Connected Word Recognition Using a
Level Building Dynamic Time Warping Algorithm

C. S. Myers
L. R. Rabiner

Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

The technique of dynamic time warping has proven itself reliable and robust for a wide variety of isolated word recognition tasks. Recently extensions of the algorithm have been investigated for application to the problem of connected word recognition. In this paper a level building technique is proposed for optimally aligning a test pattern, consisting of a sequence of connected words, with a sequence of isolated word reference patterns. This algorithm is shown to be significantly more efficient than the one proposed by Sakoe while solving the exact same problem. Implementation parameters for the level building algorithm are presented and the effectiveness of the proposed algorithm for connected digit recognition is experimentally verified.

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. In particular, Sakoe and Chiba [1], Itakura [2] and White and Neely [3] have shown how dynamic time warping (DTW) algorithms can be applied in isolated word recognition systems. Recently, Sakoe [4] and Rabiner and Schmidt [5], and Bridle and Brown [6] have investigated extensions of the principle of dynamic time warping in order to recognize strings of words in connected speech. Sakoe's approach, called the 2-level DP warp method, exhaustively tries to match all reference patterns to all possible subsections of the test string (the first level), and then, based on the distance scores generated in the first pass, determines the best match to the spoken string (the second pass). Rabiner and Schmidt use a sampling approach by attempting to build up candidate strings in a left to right manner. The only regions of the test string for which DTW matches are tried are those regions starting near the end of good matches at the preceding level. Thus, only certain "sample" points are used as potential beginning regions for matching a reference word to the spoken string, and the recognized string is built up word-by-word.

It is the purpose of this paper to describe another approach to dynamic time warping for connected word recognition problems. The proposed method, which we call the level building algorithm, can be shown to be related to the stack decoding algorithm of Bahl and Jelinek [7], which has been proposed for use in continuous speech recognition systems. The level building algorithm has been shown to be both an efficient implementation of the two-level DP warp algorithm of Sakoe, and more general, but as efficient as the "sampling" algorithm proposed by Rabiner and Schmidt [8,9].

In this paper we show how the level building DTW algorithm may be derived from examination of dynamic time warping for isolated word recognition and show how simple modifications may be imposed on the basic level building algorithm to account for variable length strings and also to give multiple candidate strings. A brief discussion of the implementation parameters of the level building algorithm is given and the results of an experimental evaluation of the level building algorithm in a connected digit recognition task is presented.

II. The Level Building Dynamic Time Warping Algorithm

The problem of dynamic time warping for connected word recognition is to find the sequence of L reference patterns, $\mathbf{R}_{q(1)}$, $\mathbf{R}_{q(2)}, \ldots, \mathbf{R}_{q(L)}$, which best matches a given test pattern $\mathbf{T}(m)$, $m=1,2,\ldots,M$ where $\mathbf{T}(m)$ is a vector of features for frame m and M is the number of frames in the test pattern. Each of the reference patterns $\mathbf{R}_{q(l)}$ is chosen from a set of V reference patterns $\mathbf{R}_v$, $v=1,2,\ldots,V$. In all of our work we have assumed that each reference pattern corresponds to a single word spoken in isolation and that these words may be combined in any order. The algorithm which we present here may, however, be easily extended to other units of speech besides words and also may be extended to incorporate syntactic constraints [7,10].

Formally, the problem of connected word recognition is to find that sequence of reference patterns $\mathbf{R}_{q(1)}$, $\mathbf{R}_{q(2)}, \ldots, \mathbf{R}_{q(L)}$ which minimizes $D_{q(1)q(2)\ldots q(L)}$ where $D_{q(1)q(2)\ldots q(L)}$ is the DTW distance between the test pattern $\mathbf{T}(m)$ and the super reference pattern $\mathbf{R}^s$ formed by concatenating $\mathbf{R}_{q(1)}$, $\mathbf{R}_{q(2)},\ldots$, $\mathbf{R}_{q(L)}$, i.e. $\mathbf{R}^s = \mathbf{R}_{q(1)} \oplus \mathbf{R}_{q(2)} \oplus \cdots \oplus \mathbf{R}_{q(L)}$. The dynamic time warping distance $D_{q(1)q(2)\ldots q(L)}$ is given by

$$D_{q(1)q(2)\ldots q(L)} = \min_{w(m)} \left[\sum_{m=1}^{M} d(m,w(m)) \right] \tag{1}$$

where $w(m)$ is the warping function which maps $\mathbf{T}(m)$ to $\mathbf{R}^s$ and where $d(m,w(m))$ is the local distance between frame m of $\mathbf{T}$ and frame $w(m)$ of $\mathbf{R}^s$. Typically, $w(m)$ is restricted to match the endpoints of $\mathbf{T}$ and $\mathbf{R}^s$ and also its slope is restricted to some nonnegative range. Figure 1 illustrates a typical example in which the slope of $w(m)$ is restricted to be between ½ and 2, as shown by the parallelogram.

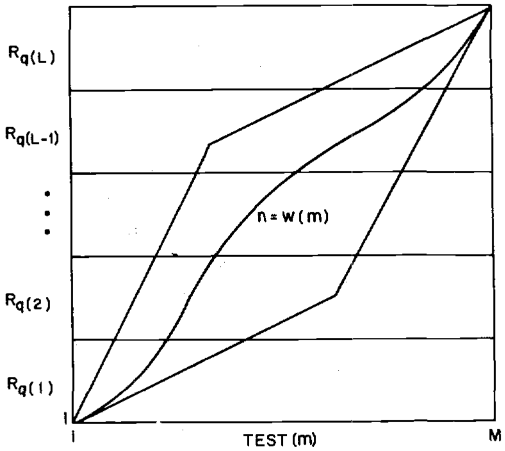

Fig. 1 Illustration of constraint endpoint DTW match of a sequence of reference patterns to a connected word test string.

For any *given* $\mathbf{R}^s$ the problem of determining $D_{q(1)q(2)\ldots q(L)}$ is the same as the constrained endpoint isolated word dynamic time alignment problem since $\mathbf{R}^s$ may be considered as a single reference pattern. Thus, in theory, it is possible to solve the connected word recognition problem by exhaustively trying all V^L possible $\mathbf{R}^s$. However, even for modest values of V or L this amount of computation is intractable.

In order to see how to efficiently solve the connected word recognition problem we must examine the way in which a DTW algorithm is generally implemented for a fixed $\mathbf{R}^s$ and $\mathbf{T}$. Figure 2a shows a typical implementation of a constrained endpoint DTW algorithm. Generally the computation to find the optimal warping path is performed in vertical stripes (i.e. m is indexed sequentially and a range on n is found in which the path is constrained to lie) as illustrated in this figure. An alternative way in which the computation may be performed is illustrated in Fig. 2b. For this case the computation is done in vertical stripes again; however, the horizontal lines formed by the end of each reference pattern in $\mathbf{R}_s$ form constraints on the regions in which the computation is done. As such, the computation is initially done in vertical stripes until the partial region G_1 is completed. In order to correctly start up the computation for the second reference pattern (i.e. in the region G_2) the accumulated distance scores for all paths that end at the first horizontal line (denoted by the heavy dots) must be retained at the end of the first level and used as initial conditions for the second level. In this manner, the entire dynamic time alignment may be carried out by levels (i.e. by successive concatenation of reference patterns within $\mathbf{R}_s$) in a series of computations.

The significance of the above results is that the level building approach to finding the best dynamic time alignment path can be extended to the case of more than one reference pattern at each level, as illustrated in Figure 3. Figure 3 shows how a set of V reference patterns can be tried at the first level to find the best set of partial matches to a portion of the test. As shown in Fig. 3a, for reference pattern $\mathbf{R}_1$, the algorithm must keep track of the accumulated distance for all paths that end at the grid points (m,N) - i.e. at the end of reference pattern 1. The range of m for which such paths end is $m_{11}(1) \leq m \leq m_{12}(1)$, as determined by the intersection of the line $n = N_1$ with the lower and upper warping function constraint lines. Similarly, as shown in Figure 3b, for reference pattern $\mathbf{R}_2$, the algorithm must keep track of the accumulated distance for all paths that end at the end of reference pattern 2. This process is repeated for all reference patterns $\mathbf{R}_v$, $v=1,2,...,V$ and an overlap range on m, $m_1(1) \leq m \leq m_2(1)$ is determined.

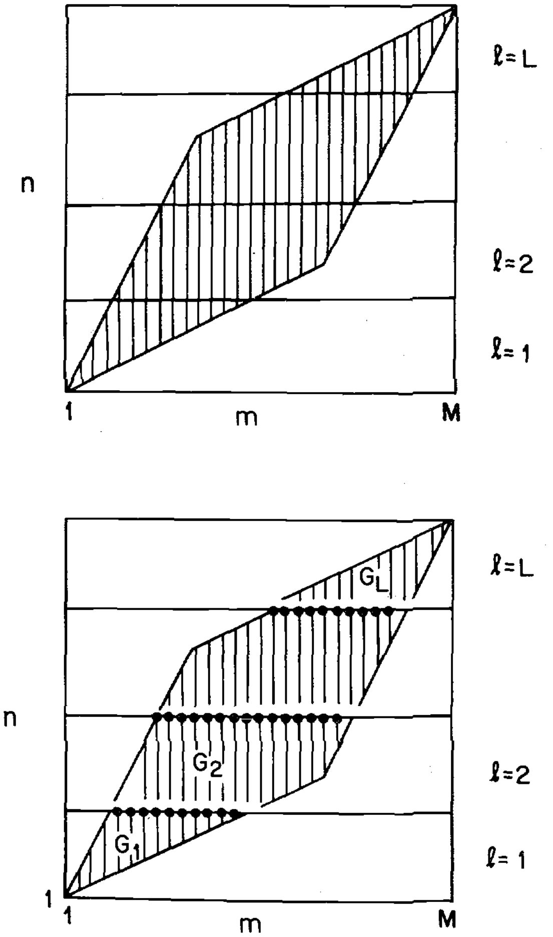

Fig. 2 Two possible implementations of constrained endpoint DTW algorithm.

For each value of m in the ending range $m_1(\ell) \leq m \leq m_2(\ell)$, at level ℓ, we must keep track of 3 quantities, namely

1. Minimum accumulated distance, $\tilde{D}_\ell^B(m) = \min\limits_v [\tilde{D}_\ell(m)]$ where $\bar{D}_\ell^v(m)$ is the accumulated distance for the v^{th} reference pattern, at level ℓ, ending at frame m of the test pattern.
2. Best reference, $W_\ell(m) = \operatorname*{argmin}\limits_v [\tilde{D}_\ell^v(m)]$ where $\operatorname*{argmin}\limits_v [f(x)]$ is that value of x that minimizes $f(x)$.
3. Backtracking pointer, $\tilde{F}_\ell^B(m) = \tilde{F}_\ell^{W_\ell(m)}(m)$ where $\tilde{F}_\ell^v(m)$ points to the frame of the test pattern at level $\ell-1$ at which the best path to the test frame m, at level ℓ, using reference $\mathbf{R}^v$ ended, i.e. the best path to frame m of the test pattern, at the end of the ℓ^{th} level, using reference $\mathbf{R}_v$, began at frame $\tilde{F}_\ell^v(m) + 1$. For level $\ell=1$, it should be clear that $\tilde{F}_\ell^v(m) = 0$ since all paths started at frame 1 of the test pattern.

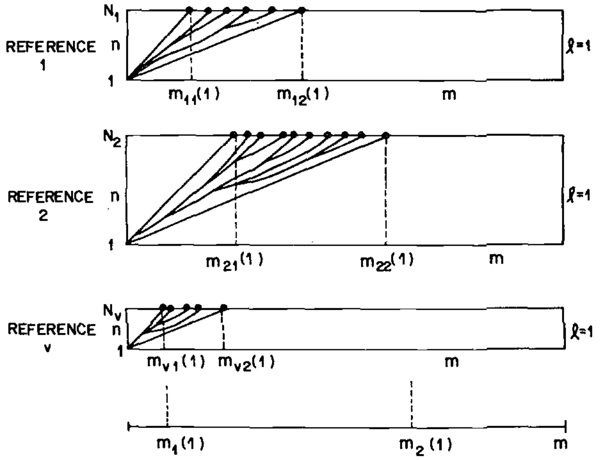

Fig. 3 Sequence illustrating the match regions and the resulting ending regions used in the first level of the level building DTW algorithm.

The purpose of the vector of distances $\tilde{D}_\ell^B(m)$ is to initialize the next level of the dynamic time alignment procedure just as discussed in the case of isolated word dynamic time warping. The associated word vector $W_\ell(m)$ is used to record the best word associated with each accumulated distance. $\tilde{D}_\ell^B(m)$ is used to determine the optimal reference string once the optimal path has been determined. The vector of back pointers $\tilde{F}_\ell^B(m)$ is used to determine the optimal path by tracing the path back from the end (level L and frame M). Myers and Rabiner have described a method by which the backpointers may be computed simultaneously with the accumulated distance vector with only a small increase in storage costs [8].

Figure 4 illustrates the operation of the level building algorithm on the second level. Here the distance vector $\tilde{D}_1^B(m)$ gives a set of initial conditions and paths may now begin at any frame within the starting region $m_1(1) \leq m \leq m_2(1)$. The algorithm keeps track of the total accumulated distance for each path and determines a new starting range for the next level.

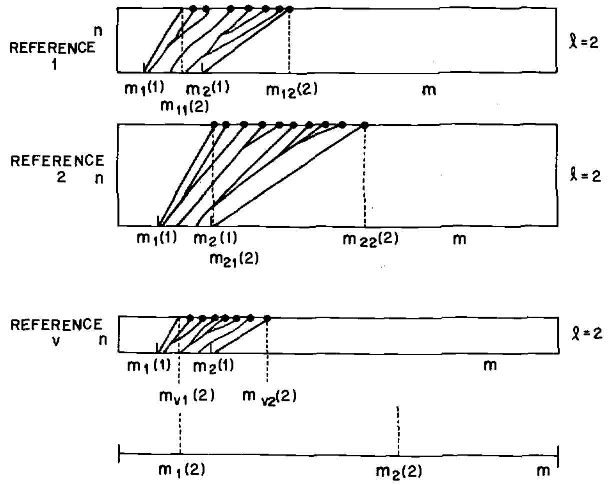

Fig. 4 Sequence illustrating the match regions and resulting ending regions used in the second level of the level building DTW algorithm.

Figure 5 illustrates the level building algorithm for a simple example in which there are only 2 reference patterns, A and B, each of equal length. It is assumed that a string of length L=4 is known to have been spoken. Figure 5 shows that at the end of the first level there are 6 possible ending values of m and the reference pattern giving the smallest distance is denoted along the horizontal line at the end of the level. Similarly, at levels 2 and 3, the best path to each possible ending frame are noted by the reference, at that level, which gave the minimum accumulated distance. Finally, at level 4, only a single path is retained, as this is the optimal path. To determine the best matching string, we must backtrack the path ending at m=M to give the sequence BAAB as the optimal sequence of 4 reference patterns to match the test pattern. Also denoted on Figure 5 are the test frame values e_ℓ, corresponding to the end of each reference in the best matching sequence. In principle, these values, e_ℓ, could be used as best estimates of segmentation points between entries in the test pattern.

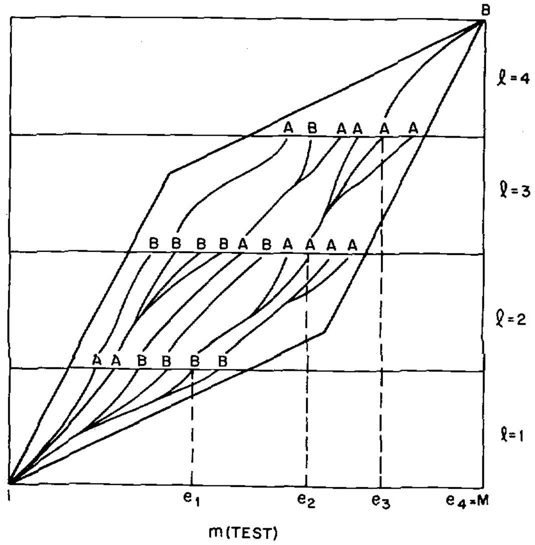

Fig. 5 Illustration of a 4 level DTW match.

Certain simple extensions of the level building algorithm provide for much more general usage. It is possible to get the best string of any length by simply comparing the accumulated distance scores associated with each of the potential lengths. In addition, it is possible to generate more than one candidate string of each length. By keeping track of not only the best reference pattern at each level and ending frame but also the second best, it is possible, in the traceback stage, to generate several second choice candidates by using at each level, the second best candidate rather than the best candidate at that level.

Further modifications to the basic structure of the level building algorithm may included in order to increase both its efficiency and its flexibility. To describe these modifications we have defined a set of variables which influence the performance of the level building algorithm. Some of these variables are illustrated in Figure 6. The variables δ_{R_1} and δ_{R_2} define regions, at the beginning and end of each reference pattern, in which the local path can begin or end - i.e. paths need not begin at frame 1 of each reference nor end at the last frame, but instead the best beginning and ending frames, within the specified regions, are found and used for each path.

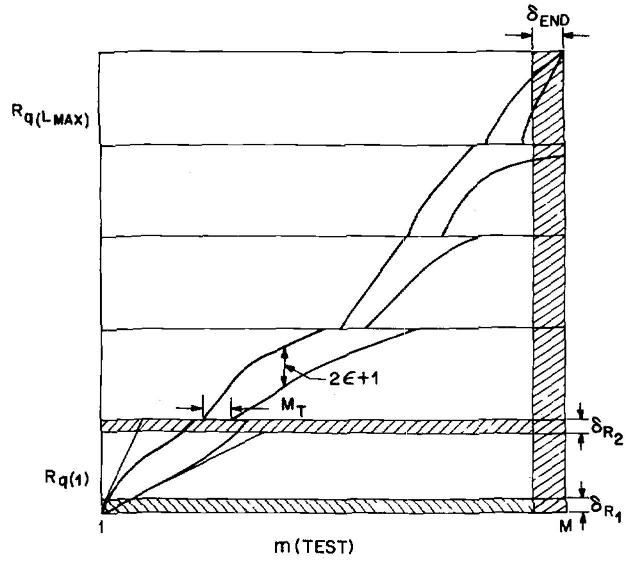

Fig. 6 Illustration of the level building parameters δ_{R_1}, δ_{R_2}, δ_{END}, M_T and ϵ.

Similarly, the parameter δ_{END} defines a region at the end of the test pattern in which a total match can end, rather than strictly requiring each path to end at the frame m=M. This added flexibility allows for some margin of error in determining the ending frame of the test pattern.

The parameters M_T and ϵ are range reduction parameters which reduce the size of the local regions, G_ℓ, in which the dynamic path is constrained to lie. The parameter M_T ($M_T \geq 1$) is used to reduce the initial starting range at each level to only those frames for which the average accumulated distance to that frame, $\tilde{D}^B_\ell(m)/m$, is within a factor of M_T of the best average accumulated distance ϕ_ℓ, where

$$\phi_\ell = \min_m \left[\frac{\tilde{D}^B_\ell(m)}{m}\right] \tag{2}$$

Similarly, the parameter ϵ is used to reduce the size of the region G_ℓ during the computation by defining a range of size $2\epsilon + 1$ for each vertical strip. At each frame m along the test, the range of n is determined by examining a region within ± ϵ of the minimum accumulated distance in column $m-1$. This parameter was originally used in the UELM DTW algorithm and is discussed in detail by Rabiner and Schmidt [5].

In addition to the illustrated parameters, the parameters *TMIN* and *TMAX* are used to terminate the DTW search on a given reference pattern if the best distance in the vertical strip for frame *m* exceeds $TMIN \cdot m + TMAX$. These parameters are used to eliminate non-productive searches.

III. Computational Comparison of Connected Word DTW Algorithms

It is very important to note that the level building algorithm solves the exact same problem as Sakoe's 2 level DP-Matching algorithm. This has been formally shown by Myers and Rabiner [8]. As such, it is worthwhile to compare the computational burden of the level building, 2 level DP-warping and sampling DTW algorithms. In Table 1a we show the number of basic time warps, the average size of each time warping region, the product of these two and the amount of temporary storage required by each of these algorithms. In this table, the reduced level building algorithm refers to level building with the addition of the parameters M_T and ϵ, $\bar{N}$ refers to the average reference length, R refers to a range parameter, similar to ϵ, in the 2 level DP-warping algorithm, and $\bar{\gamma}$ refers to the typical branching factor in the sampling DTW algorithm.

Table 1b gives a numerical comparison of the computation for some typical parameter values. It can be seen that the reduced level building algorithm requires only 1/30th of the computation required by the 2-level DP Warping method and that, even the full level building algorithm requires 1/15th computation of the two-level DP warping method.

DTW Algorithm

	Level Building	Two-Level DP Warp	Sampling	Reduced Level Building
Number of Basic Time Warps	$L_{MAX} \cdot V$	$M \cdot V$	$L \cdot V \cdot \bar{\gamma}$	$L \cdot V$
Size of Time Warps	$\bar{N} \cdot M/3$	$\bar{N} \cdot (2R+1)$	$\bar{N} \cdot (2\bar{\epsilon}+1)$	$\bar{N} \cdot (2\epsilon+1)$
Total Computation For Distances	$L_{MAX} \cdot V \cdot \bar{N} \cdot M/3$	$M \cdot V \cdot \bar{N} \cdot (2R+1)$	$L \cdot V \cdot \bar{\gamma} \cdot \bar{N}\, (2\bar{\epsilon}+1)$	$L \cdot V \cdot \bar{N}\, (2\epsilon+1)$
Storage	$3 \cdot M \cdot L_{MAX}$	$2 \cdot M \cdot (2R+1)$	0	$3 \cdot M \cdot L_{MAX}$

Table 1a

Computational Comparisons of Connected Word DTW Algorithm

DTW Algorithm

	Level Building	Two-Level DP Warp	Sampling	Reduced Level Building
Number of Basic Time Warps	50	1200	60	40
Size of Time Warps	1400	875	595	875
Total Computation For Distances	70,000	1,050,000	35,700	35,000
Storage	1800	6000	0	1800

Table 1b

Typical Computational Requirements For The Case $L_{MAX} = 5$, $V=10$, $M=120$, $\bar{N}=35$, $\bar{\epsilon}=8$, $\epsilon=12$, $\gamma=1.5$, $L=4$

IV. Experimental Evaluation of the Level Building Algorithm

In order to evaluate the performance of the level building DTW algorithm we performed a series of connected digit recognition experiments. In these experiments the set of recordings described by Rabiner and Schmidt were used [5]. In this set each of six talkers spoke 80 randomly generated sentences of from 2 to 5 digits per sentence. All recordings were made over dialed-up telephone lines, sampled at a 6.67 kHz sampling rate, preemphasized by a first order network and analyzed by an 8 pole LPC analysis. This analysis was performed with a 300 sample window once every 100 samples for a frame rate of 67 frames per second. Following feature extraction the beginning and ending points of the test patterns were determined. No further segmentation was used. The set of reference patterns consisted of either a single isolated word per digit in a speaker trained mode or twelve isolated word templates per digit formed by a clustering procedure for a speaker independent mode.

The results of the connected digit recognition tests are shown in Table 2 and Figures 7 and 8. For the speaker trained recognizer giving a string recognition rate of 95.2% was obtained. Figures 7 and 8 show the effects of varying the parameters of the level building DTW algorithm around the "operating point" used to give the results of Table 2 for the speaker trained system. Figure 7 shows the results of varying δ_{R_1} and δ_{R_2}. A distinct minimum string error rate is obtained for the parameters $\delta_{R_1} = 4$, $\delta_{R_2} = 6$ with small increases in error rate obtained for nearby values of these parameters. The importance of the δ_{R_1}, δ_{R_2} is clear from this figure since fairly significant values of δ_{R_1} and δ_{R_2} were used to achieve the low error rate.

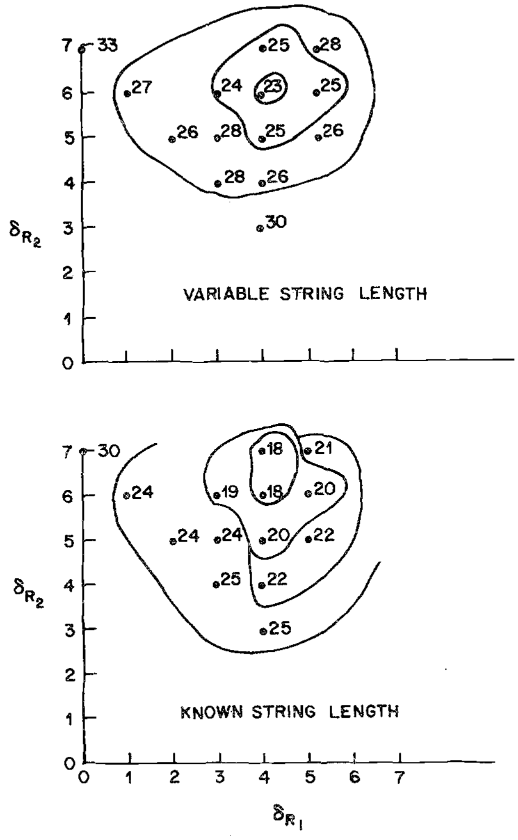

Fig. 7 Contour plot of the number of string errors as a function of δ_{R_1} and δ_{R_2}.

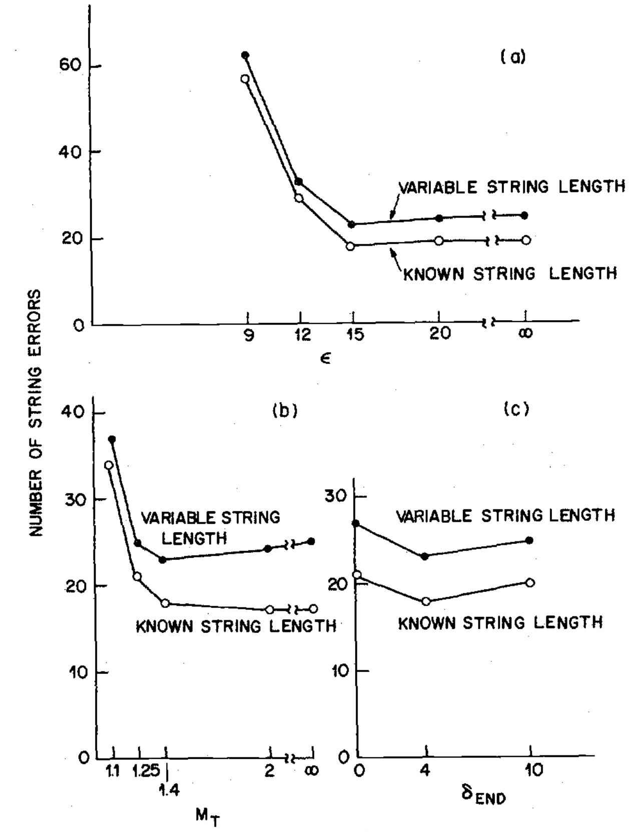

Fig. 8 Plots of the number of string errors vs. ϵ, M_T and δ_{END}.

CONNECTED DIGIT RECOGNITION SCORES

	STRING ERROR RATE	WORD ERROR RATE	STRING ERROR RATE-KNOWN LENGTH
SPEAKER TRAINED	4.8%	0.7%	3.8%
SPEAKER INDEPENDENT	4.6%	0.9%	3.5%

6 TALKERS - 3 MALE, 3 FEMALE
80 STRINGS PER TALKER
20 EACH OF LENGTH 2,3,4,5 DIGITS
BALANCED DIGITS WITHIN STRINGS

Table 2 Connected digit recognition test scores.

Figure 8 shows the effects of varying ϵ, M_T and δ_{END} on the number of string errors. All three parts of this figure show the interesting result that a finite optimum exists for each of these parameters. This is particularly encouraging since finite values for ϵ and M_T reduce the amount of computation required by the level building algorithm.

The results of the recognition tests using speaker independent templates are also given in Table 2. In this case a string recognition rate of 95.4%, slightly better than the speaker trained case, was obtained. These results compare favorably with the results presented by Rabiner and Schmidt in which they reported, for the same set of data, a 6.7% string error rate for the speaker trained case and a 9.0% string error rate for the speaker independent case. Since the level building algorithm is equivalent to the 2-level DP matching algorithm with $\delta_{R_1} = 0$ and $\delta_{R_2} = 0$ it is clear that the increased flexibility of the level building algorithm improves the performance.

V. Summary

We have described a level building approach to connected word recognition using isolated word reference patterns. The method was shown to be relatively efficient and flexible. In addition, we demonstrated that the algorithm is capable of connected digit recognition string accuracies of 95-96% for both speaker trained and speaker independent systems.

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] 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, pp. 588-595, Dec. 1979.

[5] L. R. Rabiner and C. E. Schmidt, "Application of Dynamic Time Warping to Connected Digit Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, pp. 377-388, August 1980.

[6] J. S. Bridle and M. D. Brown, "Connected Word Recognition Using Whole Word Templates," *Proc. Autumn Conf. Institute of Acoustics*, 1979.

[7] L. R. Bahl and F. Jelinek, "Decoding for Channels with Insertions, Deletions, and Substitutions with Applications to Speech Recognition," *IEEE Trans. on Info. Theory*, Vol. IT-21, pp. 404-411, July 1975.

[8] C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, to appear.

[9] C. S. Myers and L. R. Rabiner, "Connected Digit Recognition Using a Level Building DTW Algorithm," submitted for publication.

[10] C. S. Myers and S. E. Levinson, "Connected Word Recognition Based on Syntactical Analysis and a Level Building Dynamic Time Warping Algorithm," Proceedings of ICASSP, 1981.