shift}=1, 2, 5, 10, 15, 20, 25, 30, 40, 50 frames) where the length of the URP was set to 50 frames (N_{CDP}=50). The main results are shown in Fig. 4.

Frame shifting was introduced to reduce the calculation burden. The smaller the shifting number, the better the detection performance becomes because the time resolution is improved. 40 and 50 frames shifts cause a serious decline in the performance, but 5 frame shifts do not cause a decline in the performance at all and the decline at 25 frame shifts is very small. On the other hand, the reduction of the calculation burden is huge. For example, it becomes 1/25 at

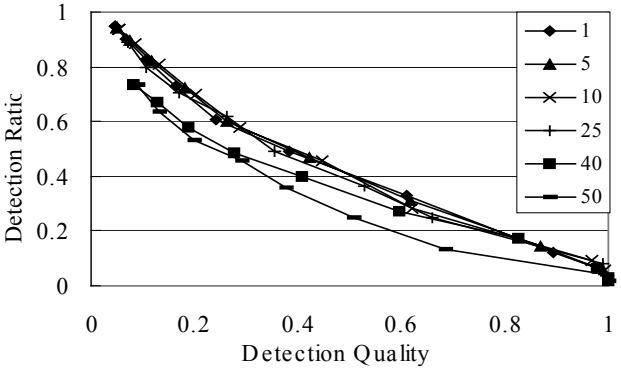

Fig. 4 Detection performance according to the number of frame shifts in Shift CDP

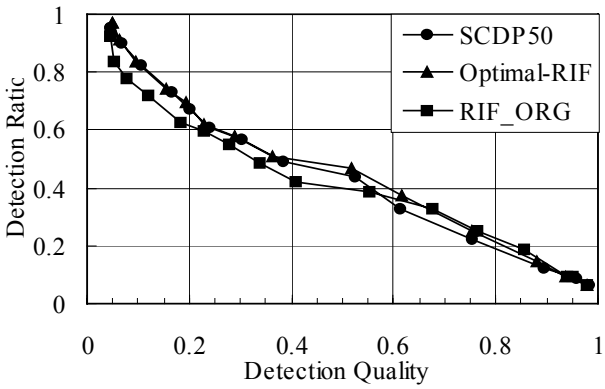

Fig. 5 Detection performance comparison among Shift CDP, RIFCDP, Optimal-RIF

25 frame shifting. In this case, two URPs cover each frame in the reference pattern R and such redundancy is thought to work well for the detection.

The performance was compared to other methods. There are two other methods. One is our former method, RIFCDP, and another is an algorithm that searches for the optimal path and is expected to obtain the best performance although the calculation burden is extraordinary. It is called Optima-RIF, which performs CDP at each reference frame for 100 patterns whose length varies from 50 to 150 frames. As shown in Fig. 5, the performance of Shift CDP is comparable with Optimal -RIF and better than RIFCDP.

In these experiments, pauses are not regarded as similar sections although they have the same signal level and are detected as similar sections with a high adjustment degree. It is easy to suppress pause detection by giving some penalty to local distances for no power frames. The detection performance can then be improved to 70% at Detection Quality of 40% [1].

3.3. Calculation Burden

To compare the calculation burden, actual calculation time was measured using a workstation (Sun Ultra5, 360MHz). Fig. 5 shows the results of actual calculation time when the length of a reference and an input pattern are set to 5 and 20 seconds respectively. For the reference, Optimal-RIF requires 50 times the processing time of RIFCDP. Shift CDP is faster than RIFCDP at even one frame shift. SCDP10 in Fig. 5 shows the case of Shift CDP at 10 frame shifts. The calculation time for Shift CDP decreases linearly in proportion to the number of frame shifts . At 25 frame shifts, Shift CDP(SCDP25) requires 1/25 the processing time of SCDP1, and 1/15 the processing time and 1/75 the memory use of RIFCDP. SCDP25 realizes real-time processing for a 25 second reference pattern with the above single CPU. In this case, the Shift CDP calculation burden becomes half of the local distance. It is possible to deal with parallel processing for the Shift CDP algorithm that has no surplus processing.

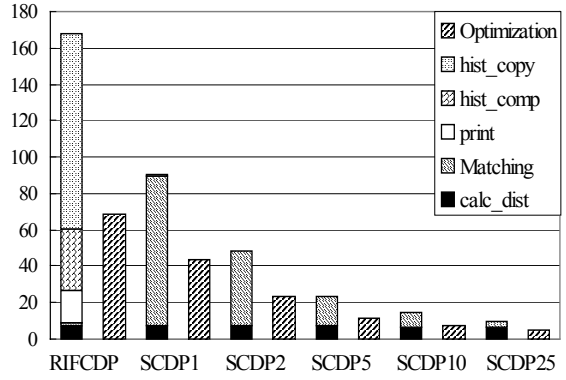

Fig. 6 Calculation time for Shift CDP

4. CONCLUSION

This paper has proposed a new fast algorithm for spotting similar sections between arbitrary sections from two time sequence data sets synchronously with input. This method is called Shift Continuous DP, which makes it possible for arbitrary sections in a reference pattern to be extracted for real-time information retrieval. Evaluation experiments verified that this method can detect similar sections between arbitrary sections of a reference pattern and arbitrary sections of input speech. The experiments described here used only one speaker. A future evaluation should include object dialog from several speakers and different groups of several speakers. In addition, while the object of the research described in this paper was speech oriented, the algorithm can be applied to other areas such as moving images, and we plan on evaluating such applications. Lastly, a speech information retrieval system is now under construction, which enables real-time extraction of a user's endless speech query using the Shift CDP algorithm on a multi-CPU machine.

REFERENCE

[1] Y. Itoh, J.Kiyama and R.Oka, "A proposal for a new algorithm of reference interval-free Continuous DP for real-time speech or text retrieval",ICASSP, vol.1,pp.486-489, Oct. 1996.

[2] J. Kiyama, Y. Itoh and R.Oka, "Automatic Detection of Topic Boundaries and Keywords in Arbitrary Speech Using Incremental Reference Interval-free Continuous DP", ICSLP, vol.3, pp.1946-1949, Oct 1996.

[3] H. Kojima, Y. Itoh and R.Oka, "Location identification of a mobile robot by applying reference interval-free continuous dynamic programming to time-varying images," Third Int.'l Symposium on Intelligent Robotics Systems, Nov. 1995.

[4] T. Nishimura, N. Sekimoto, J. Xin Zhang, M. Ihara, T.Akasaka, H.Takahashi and R.Oka, "Methodology for retrieving time sequence pattern," IWHIT/SM'99, pp.1-9, Oct.1999

[5] M. Sugiyama, "Fast and Robust Segment Searching Algorithms," SPECOM'99, Oct. 1999.

[6] G.A.Smith, H.Murase, K.Kashino, "Quick audio retrieval using active search," ICASSP, vol.6,pp.3777-3780, May 1998.

[7] R.C.Rose, E.I.Cang and R.P.Lippmann, "Techniques for Information Retrieval from Voice Messages," ICASSP, Vol.I, pp.317-320, Apr.1991.

[8] F.Chen and M.Withgott,"The use of emphasis to automatically summarize a spoken discourse," ICASSP, vol.I, pp.229-232, May 1992.