Dynamic Programming Algorithm Optimization for Spoken Word Recognition

HIROAKI SAKOE AND SEIBI CHIBA

Abstract—This paper reports on an optimum dynamic programming (DP) based time-normalization algorithm for spoken word recognition. First, a general principle of time-normalization is given using time-warping function. Then, two time-normalized distance definitions, called symmetric and asymmetric forms, are derived from the principle. These two forms are compared with each other through theoretical discussions and experimental studies. The symmetric form algorithm superiority is established. A new technique, called slope constraint, is successfully introduced, in which the warping function slope is restricted so as to improve discrimination between words in different categories. The effective slope constraint characteristic is qualitatively analyzed, and the optimum slope constraint condition is determined through experiments. The optimized algorithm is then extensively subjected to experimental comparison with various DP-algorithms, previously applied to spoken word recognition by different research groups. The experiment shows that the present algorithm gives no more than about two-thirds errors, even compared to the best conventional algorithm.

I. INTRODUCTION

IT is well known that speaking rate variation causes nonlinear fluctuation in a speech pattern time axis. Elimination of this fluctuation, or time-normalization, has been one of the central problems in spoken word recognition research. At an early stage, some linear normalization techniques were examined, in which timing differences between speech patterns were eliminated by linear transformation of the time axis. Reports on these efforts indicated that any linear transformation is inherently insufficient for dealing with highly complicated fluctuation nonlinearity as well as that time-normalization significantly improves recognition accuracy.

DP-matching, discussed in this paper, is a pattern matching algorithm with a nonlinear time-normalization effect. In this algorithm, the time-axis fluctuation is approximately modeled with a nonlinear warping function of some carefully specified properties. Timing differences between two speech patterns are eliminated by warping the time axis of one so that the maximum coincidence is attained with the other. Then, the time-normalized distance is calculated as the minimized residual distance between them. This minimization process is very efficiently carried out by use of the dynamic programming (DP) technique. The basic idea of DP-matching has been reported in several publications [1]–[3], where it has been shown by preliminary experiment on Japanese digit words that a recognition accuracy as high as 99.8 percent has been achieved, indicating the DP-matching effectiveness.

This paper reports an optimum algorithm for DP-matching through theoretical discussions and experimental studies. In-

vestigations were made, based on the assumption that speech patterns are time-sampled with a common and uniform sampling period, as in most general cases. One of the problems discussed in this paper involves the relative superiority of either a symmetric form of DP-matching or an asymmetric one. In the asymmetric form, time-normalization is achieved by transforming the time axis of a speech pattern onto that of the other. In the symmetric form, on the other hand, both time axes are transformed onto a temporarily defined common axis. Theoretical and experimental comparisons show that the symmetric form gives better recognition than the asymmetric one. Another problem discussed concerns slope constraint technique. Since too much of the warping function flexibility sometimes results in poor discrimination between words in different categories, a constraint is newly introduced on the warping function slope. Detailed slope constraint condition is optimized through experimental studies. As a further investigation, the optimized algorithm is experimentally compared with several varieties of the DP-algorithm, which have been applied to spoken word recognition by some research groups [3]–[6]. The optimized algorithm superiority is established, indicating the validity of this investigation.

II. DP-MATCHING PRINCIPLE

A. General Time-Normalized Distance Definition

Speech can be expressed by appropriate feature extraction as a sequence of feature vectors.

$$A = a_1, a_2, \text{------}, a_i, \text{------}, a_I$$

$$B = b_1, b_2, \text{------}, b_j, \text{------}, b_J. \qquad (1)$$

Consider the problem of eliminating timing differences between these two speech patterns. In order to clarify the nature of time-axis fluctuation or timing differences, let us consider an i-j plane, shown in Fig. 1, where patterns A and B are developed along the i-axis and j-axis, respectively. Where these speech patterns are of the same category, the timing differences between them can be depicted by a sequence of points $c = (i, j)$:

$$F = c(1), c(2), \text{------}, c(k), \text{------}, c(K), \qquad (2)$$

where

$$c(k) = (i(k), j(k)).$$

This sequence can be considered to represent a function which approximately realizes a mapping from the time axis of pattern A onto that of pattern B. Hereafter, it is called a warping function. When there is no timing difference between these
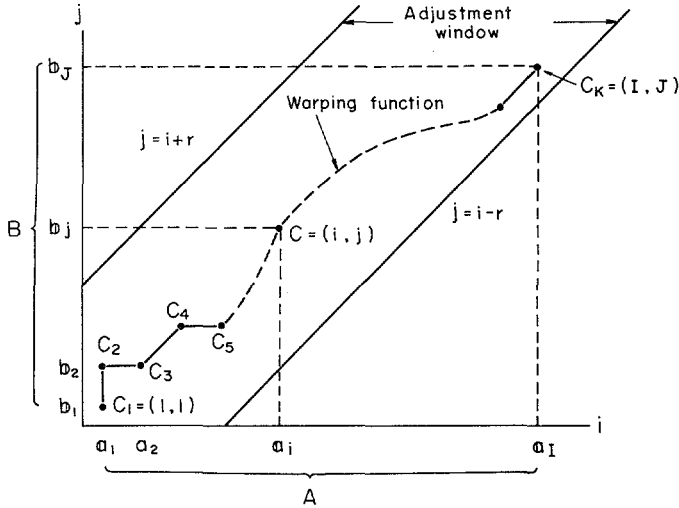
Fig. 1. Warping function and adjustment window definition.

patterns, the warping function coincides with the diagonal line $j = i$. It deviates further from the diagonal line as the timing difference grows.

As a measure of the difference between two feature vectors a_i and b_j, a distance

$$d(c) = d(i,j) = \| a_i - b_j \| \tag{3}$$

is employed between them. Then, the weighted summation of distances on warping function F becomes

$$E(F) = \sum_{k=1}^{K} d(c(k)) \cdot w(k) \tag{4}$$

(where $w(k)$ is a nonnegative weighting coefficient, which is intentionally introduced to allow the $E(F)$ measure flexible characteristic) and is a reasonable measure for the goodness of warping function F. It attains its minimum value when warping function F is determined so as to optimally adjust the timing difference. This minimum residual distance value can be considered to be a distance between patterns A and B, remaining still after eliminating the timing differences between them, and is naturally expected to be stable against time-axis fluctuation. Based on these considerations, the time-normalized distance between two speech patterns A and B is defined as follows:

$$D(A,B) = \underset{F}{\text{Min}} \left[\frac{\sum_{k=1}^{K} d(c(k)) \cdot w(k)}{\sum_{k=1}^{K} w(k)} \right] \tag{5}$$

where denominator $\sum w(k)$ is employed to compensate for the effect of K (number of points on the warping function F).

Equation (5) is no more than a fundamental definition of time-normalized distance. Effective characteristics of this measure greatly depend on the warping function specification and the weighting coefficient definition. Desirable characteristics of the time-normalized distance measure will vary, according to speech pattern properties (especially time axis expression of speech pattern) to be dealt with. Therefore, the present problem is restricted to the most general case where the following two conditions hold:

Condition 1: Speech patterns are time-sampled with a common and constant sampling period.

Condition 2: We have no *a priori* knowledge about which parts of speech pattern contain linguistically important information.

In this case, it is reasonable to consider each part of a speech pattern to contain an equal amount of linguistic information.

B. Restrictions on Warping Function

Warping function F, defined by (2), is a model of time-axis fluctuation in a speech pattern. Accordingly, it should approximate the properties of actual time-axis fluctuation. In other words, function F, when viewed as a mapping from the time axis of pattern A onto that of pattern B, must preserve linguistically essential structures in pattern A time axis and vice versa. Essential speech pattern time-axis structures are continuity, monotonicity (or restriction of relative timing in a speech), limitation on the acoustic parameter transition speed in a speech, and so on. These conditions can be realized as the following restrictions on warping function F (or points $c(k) = (i(k), j(k))$).

1) Monotonic conditions:

$$i(k - 1) \leqq i(k) \text{ and } j(k - 1) \leqq j(k).$$

2) Continuity conditions:

$$i(k) - i(k - 1) \leqq 1 \text{ and } j(k) - j(k - 1) \leqq 1.$$

As a result of these two restrictions, the following relation holds between two consecutive points.

$$c(k - 1) = \begin{cases} (i(k), j(k) - 1), \\ (i(k) - 1, j(k) - 1), \\ \text{or } (i(k) - 1, j(k)). \end{cases} \tag{6}$$

3) Boundary conditions:

$$i(1) = 1, j(1) = 1, \text{ and}$$

$$i(K) = I, j(K) = J. \tag{7}$$

4) Adjustment window condition (see Fig. 1):

$$| i(k) - j(k) | \leqq r \tag{8}$$

where r is an appropriate positive integer called window length. This condition corresponds to the fact that time-axis fluctuation in usual cases never causes a too excessive timing difference.

5) Slope constraint condition:

Neither too steep nor too gentle a gradient should be allowed for warping function F because such deviations may cause undesirable time-axis warping. Too steep a gradient, for example, causes an unrealistic correspondence between a very short pattern A segment and a relatively long pattern B segment. Then, such a case occurs where a short segment in consonant or phoneme transition part happens to be in good coincidence with an entire steady vowel part. Therefore, a restriction called a slope constraint condition, was set upon the warping function F, so that its first derivative is of discrete form. The slope constraint condition is realized as a restriction on the possible relation among (or the possible configuration of) several consecutive points on the warping function, as is shown in Fig. 2(a) and (b). To put it concretely, if point $c(k)$ moves forward in the direction of i (or j)-axis consecutive m times, then point
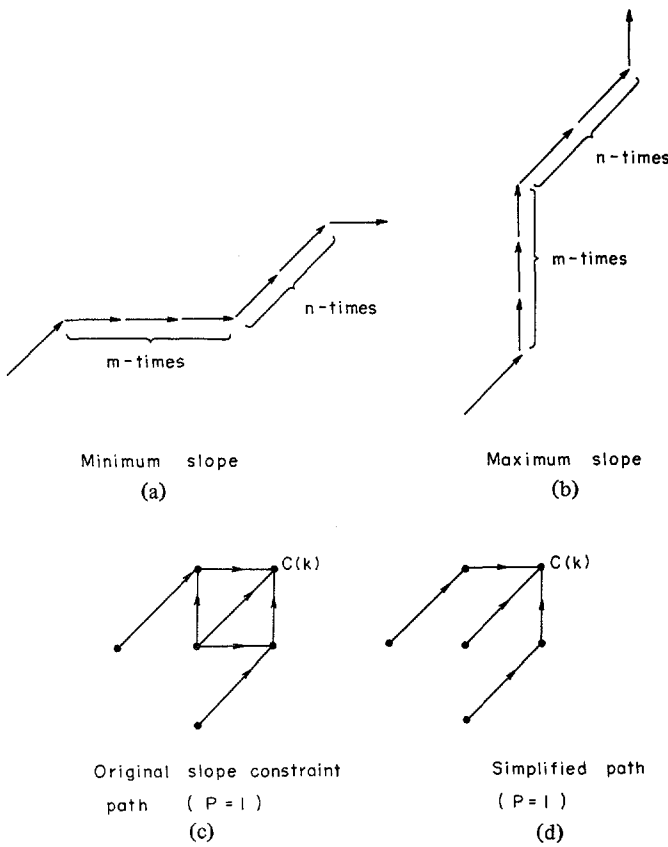
Fig. 2. Slope constraint on warping function.

$c(k)$ is not allowed to step further in the same direction before stepping at least n times in the diagonal direction. The effective intensity of the slope constraint can be evaluated by the following measure

$$P = n/m. \tag{9}$$

The larger the P measure, the more rigidly the warping function slope is restricted. When $p = 0$, there are no restrictions on the warping function slope. When $p = \infty$ (that is $m = 0$), the warping function is restricted to diagonal line $j = i$. Nothing more occurs than a conventional pattern matching no time-normalization. Generally speaking, if the slope constraint is too severe, then time-normalization would not work effectively. If the slope constraint is too lax, then discrimination between speech patterns in different categories is degraded. Thus, setting neither a too large nor a too small value for p is desirable. Section IV reports the results of an investigation on an optimum compromise on p value through several experiments.

In Fig. 2(c) and (d), two examples of permissible point $c(k)$ paths under slope constraint condition $p = 1$ are shown. The Fig. 2(c) type is directly derived from the above definition, while Fig. 2(d) is an approximated type, and there is another constraint. That is, the second derivative of warping function F is restricted, so that the point $c(k)$ path does not orthogonally change its direction. This new constraint reduces the number of paths to be searched. Therefore, the simple Fig. 2(d) type is adopted afterward, except for the $p = 0$ case.

C. Discussions on Weighting Coefficient

Since the criterion function in (5) is a rational expression, its maximization is an unwieldy problem. If the denominator

in (5)

$$N = \sum_{k=1}^{K} w(k) \tag{10}$$

(called normalization coefficient) is independent of warping function F, it can be put out of the bracket, while simplifying the equation as follows:

$$D(A, B) = \frac{1}{N} \min_{F} \left[\sum_{k=1}^{K} d(c(k)) \cdot w(k) \right]. \tag{11}$$

This simplified problem can be effectively solved by use of the dynamic programming technique. There are two typical weighting coefficient definitions which enable this simplification. They are as follows.

1) Symmetric form:

$$w(k) = (i(k) - i(k - 1)) + (j(k) - j(k - 1)), \tag{12}$$

then

$$N = I + J, \tag{13}$$

where I and J are lengths of speech patterns A and B, respectively [see (1)].

2) Asymmetric form:

$$w(k) = (i(k) - i(k - 1)), \tag{14}$$

then

$$N = I. \tag{15}$$

(Or equivalently, $w(k) = (j(k) - j(k - 1))$, then $N = J$.)

The basic concepts of the symmetric and asymmetric forms were originally defined by Sakoe and Chiba [3]. The problem of their relative superiority has been left unsolved.

If it is assumed that time axes i and j are both continuous, then, in the symmetric form, the summation in (5) means an integration along the temporarily defined axis $l = i + j$. In the asymmetric form, on the other hand, the summation means an integration along time axis i. As a result of this difference, time-normalized distance is symmetric, or $D(A, B) = D(B, A)$, in the symmetric form, though not in the asymmetric form. Another more important result, caused by the difference in the integration axis, is that, as is shown in Fig. 3, weighting coefficient $w(k)$ reduces to zero in the asymmetric form, when the point in warping function steps in the direction of j-axis, or $c(k) = c(k - 1) + (0, 1)$. This means that some feature vectors b_j are possibly excluded from the integration in the asymmetric form. On the contrary, in the case of symmetric form, minimum $w(k)$ value is equal to 1, and no exclusion occurs. Since discussions here are based on the assumption that each part in a speech pattern should be treated equally, an exclusion of any feature vectors from integration should be avoided as long as possible. It can be expected, therefore, that the symmetric form will give better recognition accuracy than the asymmetric form. However, it should be noted that the slope constraint reduces the situation where the point in warping function steps in the j-axis direction. The difference in performance between the symmetric one and asymmetric one will gradually vanish as the slope constraint is intensified.
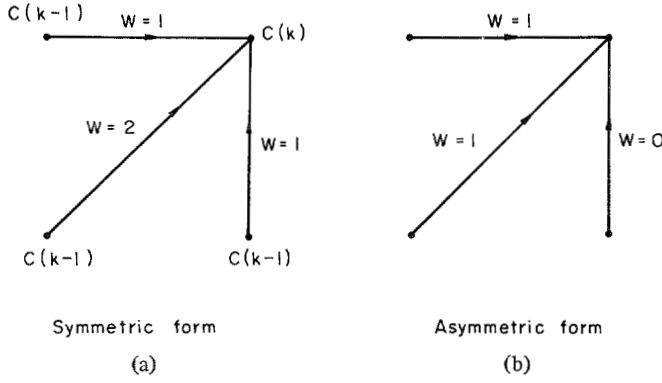
Fig. 3. Weighting coefficient $W(k)$ for both symmetric and asymmetric forms.

III. PRACTICAL DP-MATCHING ALGORITHM

A. DP-Equation

A simplified definition of time-normalized distance $D(A, B)$ given by (11) is one of the typical problems to which the well-known DP-principle [7] can be applied. The basic algorithm for calculating (11) is written as follows.

Initial condition:

$$g_1\,(c(1)) = d(c(1)) \cdot w(1). \tag{16}$$

DP-equation:

$$g_k\,(c(k)) = \min_{c\,(k-1)}\,[g_{k-1}\,(c(k-1)) + d(c(k)) \cdot w(k)]. \tag{17}$$

Time-normalized distance:

$$D(A, B) = \frac{1}{N}\,g_K\,(c(K)). \tag{18}$$

It is implicitly assumed here that $c(0) = (0, 0)$. Accordingly, $w(1) = 2$ in the symmetric form, and $w(1) = 1$ in the asymmetric form. By realizing the restriction on the warping function described in Section II-B and substituting (12) or (14) for weighting coefficient $w(k)$ in (17), several practical algorithms can be derived. As one of the simplest examples, the algorithm for symmetric form, in which no slope constraint is employed (that is $P = 0$), is shown here.

Initial condition:

$$g(1, 1) = 2d(1, 1). \tag{19}$$

DP-equation:

$$g(i, j) = \min \begin{bmatrix} g(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-1, j) + d(i, j) \end{bmatrix}. \tag{20}$$

Restricting condition (adjustment window):

$$j - r \leqq i \leqq j + r. \tag{21}$$

Time-normalized distance:

$$D(A, B) = \frac{1}{N}\,g(I, J), \quad \text{where } N = I + J. \tag{22}$$

Calculation details are briefly depicted in Section III-B.

The algorithm, especially the DP-equation, should be modified when the asymmetric form is adopted or some slope constraint is employed. In Table I, algorithms are summarized for both symmetric and asymmetric forms, with various slope constraint conditions. In this table, DP-equations for asymmetric forms are shown in some improved form. The first expression in the bracket of the asymmetric form DP-equation for $P = 1$ (that is, $g(i - 1,\ j - 2) + (d(i,\ j - 1) + d(i,\ j))/2$) corresponds to the case where $c(k - 1) = (i(k),\ j(k) - 1)$ and $c(k - 2) = (i(k - 1) - 1,\ j(k - 1) - 1)$. Accordingly, if the definition in (14) is strictly obeyed, $w(k)$ is equal to zero while $w(k - 1)$ is equal to 1, thus completely omitting the $d(c(k))$ from the summation. In order to avoid this situation to a certain extent, the weighting coefficient $w(k - 1) = 1$ is divided between two weighting coefficients $w(k - 1)$ and $w(k)$. Thus, $(d(i,\ j - 1) + d(i,\ j))/2$ is substituted for $d(i,\ j - 1) + 0 \cdot d(i,\ j)$ in this expression. Similar modifications are applied to other asymmetric form DP-equations. In fact, it has been established, by a preliminary experiment, that this modification significantly improves the asymmetric form performance.

B. Calculation Details

DP-equation or $g(i, j)$ must be recurrently calculated in ascending order with respect to coordinates i and j, starting from initial condition at $(1, 1)$ up to (I, J). The domain in which the DP-equation must be calculated is specified by

$$1 \leqq i \leqq I, \ 1 \leqq j \leqq J,$$

and

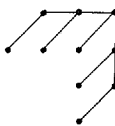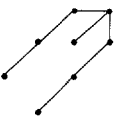$$j - r \leqq i \leqq j + r \text{ (adjustment window)}.$$

A practical procedure for calculating the time-normalized distance is shown in Fig. 4 as a flowchart.

IV. EXPERIMENTS AND RESULTS

A. Experiment Outline

In order to quantitatively evaluate various types of DP-matching, several recognition experiments were conducted. The speech analyzer used through these experiments was a 10-channel bandpass filter bank which covered up to a 5.9-kHz frequency range. The output of each channel was time-sampled every 18 ms and was digitized in order that it could be fed into the digital computer (NEAC-3100). Automatic gain control effect was realized by dividing each filter output level by their sum total, at every sampling period. The so-called time-frequency amplitude pattern thus obtained was stored on a digital magnetic tape file. Recognition experiments were conducted for the speech pattern read out of this file. The recognition scheme used was the forced decision pattern matching method, where the input pattern (unknown) was decided to be of the same category as the reference pattern to which the maximum coincidence (that is the minimum time-normalized distance) was achieved. Distance $d(i, j)$ was measured by the Chebyshev norm, which was employed in the previous work [2]. Reference patterns were adapted to each speaker. That is, one repetition of the complete vocabulary, pronounced by each speaker, was used as the reference patterns for each speaker.

TABLE I
SYMMETRIC AND ASYMMETRIC DP-ALGORITHMS WITH SLOPE CONSTRAINT CONDITION $P = 0, \frac{1}{2}, 1,$ AND 2

P	Schematic explanation	Symmetric / Asymmetric		DP-equation $g(i, j) =$
0		Symmetric	min	$\begin{bmatrix} g(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-1,j)+d(i,j) \end{bmatrix}$
		Asymmetric	min	$\begin{bmatrix} g(i,j-1) \\ g(i-1,j-1)+d(i,j) \\ g(i-1,j)+d(i,j) \end{bmatrix}$
1/2		Symmetric	min	$\begin{bmatrix} g(i-1,j-3)+2d(i,j-2)+d(i,j-1)+d(i,j) \\ g(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-2,j-1)+2d(i-1,j)+d(i,j) \\ g(i-3,j-1)+2d(i-2,j)+d(i-1,j)+d(i,j) \end{bmatrix}$
		Asymmetric	min	$\begin{bmatrix} g(i-1,j-3)+(d(i,j-2)+d(i,j-1)+d(i,j))/3 \\ g(i-1,j-2)+(d(i,j-1)+d(i,j))/2 \\ g(i-1,j-1)+d(i,j) \\ g(i-2,j-1)+d(i-1,j)+d(i,j) \\ g(i-3,j-1)+d(i-2,j)+d(i-1,j)+d(i,j) \end{bmatrix}$
1		Symmetric	min	$\begin{bmatrix} g(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-2,j-1)+2d(i-1,j)+d(i,j) \end{bmatrix}$
		Asymmetric	min	$\begin{bmatrix} g(i-1,j-2)+(d(i,j-1)+d(i,j))/2 \\ g(i-1,j-1)+d(i,j) \\ g(i-2,j-1)+d(i-1,j)+d(i,j) \end{bmatrix}$
2		Symmetric	min	$\begin{bmatrix} g(i-2,j-3)+2d(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-3,j-2)+2d(i-2,j-1)+2d(i-1,j)+d(i,j) \end{bmatrix}$
		Asymmetric	min	$\begin{bmatrix} g(i-2,j-3)+2(d(i-1,j-2)+d(i,j-1)+d(i,j))/3 \\ g(i-1,j-1)+d(i,j) \\ g(i-3,j-2)+d(i-2,j-1)+d(i-1,j)+d(i,j) \end{bmatrix}$

Experiments were conducted in three parts. The first part was carried out with the objectives of comparing the performances of symmetric form DP-matching and asymmetric form DP-matching, and optimizing the slope constraint condition. In the second part, further optimization of the slope constraint condition was investigated. In the final part of the experiments, the algorithm thus optimized was compared with several DP-algorithms proposed by different research groups.

B. Experiment (I)

The objective of this experiment was to compare symmetric form DP-matching and asymmetric form DP-matching performances, and to determine the best compromise for the slope constraint intensity (parameter P). Speech data used in this experiment were Japanese digit words (see Table II) isolatedly spoken by 10 male speakers. Six repetitions of the 10 digit words were made by each speaker. Then, for each speaker, each of the six repetitions was used as a reference pattern set. For each reference pattern set, the remaining five repetitions were supplied to recognition. Therefore, 10 (persons) × 6 (reference pattern sets) × 50 (input patterns) = 3000 (recognition tests) were conducted. The DP-matching subjected to this experiment covered both symmetric and asymmetric forms, with slope constraint condition of $P = 0, \frac{1}{2}, 1,$ and 2. In each case, window length r was set equal to 6, which covered the utmost ±108 ms timing difference. A linear time-normalization method was also tested where the time axis of the input pattern was adjusted to that of the reference pattern with linear transformation.

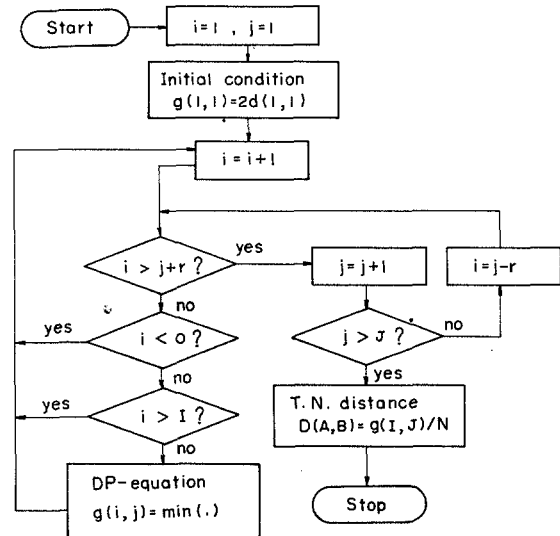Results are shown in Fig. 5 as two error rate curves. In this

Fig. 4. DP-matching flowchart.

figure, it can be seen that the performance of the asymmetric form DP-matching is evidently inferior to that of the symmetric one, and that the difference in performance between them tends to vanish gradually as the slope constraint is intensified. It can also be seen that symmetric form DP-matching performance is utterly unaffected by a slope constraint of up to $P = 1$. On the other hand, the asymmetric form DP-matching performance is very effectively improved by slope constraint. The optimum condition is $P = 1$. When the slope constraint is intensified beyond $P = 1$, the performance of the asym-

TABLE II
Ten Japanese Digits and Their Phonemic Transcriptions

0	1	2	3	4	5	6	7	8	9
rei	itʃi	ni	san	jon	go	roku	nana	hatʃi	kju

metric form, as well as that of the symmetric one tends to be degraded. Since extremely intensified slope constraint does not give any time-normalization, it is naturally understood that further growth in slope constraint will result in some worse performance than that of linear time-normalization method (0.8 percent error).

C. Experiment (II)

In order to further examine the effect of the slope constraint on the symmetric form DP-matching performance, another experiment was carried out for a 50-Japanese geographical name vocabulary. This vocabulary includes such confusing pairs of words as "Chiba"-"Shiga", "Okayama"-"Wakayama", "Fukushima"-"Tokushima", and "Hyogo"-"Kyoto". Each of two male speakers and two female speakers uttered six repetitions of the complete vocabulary. The first repetitions of each speaker were used as the reference patterns. The remaining five repetitions of each speaker were used as unknown input patterns, thus providing 4 (persons) × 50 (vocabulary size) × 5 (repetitions) = 1000 (recognition tests). The window length r here was set equal to 8, which covered utmost ±144 ms timing difference.

Results are shown in Fig. 6 for each of the slope constraint conditions. These results show that the slope constraint has a marked effect on the performance of the symmetric form DP matching, too. Optimum performance is also attained at $P = 1$.

D. Experiment (III)

Various DP-algorithms have been applied to spoken word recognition, by different research groups. Four typical ones, including those proposed by Sakoe and Chiba [3], Velichko and Zagoruyko [4], White and Neely [5], and Itakura [6], were subjected to comparison with the algorithms described in this paper. Details of each algorithm are summarized in Table III. Some modifications were made to equalize the experimental condition, but these modifications are not harmful to time-normalizing abilities of algorithms. Both the Japanese digit data and Japanese geographical name data were again used as test data. Results are shown in Table IV. From these results, it can be observed that the symmetric form DP-matching with slope constraint $P = 1$, described in this paper, is the best of various DP-algorithms applied to spoken word recognition.

V. Discussions

From the results shown in Fig. 5, it can be observed that the symmetric form DP-matching performance is significantly superior to that of the asymmetric one. It can also be seen that the difference in performance between them tends to decrease as the slope constraint grows. These observations completely agree with the theoretical discussions presented in Section II, indicating the validity of this investigation.
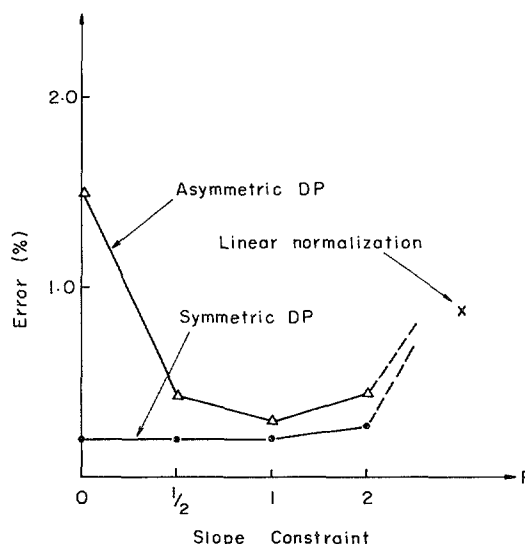

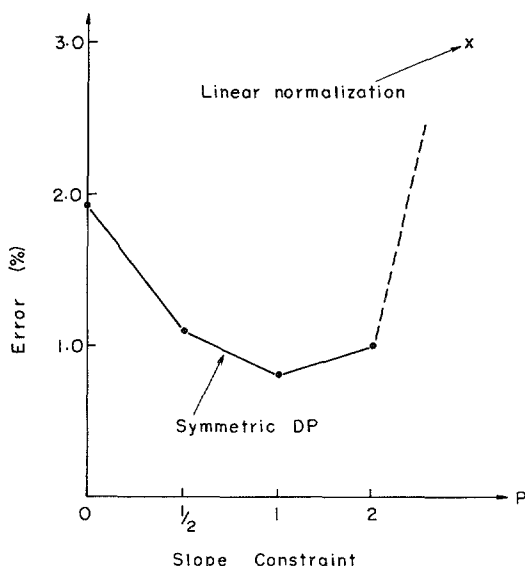Fig. 5. Experiment (I) results (for Japanese digit words).

Fig. 6. Experiment (II) results (for 50 Japanese geographical names).

From Figs. 5 and 6, it can be determined that the slope constraint condition with $P = 1$ is the optimum point for both symmetric and asymmetric forms. Moreover, as can be seen from Table I, the DP-equation for $P = 1$ is of the most simple form, next to that for $P = 0$. Thus, the slope constraint condition with $P = 1$ is favorable for computational economy, as well as for best performance. The slope constraint hardly affects the performance of the symmetric form DP-matching in case of Japanese digit vocabulary. This is perhaps due to the fact that, in case of Japanese digit words, the vocabulary size is so small and the separation between the words in the vocabulary is inherently so good that an optimally high recognition accuracy was achieved even without slope constraint. Nevertheless, the usefulness of the slope constraint for the symmetric form was established by the experiment on the Japanese geographical name data. Summing up these discussions, the symmetric form with slope constraint condition $P = 1$ is the optimum condition, when the speech patterns are time-sampled with a common and uniform sampling period.

TABLE III
FOUR VARITIES OF DP-ALGORITHMS SUBJECTED TO COMPARISON IN EXPERIMENT (III)

Algorithm	Initial Condition $g(1, 1) =$	Normalization Coefficient N	DP-equation $g(i, j) =$
Sakoe and Chiba [3]	$d(1, 1)$	I	$\min \begin{bmatrix} g(i-1, j) + d(i, j) \\ g(i-1, j-1) + d(i, j) \\ g(i-1, j-2) + d(i, j) \end{bmatrix}$
Velichko and Zagoruyko [4]	$a(1, 1)$	$\max \, [I, J]$	$\max \begin{bmatrix} g(i, j-1) \\ g(i-1, j-1) + a(i, j) \\ g(i-1, j) \end{bmatrix}$ where $a(i, j) = 1-d(i, j)$
White and Neely [5]	$d(1, 1)$	$(I + J)$	$\min \begin{bmatrix} g(i-1, j) + d(i, j) \\ g(i-1, j-1) + d(i, j) \\ g(i, j-1) + d(i, j) \end{bmatrix}$
Itakura [6]	$d(1, 1)$	I	$\min \begin{bmatrix} g(i-1, j) + \alpha \cdot d(i, j) \\ g(i-1, j-1) + d(i, j) \\ g(i-1, j-2) + d(i, j) \end{bmatrix}$ where $\alpha = \infty \quad (i(k-1) = i(k-2))$ $\alpha = 1 \quad (i(k-1) \neq i(k-2))$

The optimized algorithm was experimentally compared with several DP-algorithms, including those reported by other research groups. Results show that the present algorithm gives considerably better performance, for both Japanese digit data and Japanese geographical name data. This superiority can be attributed to careful investigations made to realize a pattern matching algorithm with rational characteristics of comparing each part of speech pattern evenly as far as possible.

As for the computational time, NEAC-3100 computer (index modified addition/substruction execution time is 8 μs) required about 3 s for each digit word recognition and about 30 s for geographical name recognition. These computational times scarcely depended upon the employed DP-equation. Recent high-speed digital integrated circuit devices and pipeline processor techniques made it feasible to realize the real time DP-matching operation. Actually, a DP-matching processor has been constructed which recognizes 60 geographical names in 300 ms after the utterance [8].

VI. CONCLUSIONS

The optimum DP-algorithm, applied to speech recognition, was investigated. Two forms of pattern matching method, symmetric and asymmetric forms, were proposed along with a new technique called slope constraint. These varieties were then compared through theoretical and experimental investigations. Conclusions are as follows.

1) The symmetric form gives better recognition accuracy than the asymmetric form.

2) Slope constraint is actually effective. Optimum performance is attained when the slope constraint condition is $P = 1$.

The validity of these results was ensured by a good agreement between theoretical discussions and experimental results.

The optimized algorithm was then experimentally compared with several other DP-algorithms applied to spoken word recognition by different research groups, and the superiority of the algorithm described in this paper was established.

ACKNOWLEDGMENT

The authors wish to thank Y. Kato for his encouragement throughout this research.

TABLE IV
EXPERIMENT (III) RESULTS

Test data / Algorithm	Error rate (%)	
	Japanese digit words	50 Japanese Geographical names
Sakoe and Chiba Symmetric P = 1 [in this paper]	0.2	0.8
Sakoe and Chiba Asymmetric P = 1 [in this paper]	0.3	1.3
Sakoe and Chiba [3]	0.3	1.5
Velichko and Zagoruyko [4]	2.0	2.7
White and Neely [5]	0.33	1.3
Itakura [6]	0.4	1.3
Linear method	0.87	5.9

REFERENCES

[1] H. Sakoe and S. Chiba, "A similarity evaluation of speech patterns by dynamic programming" (in Japanese), presented at the Dig. 1970 Nat. Meeting, Inst. Electron. Comm. Eng. Japan, p. 136, July 1970.
[2] ——, "A dynamic programming approach to continuous speech recognition," in *1971 Proc. 7th ICA, Paper 20 C13*, Aug. 1971.
[3] ——, "Comparative study of DP-pattern matching techniques for speech recognition" (in Japanese), in *1973 Tech. Group Meeting Speech, Acoust. Soc. Japan, Preprints (S73-22)*, Dec. 1973.
[4] V. M. Velichko and N. G. Zagoruyko, "Automatic recognition of 200 words," *Int. J. Man-Machine Stud.*, vol. 2, p. 223, June 1970.
[5] G. White and R. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 183–188, Apr. 1976.
[6] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.
[7] R. Bellman and S. Dreyfus, *Applied Dynamic Programming*. New Jersey: Princeton Univ. Press, 1962.
[8] S. Tsuruta, H. Sakoe, and S. Chiba, "Real-time speech recognition system by minicomputer with DP processor" (in Japanese), in *1974 Tech. Group Meeting Speech, Acoust. Soc. Japan, Preprints (S74-30)*, Dec. 1974.