An Investigation of the Use of Dynamic Time Warping for Word Spotting and Connected Speech Recognition

C. S. Myers

L. R. Rabiner

A. E. Rosenberg

Acoustics Research Department
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

Several variations on algorithms for dynamic time warping have been proposed for speech processing applications. In this paper two general algorithms that have been proposed for word spotting and connected word recognition are studied. These algorithms are called the fixed range method and the local minimum method. The characteristics and properties of these algorithms are discussed. It is shown that, in several simple performance evaluations, the local minimum method performed considerably better then the fixed range method. Explanations of this behavior are given and an optimized method of applying the local minimum algorithm to word spotting and connected word recognition is described.

I. Introduction

The technique of dynamic programming for the time registration of a reference and a test pattern has found widespread use in the area of automatic speech recognition. Work by Sakoe and Chiba [1], Itakura [2] and White and Neely [3] has shown that dynamic time warping (DTW) algorithms may be successfully applied in isolated word recognition systems. Bridle [4] and Christiansen and Rushforth [5] have demonstrated effective DTW algorithms for word spotting, and recently, Sakoe [6] and Rabiner and Schmidt [7] have successfully applied time warping techniques to connected digit recognition. There are many factors which determine the performance of a DTW algorithm for such applications. It is the purpose of this paper to study these factors, and to determine "optimal" choices for some of the parameters of a time warping algorithm for word spotting and for connected word recognition applications.

The basic difficulty in both word spotting and connected word recognition (based on matching isolated word reference template patterns) is illustrated in Figure 1. Here we show the time pattern of the log intensity for two speech utterances, "3," "8," in part a, and "38" in part b. The utterance in part a is spoken with a discernible pause between the "3" and the "8," while the utterance of part b is spoken with no discernible pause between the "3" and the "8." Dynamic time warping algorithms, as they have been applied to isolated word recognition applications, require a reliable set of beginning and ending points in order to match a reference pattern to an unknown test pattern. However, as seen in part b of Figure 1, a reliable segmentation for the utterance "38" is difficult to obtain. (Furthermore, even if a reliable segmentation of the connected sequence could be obtained, the coarticulation between words of the sequence would further complicate the DTW match to an isolated word reference pattern.)

The form of time warping which will be used to overcome such a segmentation problem is illustrated in Figure 2. A reference pattern, $\mathbf{R}(n)$, represented as a time function of a multi-dimensional feature vector, is to be time registered with a test pattern, $\mathbf{T}(m)$, also represented as a time function of a multi-dimensional feature vector. It is assumed that the basic template unit is a word; however, the results presented here can equally be applied to other units of speech (e.g. syllable, phone, phrase, etc.). For purposes of time alignment, it is assumed that we have determined a beginning region, of size B, (as shown in Figure 2) with potential starting frames between b_1 and b_2 ($B = b_2 - b_1 + 1$). It is also assumed that we have determined an ending region, of size E, with potential ending frames between e_1 and e_2 ($E = e_2 - e_1 + 1$). Thus, the best match to the reference pattern can begin at some frame in the specified beginning region and end at some frame in the specified ending region. It is, of course, possible to expand either the beginning or the ending region to incorporate the entire test pattern. Thus, the framework of Fig. 2 may be used for either word spotting, in which neither the beginning nor the ending frames are known, or for connected word recognition, in which the ending frame of one word is used to postulate the beginning frame of the next. Given a beginning and ending region, the purpose of the DTW algorithm is to determine (in an optimal and efficient manner) the warping function, or contour, which provides the best time alignment between the reference and the test pattern.

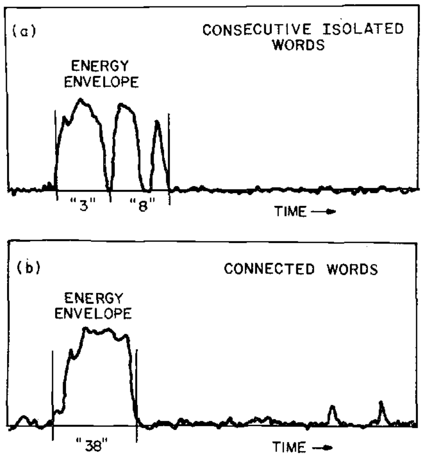

Fig. 1. Log energy for two speech utterances.

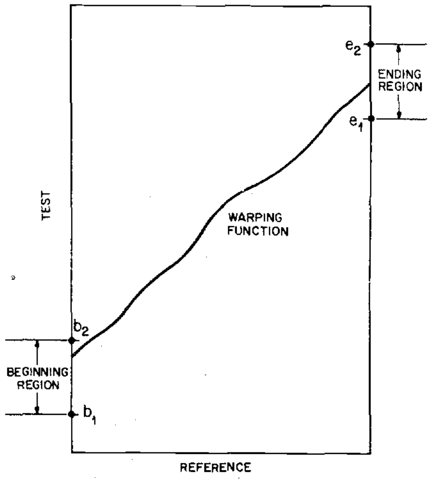

Fig. 2. Illustration of the general time warping problem.

CH1559-4/80/0000-0173$00.75 © 1980 IEEE

II. Dynamic Time Warping Algorithms

As discussed in Section I, the purpose of the DTW algorithm is to provide the optimal time alignment between a reference pattern $R(n)$, $n = 1,2,\ldots,N$, and some portion of a test pattern $T(m)$, $m = 1,2,\ldots,M$ where N and M are the lengths of the reference and test pattern respectively, in frames. (Typically a frame represents from 10 to 30 milliseconds of speech.) The time registration contour is best represented by a parameterized path, $(i(k),j(k))$, where $n = i(k)$, $m = j(k)$, and k is the index of the parameterized curve. The degree of similarity achieved by a time warping path of length K is given by the overall distance function, $D(i(k),j(k))$, where

$$D(i(k),j(k)) = \frac{\sum_{k=1}^{K} d(i(k),j(k))\,\tilde{W}(k)}{N(\tilde{W})} \qquad (1)$$

and in which $d(i(k),j(k))$ is a local distance metric used to measure the degree of dissimilarity between the feature vectors of frames $i(k)$ of the reference pattern and $j(k)$ of the test pattern, $\tilde{W}(k)$ is a set of weights used along the parameterized path, and $N(\tilde{W})$ is a normalization factor, based on the set of weights chosen.

Several parameters must be specified in Eq. (1) in order to determine the optimal time warping path. These include the initial and final values of $i(k)$ and $j(k)$, the weighting function $\tilde{W}(k)$ and the range of possible time warping functions (i.e. restrictions on the slope and shape of the time alignment path).

For the initial and final values of $i(k)$ and $j(k)$ we shall use our definition of the beginning and ending regions as follows:

$$i(1) = 1, \quad j(1) = b \quad b_1 \leqslant b \leqslant b_2 \qquad (2a)$$
$$i(K) = N, \quad j(K) = e \quad e_1 \leqslant e \leqslant e_2. \qquad (2b)$$

Hence the warping contour begins at the first reference frame and within the beginning region of the test, and ends at the last reference frame and within the ending region of the test.

We shall use several different local constraints to specify the range of the time warping path, including those proposed by Sakoe and Chiba [8] and Itakura [2]. A comprehensive discussion of this area is given by Myers [9] and will not be repeated here. Based on earlier work by Sakoe and Chiba [8], it is assumed that the overall slope of the time warping function is constrained to be between 1/2 and 2. Figure 3 illustrates the effect of this slope limitation on the range of legal paths. It is seen that for each possible starting and ending frame pair, b, e, the optimal path is constrained to lie within a parallelogram defined by lines of slope 2 to 1 and 1/2 to 1, beginning at b and ending at e.

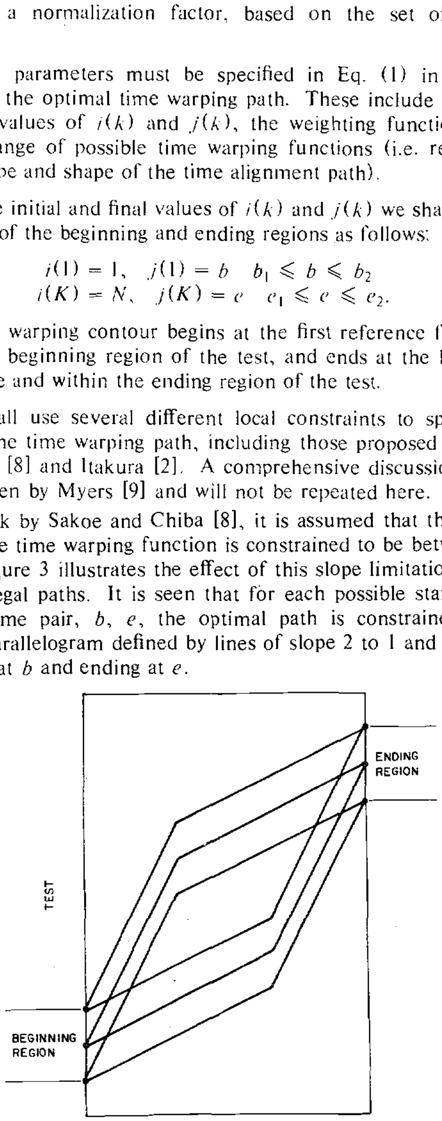

Fig. 3. Range available for a time warping function for several beginning and ending points.

In the most general case, it is necessary to determine the optimal time warping path for every possible pair of beginning and ending points, i.e.

$$\hat{D} = \min_{b_1 \leqslant b \leqslant b_2} [\min_{e_1 \leqslant e \leqslant e_2} [D(i(k),j(k)) \; s{\cdot}t{\cdot} \; j(1) = b, \; j(K) = e]] (3)$$

where $\hat{D}$ is the distance score of the best possible path using any possible beginning and ending point pair. Although the amount of computation required to solve Eq. (3) may be excessive, it is possible, with some reasonable choices of the weighting function $\tilde{W}(k)$, and the normalization $N(\tilde{W})$, to greatly simplify the computational process. If $\tilde{W}(k)$ and $N(\tilde{W})$ are chosen as

$$\tilde{W}(k) = i(k) - i(k-1) \qquad (4a)$$
$$\tilde{W}(1) = 1 \qquad (4b)$$
$$N(\tilde{W}) = N \qquad (4c)$$

then the solution to Eq. (3) may be generated by a *single* application of a DTW algorithm using the extended region shown in Figure 4.

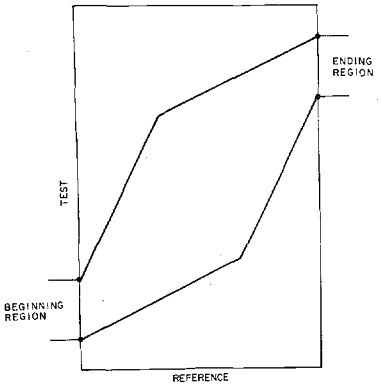

Fig. 4. Expanded range for a single time warp.

In spite of the large amount of savings in computation provided by the reduction to a single application of the DTW algorithm, the overall size of the single time warp is still rather large (greater than $N^2/3$ points in the overall range if the ending region is known and greater than $3N^2/2$ points if the ending region is unspecified). Further modifications have been suggested in order to reduce this amount of computation. In particular, Sakoe and Chiba [1] have proposed that a time warping path not be allowed to create excessive time differences, i.e. for any $i(k)$, $j(k)$ is restricted such that

$$|j(k) - i(k) - \bar{b} + 1| \leqslant R \qquad (5)$$

where $\bar{b}$ is the center of the beginning region ($\bar{b} = (b_1 + b_2)/2$) and R is the maximum time difference, in frames, which is allowed. (R is defined to cover the beginning region, i.e. $2R+1 => B$.) This limitation is illustrated in part a of Figure 5, and is referred to as the *fixed range* DTW algorithm. Another range reduction technique proposed by Rabiner, Rosenberg, and Levinson [10] is illustrated in part b of Figure 5. Here the overall time warping path is constrained to be within a fixed range around the best path so far, i.e. the local minimum. Formally,

$$|j(k) - c(k)| \leqslant \epsilon \qquad (6a)$$
$$c(k) = \operatorname*{argmin}_{j}[D_A(i(k)-1,j)] \qquad (6b)$$
$$c(1) = \bar{b} \qquad (6c)$$

where $D_A(i,j)$ is the accumulated distance to a point (i,j) via the best path, as computed by the DTW algorithm, $\operatorname*{argmin}_{x}[F(x)]$ is the value of x that minimizes $F(x)$, ϵ is the range allowed around the local minimum, also chosen to cover the beginning region, i.e. $2\epsilon + 1 => B$, and $c(k)$ is the position along the j axis of the local minimum of $D_A(i(k)-1,j)$. Thus, if $D_A(i,j)$ is computed in consecutive vertical strips (i.e. i is fixed and j is varied), then the range of one vertical strip is $\pm\epsilon$ about the local minimum of the previous vertical strip. This algorithm is referred to as the *local*

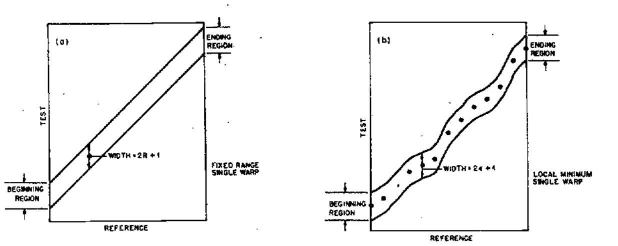

Fig. 5. Illustration of the fixed range and the local minimum DTW algorithms.

minimum DTW algorithm.

Two fundamental differences exist between these two algorithms. The fixed range DTW apriori specifies the ending region by specifying the beginning region, i.e.

$$E = B = 2R + 1 \qquad (7a)$$

$$e_1 = b_1 + N \qquad (7b)$$

$$e_2 = b_2 + N \qquad (7c)$$

while the local minimum DTW algorithm defines the ending region implicitly from the local minimum of the last vertical strip (the last frame of the reference), i.e.

$$E = 2\epsilon + 1 \qquad (8a)$$

$$e_1 = c(K) - \epsilon \qquad (8b)$$

$$e_2 = c(K) + \epsilon \qquad (8c)$$

Thus, we might expect the local minimum algorithm to be able to handle more cases than the fixed range method, i.e. those cases in which the conditions of Eqs. (7) are not valid.

The other major difference between the two time warping algorithms involves the number of time warps required to cover a beginning region. For the fixed range DTW algorithm the entire beginning region is most efficiently covered in one time warp with $2R + 1 = B$, since adjacent time warps may be merged together without loss of accuracy. However, an analogous specification of the local minimum DTW algorithm ($2\epsilon+1=B$) may not be truly optimal. Such a specification allows only one local minimum path to be followed and may result in erroneous choices for the time alignment path because the true path may be "lost," i.e. the globally optimal path need not be the locally optimal path nor even with ϵ frames of the local minimum. As such, it may be better to try several smaller local minimum time warps, thus allowing several different local minimum paths to be tried, and to compare the results of the different paths in order to determine the proper path. Such a procedure is illustrated in Figure 6. It is assumed that NTRY local minimum time warps are to be computed. Each has a local range of $\pm\epsilon$ about their respective local minima, and the centers of two adjacent time warps are initially separated by δ. The entire beginning region covered by the NTRY local minimum time warps is given by $\Delta = 2\epsilon + 1 + (\text{NTRY}-1)\cdot\delta$. To cover a specified beginning region, NTRY, ϵ and δ are chosen such that $\Delta = B$.

III. Considerations in the Dynamic Time Warping Algorithms

There are several questions raised by the two dynamic time warping algorithms which we have discussed. Among the issues which are important for both word spotting and connected word recognition applications are:

1. Which of the two algorithms, (i.e. the fixed range DTW algorithm or the local minimum DTW algorithm), gives better performance results when applied to a series of recognition and word spotting experiments?
2. In the local minimum DTW algorithm, for a given Δ, what are the optimal choices of ϵ, NTRY and δ? In particular, the main question is whether more than one time warp is required, and, if so, how should the parameters ϵ, δ, and NTRY be chosen?

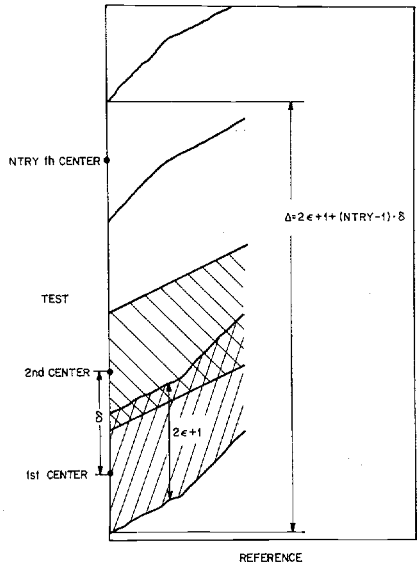

Fig. 6. Illustration of the parameters in the local minimum DTW algorithm.

In order to answer these questions we have performed several speech recognition and word spotting experiments. The results of these experiments will be described in the next section.

IV. Experimental Results

In our initial experiment we compared the recognition accuracy achieved by both the fixed range DTW algorithm and the local minimum DTW algorithm for a modified *isolated* word recognition problem. The test utterances consisted of 54 words from a vocabulary of computer terms, spoken by each of 4 talkers, for a total of 216 utterances. The test utterances were recorded over a dialed-up telephone line, digitized at 6.67 kHz, analyzed every 15 milliseconds with an 8^{th} order LPC analysis and local distance scores were calculated using Itakura's log likelihood ratio [2]. The reference patterns consisted of 2 templates per word of the vocabulary formed from a speaker independent clustering technique.

In order to test the effectiveness of the two DTW algorithms the test utterances were modified so that a beginning region could be specified as some range about the true beginning point. The ending region was left unspecified. For the sake of comparison, R and ϵ were both set equal to 8 frames and NTRY was set to 1. Figure 7 shows the recognition results for both algorithms as a function of four different local constraints.[1] We observe that the local minimum DTW algorithm performed better than the fixed range DTW algorithm for *all* local constraints.

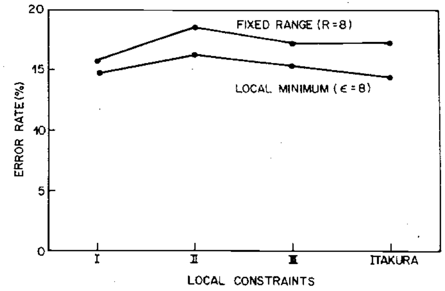

Fig. 7. Results for word recognition using both the fixed range and the local minimum DTW algorithms.

[1] Local constraints types I, II and III have been defined by Myers [9] as productions in a regular grammar and Itakura's local constraints have also been defined previously [2]. All the local constraints used have the common characteristic of limiting the slope of the warping function to between 1/2 and 2.

In another comparison we artificially (and at an arbitrary frame) imbedded an isolated digit into a connected digit utterance. We then used both DTW algorithms to "spot" the imbedded digit. The parameters used were the same as in our first experiment (ϵ=8, R=8). We tried every possible beginning region of size $2\epsilon + 1$ (=$2R$+1) and recorded the number of times that the best possible path (as determined by the lowest overall distance achieved by any starting region) was obtained. Figure 8 gives the results of this experiment. We observe that the local minimum algorithm almost always found the best path more often than the fixed range algorithm. We also observe that the local minimum DTW algorithm was able to find the best path 17 times (the maximum number possible, 2ϵ+1) in 8 out of 10 trials, while the fixed range algorithm never achieved this accuracy. From the results of these two experiments, it appears that the local minimum DTW algorithm performs better than the fixed range DTW algorithm.

In order to understand the effects of various combinations of the parameters Δ, δ, NTRY and ϵ on the performance of the local minimum DTW algorithm, a series of connected digit recognition experiments was performed. A total of 80 sequences of from 2 to 5 random digits (20 of each length), each spoken as a single connected utterance, were recorded by each of two talkers [7]. The experiment first "spotted" the initial digit in each utterance via a local minimum algorithm (ϵ=11, NTRY=1) using the known beginning point. Then an attempt was made to recognize the second digit in the utterance. The beginning region of the second digit was centered around the ending region of the first digit, as determined by the spotting procedure. Several values of ϵ, δ, Δ and NTRY were used and recognition scores were measured.

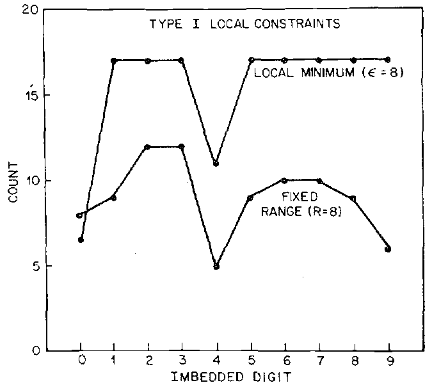

Fig. 8. Results for word spotting using both the fixed range and the local minimum DTW algorithms.

Figure 9 illustrates the major result of these experiments. For a large value of Δ, (27 in this case), the average best distance score for all NTRY warps was compiled and plotted as a function of δ, for several values of ϵ, and according to either the reference being the same word as the second word in the test string or the reference being different. Examination of Figure 9 shows that the best distance for both same words and different words increases as δ increases. However, we observe that when the reference is different from the second digit of the test utterance, the average distance is generally increasing in δ, but that when the reference and the test word are the same, the distance score is constant for small values of δ and increases beyond the critical value $\delta = 2\epsilon + 1$. This is a particularly important value of δ because for $\delta < 2\epsilon + 1$ consecutive time warps overlap in their beginning regions, and for $\delta > 2\epsilon + 1$ there are frames between two consecutive time warps which are not covered by either beginning region. From these results we can conclude that there is essentially no loss in performance of the local minimum DTW algorithm, i.e., the minimum distance path is generally found when consecutive starting points are separated by $2\epsilon + 1$, (no overlap).

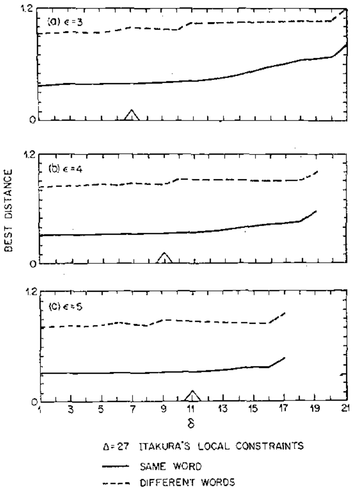

Fig. 9. Distance scores for the local minimum DTW algorithm as applied to connected word recognition.

The reason that overlapping beginning regions is unnecessary is illustrated in Figure 10. Here we show the progress of a typical case in which the starting regions overlap. By the nature of the local minimum DTW algorithm, overlapping time warps tend to converge if there is a good path common to both of their beginning regions. Such a result is directly applicable to the word spotting problem in which ϵ has already been determined to minimize some error measure (e.g. number of false alarms, number of misses, etc.) and in which the only important question is how often to sample the test. We conclude that, when several time warps must be performed in a word spotting problem in order to locate a key word which may occur more than once in the test utterance, the most efficient method is to use nonoverlapping beginning regions and a single local minimum time warp per beginning region.

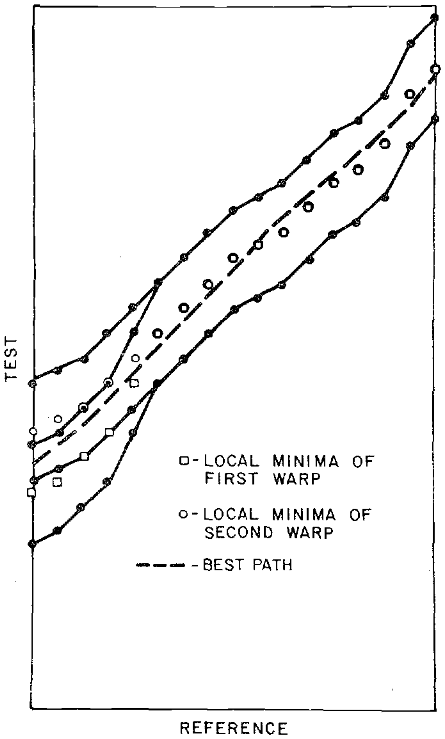

Fig. 10. Illustration of path merging for two adjacent local minimum time warps.

The connected word recognition problem is somewhat different than the word spotting problem in that the beginning region is not the entire test utterance but can be, in general, reduced to a considerably smaller range. For such a situation we would like to have a simple rule for determining the optimal choices of δ, ϵ and NTRY for a given Δ. In Figure 11 we plot recognition error rates for the second digit as a function of Δ and for two cases, $\epsilon = (\Delta-1)/2$ (NTRY=1) and the best results for any ϵ, δ and NTRY combination. We see that for smaller values of Δ, a single warp performs as well as any combination of ϵ, δ and NTRY, and that as Δ increases the differences in error rates between the best possible ϵ, δ and NTRY combination and a single warp remains less than 2.5%. Since a single time warp is computationally more efficient than several time warps and since it should be possible to define an accurate beginning region of a word to within 255 milliseconds (Δ=17 frames, at 15 milliseconds between frames), a single local minima time warp is a reasonable approach to the problem of connected word recognition using word size reference templates. In fact, though we show the minimum error rate at $\Delta = 21$, work by Rabiner and Schmidt [7] on connected digit recognition has shown that it is possible to reduce Δ, without loss in accuracy, by more judicious positioning of the beginning region than simply centering it about the ending region of the previous word.

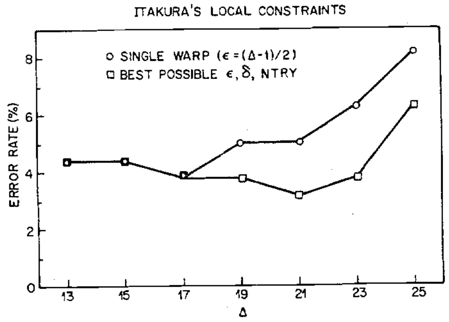

Fig. 11. Error rates for connected word recognition using the local minimum DTW algorithm.

V. Conclusion

We have shown that a local minimum dynamic time warping algorithm performs considerably better than a fixed range dynamic time warping algorithm for both word spotting and speech recognition. We have also shown that consecutive trials of a local minimum DTW algorithm need not overlap in order to achieve performance comparable to overlapping trials. Finally, we have demonstrated that, given a good estimate of the beginning region of a word, a reasonable approach to connected word recognition is to simply use one local minimum time warp to find the best time alignment over the specified beginning region.

References

[1] H. Sakoe and S. Chiba, "A Dynamic Programming Approach to Continuous Speech Recognition," *Proceedings of International Congress on Acoustics*, Budapest, Hungary, Paper 20C-13, 1971.

[2] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-23, pp. 57-72, February 1975.

[3] G. M. White and R. B. Neely, "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering and Dynamic Programming," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-24, pp. 183-188, April 1976.

[4] J. S. Bridle, "An Efficient Elastic Template Method for Detecting Given Words in Running Speech," *Proceedings of British Acoustical Society Meeting*, London, England, Paper 73SHC3, April 1973.

[5] 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, October 1977.

[6] H. Sakoe, "Two-Level DP-Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, December 1979.

[7] L. R. Rabiner and C. E. Schmidt, "Application of Dynamic Time Warping to Connected Digit Recognition," to be published.

[8] H. Sakoe and S. Chiba, "Dynamic Programming Optimization for Spoken Word Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, pp. 43-49, February 1978.

[9] C. S. Myers, "A Comparative Study of Several Dynamic Time Warping Algorithms for Speech Recognition," Masters Thesis, MIT, February 1980.

[10] L. R. Rabiner, A. E. Rosenberg and S. E. Levinson, "Considerations in Dynamic Time Warping for Discrete Word Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, pp. 575-582, December 1978.