Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition

LAWRENCE R. RABINER, FELLOW, IEEE, AARON E. ROSENBERG, MEMBER, IEEE,
AND STEPHEN E. LEVINSON, MEMBER, IEEE

Abstract—The technique of dynamic time warping for time registration of a reference and test utterance has found widespread use in the areas of speaker verification and discrete word recognition. As originally proposed, the algorithm placed strong constraints on the possible set of dynamic paths—namely it was assumed that the initial and final frames of both the test and reference utterances were in exact time synchrony. Because of inherent practical difficulties with satisfying the assumptions under which the above constraints are valid, we have considered some modifications to the dynamic time warping algorithm. In particular, an algorithm in which an uncertainty exists in the registration both for initial and final frames was studied. Another modification constrains the dynamic path to follow (within a given range) the path which is locally optimum at each frame. This modification tends to work well when the location of the final frame of the test utterance is significantly in error due to breath noise, etc. To test the different time warping algorithms a set of ten isolated words spoken by 100 speakers was used. Probability density functions of the distances from each of the 100 versions of a word to a reference version of the word were estimated for each of three dynamic warping algorithms. From these data, it is shown that, based on a set of assumptions about the distributions of the distances, the warping algorithm that minimizes the overall probability of making a word error is the modified time warping algorithm with unconstrained endpoints. A discussion of this key result along with some ideas on where the other modifications would be most useful is included.

I. Introduction

One of the most fundamental concepts in the area of speech pattern recognition is that of "time-warping" a reference to a test utterance so as to time register the two patterns. Although a wide variety of techniques are applicable to this problem, one of the most versatile of the algorithms which has been proposed is dynamic time warping [1]-[3]. Fig. 1 illustrates the general time warping problem. We denote a reference contour as $R(n)$, $0 \leqslant n \leqslant N$, and a test contour as $T(m)$, $0 \leqslant m \leqslant M$. We denote the "endpoints" of $R(n)$ as N_1 and N_2, and the endpoints of $T(m)$ as M_1 and M_2. The purpose of the time warping algorithm is to provide a mapping between the time indices n and m such that a time registration between the reference and test utterances is obtained. We denote the mapping w, between n and m as

$$m = w(n). \tag{1}$$

The function w must satisfy a set of boundary conditions at the "endpoints" of the utterances. For example, a typical assumption is that both the initial points and final points of the utterances are in time alignment, i.e.,

$$M_1 = w(N_1) \tag{2a}$$

$$M_2 = w(N_2). \tag{2b}$$

The set of boundary conditions of (2) is called the constrained endpoint (CE) set. Finally, to completely specify the warping function w, some assumptions must be made about the shape of $w(n)$. For example, $w(n)$ might be a linear function between the boundary points, in which case the time warping is a simple linear compression/expansion of one time scale to match the other one.

A more sophisticated and powerful approach to time warping is to constrain the warping function to satisfy a set of continuity conditions, e.g.,

$$w(n+1) - w(n) = 0, 1, 2 \qquad (w(n) \neq w(n-1)) \tag{3a}$$

$$= 1, 2 \qquad (w(n) = w(n-1)). \tag{3b}$$

Equations (3a) and (3b) require that $w(n)$ be monotonically increasing, with a maximum slope of 2, and a minimum slope of 0 except when the slope at the preceding frame was 0, in which case the minimum slope is 1. The boundary conditions of (2), together with the continuity conditions of (3) constrain the warping function w to lie within a parallelogram in the (n, m) plane, as shown in Fig. 2. (For convenience we assume that $N_1 = M_1 = 1$, and $N_2 = N$, $M_2 = M$ herein. Clearly there is no loss in generality due to these assumptions.) The vertices (labeled points A and B in Fig. 2) are obtained as the intersections of the lines

$$m - 1 = 2(n-1) \qquad \text{point } A \tag{4a}$$

$$m - M = (n-N)/2 \tag{4b}$$

and

$$m - 1 = \tfrac{1}{2}(n-1) \qquad \text{point } B \tag{5a}$$

$$m - M = 2(n-N). \tag{5b}$$

The dynamic warp function w is therefore constrained to follow a path inside the shaded region of Fig. 2.

A complete specification of the warping function results from a point-by-point measure of similarity between the reference contour $R(n)$ and the test contour $T(m)$. A similarity measure or distance function D must be defined for every pair of points (n, m) within the parallelogram of Fig. 2. The

Manuscript received April 6, 1978; revised July 6, 1978.

The authors are with Bell Laboratories, Murray Hill, NJ 07974.

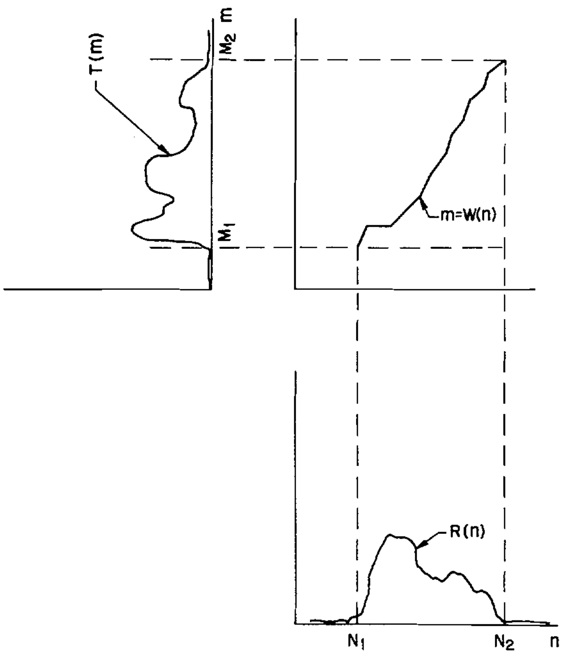


Fig. 1. Illustration of the general case of time warping.

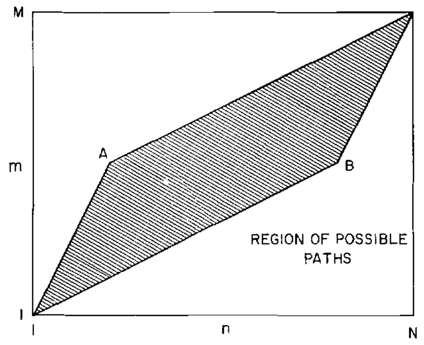


Fig. 2. Allowable region of time warping paths for a constraint on the maximum ratio of durations of test and reference utterances.

smaller the value of D, the greater the similarity between $R(n)$ and $T(m)$. Given the distance function D, the optimum dynamic path w is chosen to minimize the accumulated distance D_T along the path, i.e.,

$$D_T = \min_{\{w(n)\}} \sum_{n=1}^{N} D(R(n), T(w(n))). \tag{6}$$

An especially powerful technique for determining the optimum path w is the method of dynamic programming. Using this technique the accumulated distance to any grid point (n, m) can be recursively determined as

$$D_A(n, m) = D(n, m) + \min_{q \leq m} D_A(n-1, q) \tag{7}$$

where $D_A(n, m)$ is the minimum accumulated distance to the grid point (n, m) and is of the form

$$D_A(n, m) = \sum_{p=1}^{n} D(R(p), T(w(p))). \tag{8}$$

Given the continuity constraint of (3), (8) can be written in the form

$$\begin{aligned} D_A(n, m) = D(n, m) + \min\, [&D_A(n-1, m)\, g(n-1, m), \\ &D_A(n-1, m-1), D_A(n-1, m-2)] \end{aligned} \tag{9a}$$

where $g(n, m)$ is a weight of the form

$$g(n, m) = \begin{cases} 1 & w(n) \neq w(n-1) \\ \infty & w(n) = w(n-1). \end{cases} \tag{9b}$$

The final solution D_T of (6) is, by definition,

$$D_T = D_A(N, M). \tag{10}$$

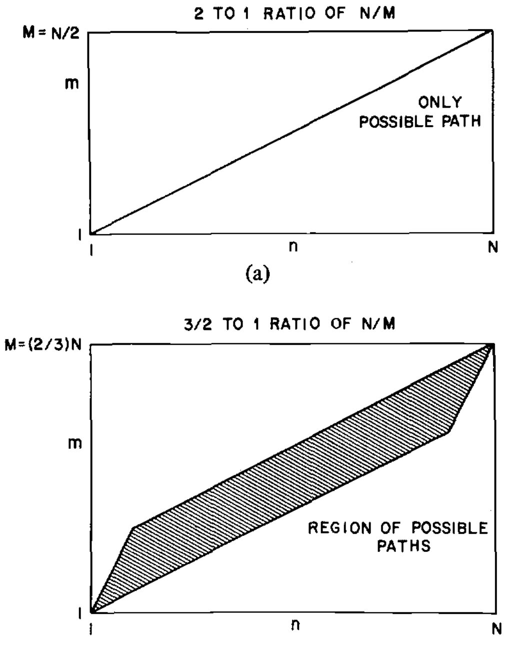
Equations (1)-(10) essentially define dynamic programming for time warping as originally defined by Sakoe and Chiba [1], [2] and modified by Itakura [3].[1] A careful examination of the assumptions used to determine the optimum time warping function leads to several issues which warrant further consideration. These issues include:

1) The assumption that the endpoints of the pair of utterances should be in exact time registration. For most applications the determination of the initial and final frames of an utterance is, at best, a highly imprecise calculation [4]. This is especially a problem for words that begin or end with weak fricatives whose energy and spectral properties are not significantly different from typical background silence. For utterances that begin and end with voiced sounds, the determination of the beginning of the utterance is generally easier to make than the determination of the end of the utterance. This is because the initiation of voicing generally has a sharp onset whereas the cessation of voicing is generally gradual. In addition, the end of an utterance is often accompanied by a short burst of "breathiness," which further complicates the endpoint calculation.

2) The restriction that the maximum allowable change in the warping function is 2 from one frame to the next. The implicit result of this restriction is that the largest ratio of durations, (M/N or N/M) between utterances which can be time registered is 2 to 1. However, if one considers the shape of the constraint region (i.e., the allowable dynamic paths) for a ratio which is exactly 2 to 1 [see Fig. 3(a)], then it is clear that no flexibility actually exists in choosing the path since only a single path is possible. Practically speaking, a ratio of about 1.5 to 1 is about the largest possible range in which a reasonable choice of warping paths exists [see Fig. 3(b)]. Thus, alternative means of removing this restriction are desirable.

3) The requirement that all possible paths within the grid of Fig. 2 be computed. For cases when N and M are large, a large number of grid points occur within the allowable regions and the computation grows proportional to $N \cdot M$. Since one would expect the optimum warping path to be reasonably close to a linear path, most of the computations at the extremities of the grid are needless. Thus, a possibly suboptimal procedure may be capable of substantially reducing computation, with little increase in total accumulated distance.

[1] Recently Sakoe and Chiba have proposed a symmetric form of a dynamic programming algorithm for time warping [6].



(a)

(b)

Fig. 3. Illustration of allowable regions of dynamic path for: (a) A 2-to-1 ratio of durations; (b) A $1\frac{1}{2}$-to-1 ratio of durations.

4) The issue as to whether the reference or test utterance controls the choice of the dynamic path. For many applications it is irrelevant which of the pair is used as the independent set of measurements (i.e., is mapped along the x axis of Fig. 10). However, for some cases it is preferred that the test utterance be mapped to the x axis since a large number of reference utterances will be matched to a single test utterance. In such cases the normalization to give average distance is independent of the duration of the reference utterance.

It is the purpose of this paper to investigate these issues by considering modified forms of the dynamic warping algorithm discussed above. In the next section we discuss the modifications and define three distinct versions of the dynamic time warping algorithm. In Section III we present results of a series of experimental comparisons among the algorithms based on a fairly large test set of spoken words. In Section IV we discuss the implications of the results for different applications.

II. Modifications to the Dynamic Time Warping Algorithm

For notational purposes we define the dynamic time warping algorithm discussed in the previous section as the CE2-1 (constrained endpoints, 2-to-1 range of slope) version. We have considered two modifications to the CE2-1 algorithm.

1) A version in which the boundary conditions of (2) are relaxed. In particular the new boundary conditions are of the form

$$1 \leqslant w(1) \leqslant \delta + 1 \qquad (11a)$$

$$M - \delta \leqslant w(N) \leqslant M \qquad (11b)$$

$$\min w(n) = 1 \qquad 1 \leqslant n \leqslant 2\delta + 1 \qquad (11c)$$

$$\max w(n) = M \qquad N - 2\delta \leqslant n \leqslant N \qquad (11d)$$

where δ represents the maximum anticipated range of mismatch (in frames) between the reference and test boundary points. For our simulations, a value of δ of 5 (frames) was used, representing a 75 ms region in which the initial and final frames could be mapped. The implementation of the boundary conditions of (11) can lead to some slight difficulties in that the warping function can now reach the final boundary of the reference prior to the last frame, i.e., it is possible that

$$w(n) = M \quad \text{for} \quad n < N \qquad (12)$$

in which case it is not physically meaningful to continue the path. For such cases the accumulated distance D_A is scaled by the factor (N/N_s) where N_s is the frame at which (12) is satisfied, so as to equalize the number of distances which enter into the total distance D_T. The resulting algorithm is referred to as the UE2-1 (unconstrained endpoints, 2-to-1 range of slope).

2) A version in which both the endpoint constraints are relaxed, and for which the allowable region of dynamic paths is constrained to follow the locally optimum path, to within a specified range. In particular, the initial endpoint satisfies (11a) and no constraint is used directly on the second endpoint. However, the range of values of m, for each value of n, is determined as

$$m_0 - \epsilon \leqslant m \leqslant m_0 + \epsilon \qquad (13)$$

where

$$m_0 = (m : D_A = \min_{\{m\}} [D_A(n-1, m)]) \qquad (14)$$

and ϵ is the specified range. The idea behind this version is to reduce computation by sharply constraining the region of allowable dynamic paths. In addition, if either the reference or test utterances is significantly longer than the other due to breath noise etc., this modification will generally be capable of choosing a path that eliminates the spurious sounds. This algorithm is referred to as the UELM (unconstrained endpoints, local minimum). A value of $\epsilon = 4$ (frames) corresponding to a 60 ms range around the minimum was used in our investigations. (It would be worthwhile, in future work, to study the effects of varying ϵ on the dynamic path.)

Fig. 4 gives a summary of the three dynamic warping algorithms we have considered. Typical warping functions and indications of the allowable regions are given in this figure. Before discussing the experimental investigations of these algorithms, it is worthwhile noting a couple of things about the algorithms. First, it is readily seen that the allowable space of dynamic warping paths for the CE2-1 algorithm is a subset of the space for the UE2-1 algorithm. However, it is not true that for *any* pair of test and reference utterances, the distance of the CE2-1 [call this D_T (CE2-1)] and the distance of the UE2-1 [call this D_T (UE2-1)] satisfy the relation

$$D_T\,(\text{CE2-1}) \geqslant D_T\,(\text{UE2-1}). \qquad (15)$$

This is because the constraint of (3b) eliminates potential paths which may occur for one version but not the other. However, as will be shown, the constraint of (15) is approximately maintained in practice. Because of the inherently dif-

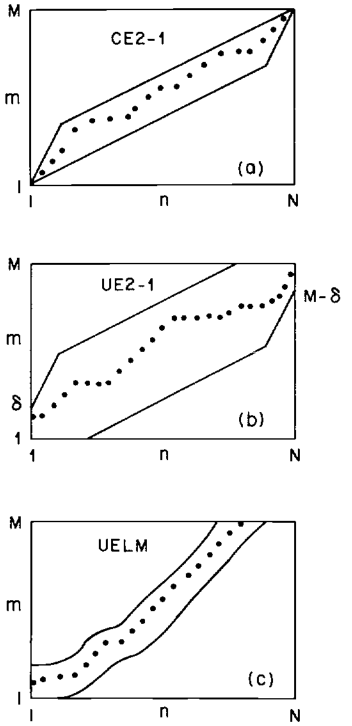


Fig. 4. Illustration of the three dynamic warping techniques used in this paper.

ferent regions of the UELM and UE2-1 or CE2-1 algorithms, no comparable statements can be made comparing the total distances for either pair of methods. With these points in mind, we now present some experimental results comparing the three algorithms.

III. Experimental Comparisons

In order to test the performance of each of the three dynamic warping algorithms of Section II, a data base of 10 isolated words spoken by 100 different speakers (50 male, 50 female) was used. Each word was sampled at a 6.67 kHz rate, digitized, and LPC analyzed using an 8 pole model. The initial and final endpoints for each word were determined semiautomatically and were essentially error free. The distance measure which was used was the log likelihood ratio proposed by Itakura [3], of the form

$$D(a_R, a_T) = \log\left[\frac{a_R V_T a_R'}{a_T V_T a_T'}\right] \qquad (16)$$

where $\boldsymbol{a}$ is the $(p+1)$ component vector containing the LPC coefficients, V is the $(p+1) \times (p+1)$ autocorrelation coefficient matrix (obtained from the speech waveform), and the subscripts R and T represent reference and test frames, respectively.

For each of the 10 words, the tokens[2] of Speaker 1 were arbitrarily chosen as the "reference" template to be matched by each of the remaining 99 tokens, using each of the three dynamic warping algorithms. In addition, matches were attempted between the "reference" templates and 100 randomly chosen words from a larger test set, none of which were the same as the template word. In this manner estimates of the probability density function (in the form of measured histograms) of dynamically warped distances both for the "correct" words, and for alternative, incorrect, possibilities could be estimated.

These measurements were made both with the selected token representing the "reference" utterance (i.e., mapped to the abscissa of the warping plane), and with the token representing the "test" utterance (i.e., mapped to the ordinate of the warping plane). There are two aspects to the question of whether the token utterance is mapped to the x axis or the y axis of the warping plane. One aspect is the question of whether or not the token utterance controls the dynamic warp, i.e., does the token utterance represent the independent variable of the warping function. Except for some trivial cases, it is readily shown that the total distances are slightly different depending on whether the "reference" or "test" controls the warp. The differences are due both to the lack of symmetry in the constraint equations [3(a) and 3(b)] on the warping function and to the possibility of the warping function coinciding with a boundary constraint for part or all of the path.

A second aspect of this issue (i.e., whether the token utterance is mapped to the abscissa or ordinate of the warping plane) is the lack of symmetry of the distance computation of (16). If the roles of reference (R) and test (T) are interchanged, the distance function becomes

$$D(a_T, a_R) = \log\left[\frac{a_T V_R a_T'}{a_R V_R a_R'}\right] \neq D(a_R, a_T). \qquad (17)$$

Thus, if one considers the warping problem as one of mapping the utterance along the y axis to the utterance along the x axis, and similarly defines the distance at a grid point as the distance from the y frame to the x frame, *both* the lack of symmetry of the distance computation *and* the differences between having reference and test as independent variables affect the results. The approach that we have considered in this paper is that when the token utterance is along the x axis, we use the distance of (16), and when the token utterance is along the y axis we use the distance of (17).

Table I gives a list of the 10 words used in this study. These words are a subset of a larger data base consisting of the letters A to Z, the digits 0 to 9, and the words STOP, ERROR, and REPEAT. This vocabulary is being used for studies of speaker independent recognition of words.

A typical plot of the total distance for each of the 99 versions of the correct word and an estimate of the resulting probability density function of the distances are given in Fig. 5 for the word A. For this example, the reference token was along the abscissa (x axis) for the dynamic warp. From top to bottom the data are for the UELM, UE2-1, and CE2-1 algorithms, respectively. Similarly Fig. 6 shows results for the word A when the test token was along the abscissa, i.e., the reverse of Fig. 5.

As seen in Fig. 5, the average total distance for the UE2-1 is smaller than the average total distance for the CE2-1 or the UELM algorithms. Also, the average total distance for the

[2] A token, as referred to throughout this paper, is a single word uttered by a single speaker.

TABLE I
WORDS USED IN THE EXPERIMENTAL INVESTIGATIONS

Word No.	*Actual Word*
1	7
2	4
3	2
4	X
5	W
6	C
7	A
8	Repeat
9	Error
10	Stop

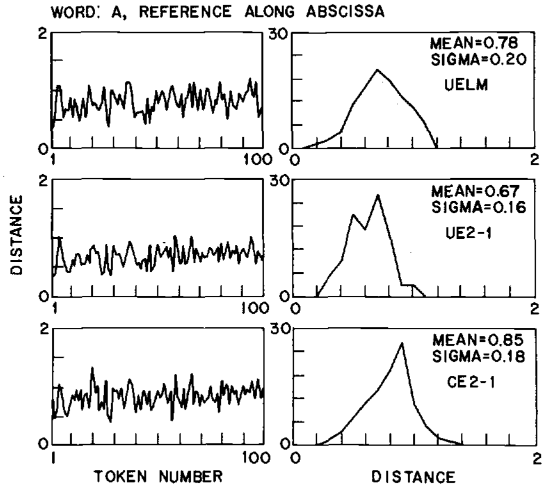


Fig. 5. Total distances and distance histograms for the three dynamic warping algorithms for the word A. Reference utterance along the abscissa of the warping plane.

UELM is always smaller than the average total distance for the CE2-1. These results tend to confirm our initial observation that constrained endpoints can lead to larger total distances than the unconstrained case for the reasons discussed previously. Measurements of the standard deviations of the distributions of Fig. 5 indicate somewhat smaller and less consistent variations among the three algorithms. To a first approximation the standard deviations of the three distributions are approximately equal.

By comparing the results of Figs. 6 (test along the x axis) and 5 (reference along the x axis), it can be seen that no completely consistent difference, for all three algorithms, is obtained. However, a slight reduction in both average distance and in standard deviation is obtained when the test utterance is along the x axis.

Fig. 7 presents scatter plots of the total distance obtained for each of the three dynamic warping algorithms. Data for all 990 cases (10 words X 99 speakers) are presented. It can now readily be seen that even though, on average, the total distance of the UELM is less than the total distance of the CE2-1 algorithm, there exists a significant number of individual cases in which this inequality does not hold. In such cases the optimum dynamic path does not follow the local dynamic path within the specified range. It can also be seen that, as mentioned earlier, for every single case the UE2-1 algorithm gives a total distance which is at least as small as that of the CE2-1.

The preliminary indication from the data discussed above is that the modifications to the CE2-1 dynamic warping algorithm serve to reduce the total distance and hence, improve the performance of the method. Although this is indeed the case for most practical applications (especially in the area of speech or speaker recognition), a second aspect of the problem must be considered—namely the performance of these algorithms when the reference and test utterances are different. If concomittant reductions in distance are obtained in this case, then a more sophisticated analysis is required to access the overall performance of the modified algorithms.

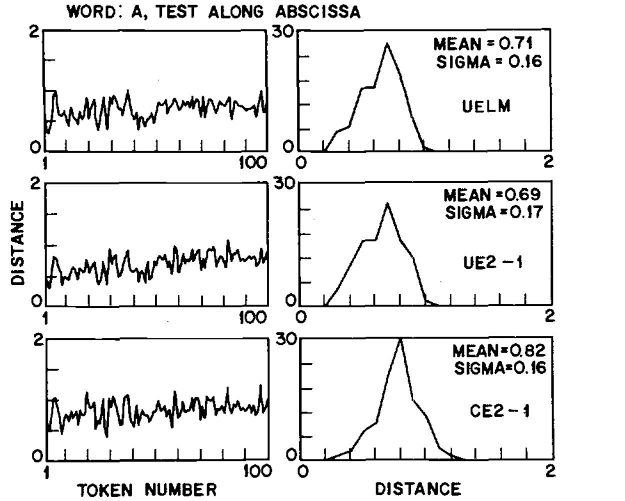


Fig. 6. Total distances and distance histograms for the three dynamic warping algorithms for the word A. Test utterance along the abscissa of the warping plane.

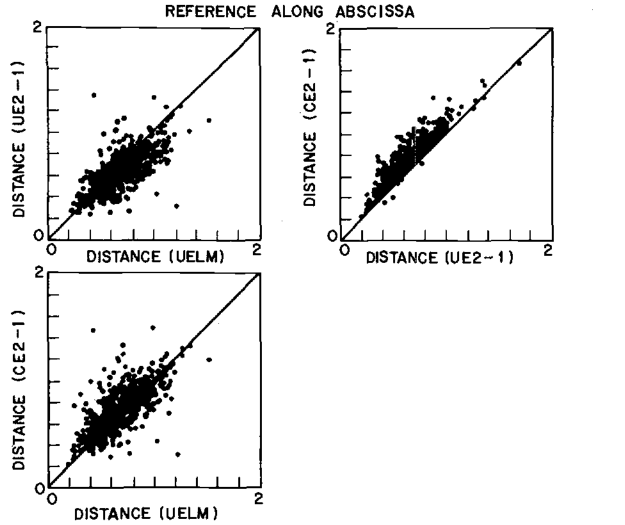


Fig. 7. Scatter plots comparing total distances of each of the three warping algorithms against each other. Reference utterance along the abscissa of the warping plane.

The data of Figs. 8-10 are essentially of the same type as those of Figs. 5-7 except that the comparisons are made between different test and reference words. Although the individual distances are significantly larger than those of Figs. 5-7, the relative ordering of the algorithms remains essentially the same in all cases. Also, the differences between results with either reference or the test utterance along the x axis are essentially random, i.e., no consistent difference is seen in the data.

Based on the results shown in Figs. 5-10, it is clear that a

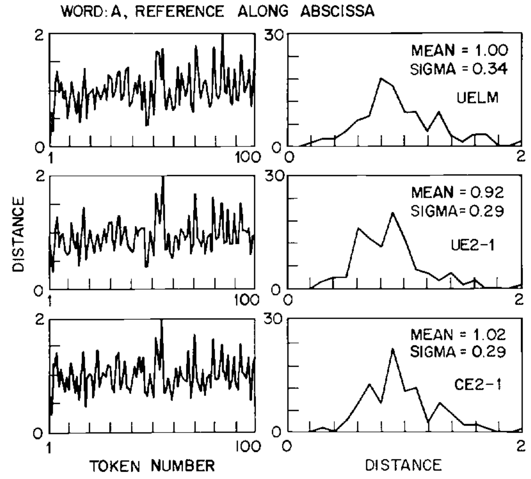


Fig. 8. Total distances and distance histograms for the three dynamic warping algorithms when the reference and test words were different for the word *A*. Reference utterance along the abscissa of the warping plane.

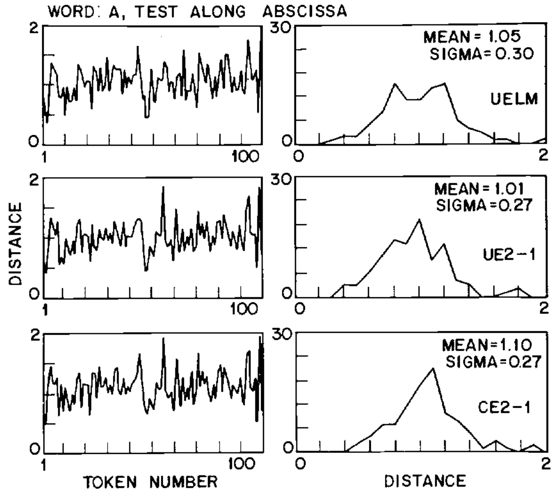


Fig. 9. Total distances and distance histograms for the three dynamic warping algorithms when the reference and test words were different for the word *A*. Test utterance along the abscissa of the warping plane.

more explicit model is required to assess the usefulness of the modifications to the basic time warping algorithm. It is assumed that the probability ($p_c(x)$) of obtaining a dynamic warped total distance of x, given that the reference and test utterances are the same word, is

$$p_c(x) = N[M_1, \sigma_1] \tag{18}$$

where $N[M_1, \sigma_1]$ is the normal distribution with mean M_1 and standard deviation σ_1. Similarly, it is assumed that the probability ($p_e(x)$) of obtaining a dynamic warped total distance of x, given that the reference and test utterances are different words, is

$$p_e(x) = N[M_2, \sigma_2]. \tag{19}$$

A threshold γ is chosen such that if $x > \gamma$, the decision is made that the reference and test words are different, whereas if $x \leqslant \gamma$, the decision is made that the reference and test words are the same. The threshold γ is chosen as the equal error

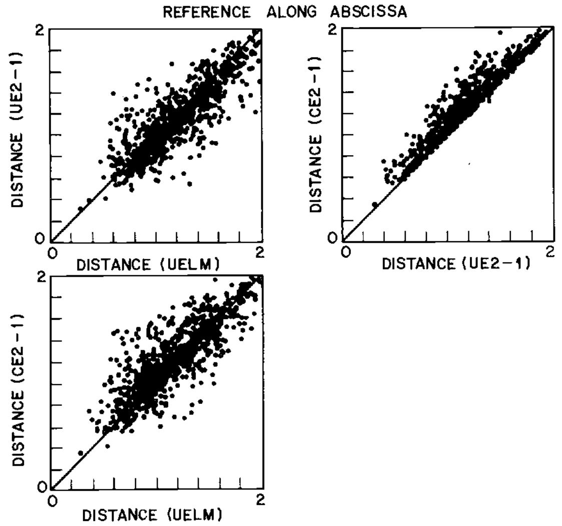


Fig. 10. Scatter plots comparing total distances of each of the three warping algorithms against each other when the test and reference words were different. Reference utterance along the abscissa of the warping plane.

threshold, i.e., the point at which the probability of a miss and the probability of a false alarm are equal. This threshold is readily determined analytically (based on M_1, σ_1, M_2, and σ_2) in the following manner. The probability of a miss p_M can be written as

$$p_M = \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-(x-M_1)^2/2\sigma_1^2}\, dx. \tag{20}$$

By making the substitution

$$y = \frac{(x - M_1)}{\sigma_1} \tag{21}$$

(20) can be written as

$$p_M = \int_{(\gamma-M_1)/\sigma_1}^{\infty} \frac{1}{\sqrt{2\pi}} e^{(y^2/2)}\, dy = \operatorname{erfc}\left(\frac{\gamma - M_1}{\sigma_1}\right) \tag{22}$$

where erfc is the standard complementary error function. Similarly, the false alarm probability (p_F) can be written as

$$p_F = \int_{-\infty}^{\gamma} \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-(x-M_2)^2/2\sigma_2^2}\, dx \tag{23}$$

which can be put in the form

$$p_F = \operatorname{erfc}\left(\frac{M_2 - \gamma}{\sigma_2}\right). \tag{24}$$

Since

$$p_F = p_M \tag{25}$$

(by definition of the equal error threshold), then (22), (24), and (25) lead to the result

$$\gamma = \frac{M_1 + \frac{\sigma_1}{\sigma_2} M_2}{1 + \frac{\sigma_1}{\sigma_2}}. \tag{26}$$

TABLE II
DATA FROM THE EXPERIMENTAL INVESTIGATIONS

		Reference Along Abscissa						Test Along Abscissa					
Word	Time Warp Method	M_1	σ_1	M_2	σ_2	γ	p_M	M_1	σ_1	M_2	σ_2	γ	p_M
7	UELM	.52	.13	1.11	.29	.70	.081	.48	.10	.99	.24	.63	.067
7	UE2-1	.50	.11	1.04	.27	.66	.078	.47	.08	.93	.22	.59	.063
7	CE2-1	.58	.11	1.14	.30	.73	.086	.55	.09	1.03	.24	.68	.073
4	UELM	.61	.31	1.39	.43	.94	.147	.53	.23	1.29	.37	.82	.102
4	UE2-1	.56	.28	1.37	.42	.88	.124	.49	.19	1.25	.34	.76	.076
4	CE2-1	.58	.28	1.42	.42	.92	.115	.52	.19	1.31	.34	.80	.070
2	UELM	.63	.18	1.07	.34	.78	.199	.51	.13	1.06	.33	.67	.115
2	UE2-1	.58	.12	1.10	.31	.73	.113	.51	.11	1.04	.30	.65	.098
2	CE2-1	.63	.13	1.19	.31	.80	.102	.58	.12	1.17	.31	.74	.085
X	UELM	.56	.14	1.33	.43	.75	.089	.61	.17	1.17	.32	.80	.127
X	UE2-1	.51	.09	1.18	.38	.64	.077	.55	.10	1.11	.30	.69	.081
X	CE2-1	.65	.13	1.37	.40	.83	.087	.72	.15	1.22	.28	.89	.123
Y	UELM	.69	.14	1.24	.29	.87	.100	.73	.14	1.31	.29	.92	.088
Y	UE2-1	.69	.13	1.27	.22	.91	.049	.73	.12	1.33	.21	.95	.035
Y	CE2-1	.73	.13	1.37	.25	.95	.046	.77	.13	1.43	.25	1.00	.041
C	UELM	.76	.17	1.10	.30	.88	.235	.65	.19	.97	.23	.79	.223
C	UE2-1	.66	.13	1.04	.29	.78	.183	.59	.12	.93	.22	.71	.159
C	CE2-1	.84	.16	1.13	.31	.94	.270	.74	.13	1.02	.24	.84	.224
A	UELM	.78	.20	1.00	.34	.86	.342	.71	.16	1.05	.30	.83	.230
A	UE2-1	.67	.16	.92	.29	.76	.290	.69	.17	1.01	.27	.85	.176
A	CE2-1	.85	.18	1.02	.29	.92	.359	.82	.16	1.10	.27	.92	.257
REPEAT	UELM	.64	.19	1.36	.46	.85	.134	.69	.19	1.42	.40	.93	.108
REPEAT	UE2-1	.71	.18	1.31	.37	.91	.138	.78	.19	1.36	.32	1.00	.128
REPEAT	CE2-1	.76	.20	1.39	.38	.98	.139	.87	.20	1.44	.35	1.06	.138
ERROR	UELM	.65	.15	1.19	.38	.80	.154	.66	.18	1.39	.43	.88	.115
ERROR	UE2-1	.66	.18	1.20	.37	.84	.163	.64	.18	1.35	.37	.87	.098
ERROR	CE2-1	.69	.19	1.26	.37	.88	.154	.67	.20	1.46	.41	.93	.098
STOP	UELM	.45	.11	1.17	.34	.63	.055	.48	.12	1.11	.33	.65	.081
STOP	UE2-1	.42	.09	1.12	.33	.57	.048	.45	.09	1.06	.32	.58	.069
STOP	CE2-1	.46	.09	1.17	.34	.60	.058	.49	.09	1.11	.31	.63	.061

Equation (26) can be used to give γ analytically (once M_1, M_2, σ_1, and σ_2 are known) and either (22) or (24) is used to give p_M or p_F.

In order to use the model described above, the distributions of total distance must be normally distributed. A test of normality (the Kolmogorov-Smirnov test [5]) was used on all the sets of distances used in this experiment, and the results indicated that in almost all cases the assumption of normality was a valid hypothesis. Thus, the use of the statistical model given above was valid.

Table II presents the raw data (i.e., values of $M_1, \sigma_1, M_2, \sigma_2$) for each of the three time warping algorithms, and for cases when both the reference and test utterances were along the x axis. Results are presented for the ten words used in the test. Also included in the table are computed values of threshold, γ and equal error probability $p_M = p_F$.[3] Table III gives the ordered results for the three algorithms for each word in terms of the overall probability of a miss. The most preferred algorithm (rank of 1 or position 1) is the one with the smallest probability of a miss. A summary of the results is given at the bottom of Table III.

Based on the results of Tables II and III, the following conclusions can be drawn:

1) The UE2-1 algorithm performed as well or better than the UELM or the CE2-1 algorithms for almost all the words in the test set. This result was independent of whether the test or the reference was along the abscissa.

2) The CE2-1 algorithm performed as well or better than the UELM algorithm for almost all the words in the test set.

3) All three algorithms tended to perform better when the test utterance was along the abscissa than when the reference utterance was along the abscissa.

IV. DISCUSSION

Based on the experimental data of the preceding section it would seem reasonable to conclude that the modifications to the basic dynamic time warping algorithm lead to improvements in the performance of this method. The most important modification, of course, is the removal of the endpoint constraint. However, the second modification (the UELM algorithm) did not lead to improved performance over the original algorithm—in fact, it led to somewhat degraded performance. Before concluding that the second modification was unsuccessful, several factors which affect the performance of the method should be pointed out. First, the data set on which the tests

[3] Values of p_M were verified experimentally to be within ±2 percent for almost all the cases presented in Table II, thus providing additional evidence that the assumed Gaussian model is valid for this problem.

TABLE III
SUMMARY OF RESULTS

	Reference Along Abscissa			*Test Along Abscissa*			*Lower Error Rate*		
		Rank			Rank				
Word	1	2	3	1	2	3	UELM	UE2-1	CE2-1
7	UE2-1	UELM	CE2-1	UE2-1	UELM	CE2-1	T	T	T
4	CE2-1	UE2-1	UELM	CE2-1	UE2-1	UELM	T	T	T
2	CE2-1	UE2-1	UELM	CE2-1	UE2-1	UELM	T	T	T
X	UE2-1	CE2-1	UELM	UE2-1	CE2-1	UELM	R	R	R
Y	CE2-1	UE2-1	UELM	UE2-1	CE2-1	UELM	T	T	T
C	UE2-1	UELM	CE2-1	UE2-1	UELM	CE2-1	T	T	T
A	UE2-1	UELM	CE2-1	UE2-1	UELM	CE2-1	T	T	T
REPEAT	UELM	UE2-1	CE2-1	UELM	UE2-1	CE2-1	T	T	T
ERROR	UELM	CE2-1	UE2-1	UE2-1	CE2-1	UELM	T	T	T
STOP	UE2-1	UELM	CE2-1	CE2-1	UE2-1	UELM	R	R	R

		Reference Along Abscissa					*Test Along Abscissa*		
		UELM	UE2-1	CE2-1			UELM	UE2-1	CE2-1
		No. of Occurrences of Rank					No. of Occurrences of Rank		
	1	2	5	3		1	1	6	3
Rank	2	4	4	2	Rank	2	3	4	3
	3	4	1	5		3	6	0	4

were run was analyzed semiautomatically, i.e., an automatic endpoint technique provided initial estimates of initial and final frames. However, the experimenter could readily change the endpoints manually if either one was significantly in error. Thus, one of the key features of the UELM algorithm, namely the ability to eliminate spurious sections of sound at the end of an utterance, was never really taken advantage of in this test set. A second point to note is that the UELM algorithm is inherently a natural candidate for a procedure to be used in word spotting applications in which no endpoints at all are specified [7], [8]. For such cases the technique of following the local minimum of the dynamic path is quite reasonable. Thus, although the results presented here were discouraging for the UELM algorithm, further investigations are necessary to examine other potential applications of this technique.

A second result that was fairly conclusive was that lower error rates were achieved when the test utterance was along the abscissa as opposed to the reference utterance. Unfortunately, we have no simple explanation of this result. It is clear that when the test and reference utterances are the same, there is essentially no difference between combinations of reference and test warping curves. However, when the reference and test utterances are different, the differences tend to be emphasized when the test is mapped to the abscissa. Thus, the better overall performance occurs with the test utterance mapped on the abscissa. Interestingly, this is a natural way of handling unknown words in a true word recognition environment since then all the variable duration templates are mapped to the fixed duration test word; hence, no normalization of distance is required.

One additional issue in the performance comparisons between the three dynamic warping algorithms is their computation time. For the algorithms we have discussed and for typical word durations (i.e., $N = M = 40$) the fastest algorithm is the UELM since only a fixed number of distance calculations $(2\epsilon + 1)$ are required per frame of the test (or reference) utterance. The next fastest is the CE2-1 in which the number of distance calculations is on the order of $NM/3$, i.e., the average number of calculations per frame grows proportional to M. Thus, as M (or N) gets bigger, the relative difference in computation time between the UELM and CE2-1 algorithms increases rapidly. Finally, the UE2-1 algorithm requires the most computation for distance calculations. Similar to the CE2-1 method, the relative computation per frame grows linearly with M (or N) but with a larger proportionality constant than for the CE2-1 method.

One final point should be made about the set of algorithms we have investigated. In spite of the desirability of removing the frame size ratio limitation (i.e., M/N or N/M must be less than 2 to 1), the only algorithm which effectively had this capability was the UELM algorithm. The reasons for this are that if we increase the allowable frame ratio, the computation increases significantly, and the control over the type of dynamic path which is used becomes somewhat unwieldly. Furthermore, trying to interpolate or decimate the parameter sets typically used (i.e., LPC parameters) by ratios greater than 2 to 1 tends to lead to severe aliasing distortion. Hence, we have not relaxed this restriction in our work.

V. SUMMARY

In this paper we have shown how some simple modifications can be made to a dynamic time warping algorithm (for speech processing applications) to increase the flexibility and improve the performance of these methods. The modifications consisted of relaxing the time registration constraints at the endpoints of the test and reference utterances, and allowing the dynamic path to follow a locally optimum path at each frame. Using a fairly large data set of isolated words, we have compared and contrasted the performance of three distinct algorithms based on a model of how such techniques would be practically applied. The results indicated improvements in many cases. Implications of the results, for different applications, were discussed.

REFERENCES

[1] H. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *Proc. Int. Cong. Acoust.*, Budapest, Hungary, Paper 20C-13, 1971.

[2] ——, "Comparative study of DP-pattern matching techniques for speech recognition," Speech Res. Group, Acoust. Soc. Japan, Rep. S73-22, 1973.

[3] F. I. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67-72, Feb. 1975.

[4] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, pp. 297-315, Feb. 1975.

[5] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," *Ann. Math. Stat.*, vol. 19, pp. 279-281, 1948.

[6] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43-49, Feb. 1978.

[7] J. S. Bridle, "An efficient elastic-template method for detecting given words in running speech," *Proc. British Acoust. Soc.*, Apr. 1973.

[8] R. W. Christiansen and C. K. Rushforth, "Detecting and locating key words in continuous speech using linear predictive coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 361-367, Oct. 1977.