A MODIFICATION OVER SAKOE AND CHIBA'S DYNAMIC TIME WARPING ALGORITHM FOR ISOLATED WORD RECOGNITION

K. K. Paliwal, A. Agarwal and S. S. Sinha

Speech and Digital Systems Group
Tata Institute of Fundamental Research
Homi Bhabha Road, Bombay 400 005, India

ABSTRACT

A modification over Sakoe and Chiba's dynamic time warping algorithm for isolated word recognition is proposed. It is shown that this modified algorithm works better without any slope constraint. Also, this algorithm not only consumes less computation time but also improves the word recognition accuracy.

INTRODUCTION

The dynamic time warping (DTW) algorithm provides a procedure to align optimally in time the test and reference patterns and to give the average distance associated with the optimal warping path. Sakoe and Chiba (1) proposed a DTW algorithm for spoken word recognition and showed experimentally its superiority over other algorithms reported in the literature. In the present paper, a modification over Sakoe and Chiba's DTW algorithm (1) is proposed. The modified DTW algorithm is applied to a speaker *dependent isolated word recognition system.* It is shown that the modified algorithm works better without the slope constraint. Also, this algorithm consumes less computation time and improves the word recognition accuracy.

THE MODIFIED DTW ALGORITHM

Figure 1 shows diagrammatically the difference between Sakoe and Chiba's DTW algorithm (1) and the modified DTW algorithm proposed in this paper. The test pattern
$A = a_0, a_1, ----, a_i, ----, a_I$
and the reference pattern
$B = b_0, b_1, ----, b_j, ----, b_J$
are developed here across the i-axis and j-axis, respectively. As can be seen from this figure, the present DTW algorithm is similar to Sakoe and Chiba's algorithm except for the form of the adjustment window used for restricting the warping function. Sakoe and Chiba (1) used an adjustment window given by
$|i-j| \leq r$,
where r is a positive integer called the window length. It can be seen from Fig. 1(a) that the adjustment window includes the ending point (I, J) in it only when the window length r is greater than the difference |I-J| in the durations of the test and reference patterns. Thus, in this case, the window length has a lower limit which depends on the actual duration difference |I-J|.

In the modified DTW algorithm, the adjustment window, as shown in Fig. 1(b), always includes the ending point irrespective of its length and is given by
$|i-(j/s)| \leq r$,
where s is the slope of the line joining the beginning point (0, 0) and the ending point (I, J) and is equal to J/I. Thus, in this case, the window length can be less than the actual duration difference |I-J|. We shall experimentally show in the next section that this small change in the form of the adjustment window not only reduces the amount of computation but also improves the recognition accuracy.

RECOGNITION EXPERIMENT AND RESULTS

The aim here is to compare experimentally the performance of the modified DTW algorithm with that proposed by Sakoe and Chiba (1) when applied to a speaker dependent isolated word recognition system. The speech data used for this purpose consist of ten Hindi digits (0 through 9) spoken in isolation by a single male speaker. Forty-nine repetitions of these ten digits are recorded in an ordinary office room. The speech signal is low-pass filtered at 4 kHz and digitized at 10 kHz sampling rate, using a 12-bit analog-to-digital converter. End points of the spoken digits are detected manually. The speech signal is analyzed at the rate of 100 frames per second and three features, namely, the energy of the speech signal, zero crossing rates of the speech signal and its first derivative, are extracted directly from the speech waveform, using a rectangular window of 18 ms duration.

Each spoken digit is represented by a time pattern of feature vectors, each feature vector characterizing a single frame. The test pattern is

CH 1746-7/82/0000 · 1259 $ 00.75 © 1982 IEEE

compared against the reference patterns of all ten digits in the vocabulary and classified as belonging to the digit showing the best match. The symmetric form of DTW algorithm is used to optimally align in time the test and reference patterns and to give average distance associated with the optimal warping path. An L_1 norm between two feature vectors is used for local distance computation. For a more detailed description of the recognition experiment, see reference (2).

The recognition system uses the first ten repetitions of the spoken digits for training and the remaining thirty-nine repetitions for testing. The reference patterns are created for each digit in the vocabulary from the data in the training set by averaging (after dynamic time warping) the patterns of ten repetitions of the same digit. The recognition system is tried out on the data in the test set and performance of the recognition system is evaluated for different values of r, the window length. Sakoe and Chiba's DTW algorithm and the modified DTW algorithm are used with slope constraint conditions of P=0 (i.e., no slope constraint) and P=1. The recognition accuracy as a function of window length is shown in Fig. 2(a) for Sakoe and Chiba's DTW algorithm and in Fig. 2(b) for the modified DTW algorithm. The solid lines correspond here to P=0 and the dashed lines correspond to P=1.

Based on the recognition scores shown in Fig. 2, the following observations can be made:

1) The recognition accuracy increases at first with r, the window length, attains a maximum and decreases thereafter. The value of window length that gives the highest recognition accuracy is 9 frames (i.e., 90 ms) for Sakoe and Chiba's DTW algorithm and 3 frames (i.e., 30 ms) for the modified DTW algorithm. Since the amount of computation time required by the DTW algorithm to find an optimal warping path is proportional to (2r+1), the modified DTW algorithm requires computation time which is less than half of the computation time required by Sakoe and Chiba's DTW algorithm.

2) For Sakoe and Chiba's DTW algorithm, the recognition accuracy obtained with the slope constraint P=1 is more than that obtained without any slope constraint (i.e., P=0), thus indicating the necessity of introducing a slope constraint in this case. Sakoe and Chiba (1) also obtained similar results. These results can be explained as follows. Sakoe and Chiba's DTW algorithm, requiring relatively large window length, provides too much of the warping function flexibility. This leads to poor discrimination between different digits. So for improving the recognition results, it becomes necessary to introduce a slope constraint to limit the warping function flexibility.

3) For the modified DTW algorithm, the recognition accuracy obtained with slope constraint P=1 is less than that obtained without the slope constraint (i.e., P=0). This can be explained as follows. The modified DTW algorithm requires relatively small window length and thus limits itself the warping function flexibility. So, it is not necessary to have an additional slope constraint for limiting the warping function flexibility.

4) The modified DTW algorithm gives the optimal recognition accuracy of 99.0% with P=0 and r=3 frames, while Sakoe and Chiba's DTW algorithm gives the optimal recognition accuracy of 98.5% with P=1 and r=9 frames. Thus, the modified DTW algorithm gives better recognition accuracy than that given by Sakoe and Chiba's DTW algorithm.

The recognition experiment is repeated with reference patterns generated as the miniav centres of the clusters formed from the multiple repetitions of each digit in the training set. (The miniav centre of the cluster is obtained as the pattern whose average distance to all other patterns in the cluster is minimum). The results obtained with these reference patterns are similar to those described above for the average reference patterns. For example, the recognition accuracies with these reference patterns by using Sakoe and Chiba's DTW algorithm are 95.6% with P=0 and r=9 frames and 96.9% with P=1 and r=9 frames. Using the modified DTW algorithm, these recognition accuracies are 98.7% with P=0 and r=3 frames and 97.2% with P=1 and r=3 frames. Thus, we see here also that the modified DTW algorithm gives better results without any slope constraint, while Sakoe and Chiba's DTW algorithm works better with the slope constraint. The window length used by the modified DTW algorithm is less than that used by Sakoe and Chiba's DTW algorithm. Thus, the modified DTW algorithm is computationally more efficient than Sakoe and Chiba's DTW algorithm. Also, the recognition accuracy obtained by the modified DTW algorithm is more than that obtained by Sakoe and Chiba's DTW algorithm.

CONCLUSION

A modification over Sakoe and Chiba's DTW algorithm is proposed. The modified algorithm is applied to a speaker dependent isolated word recognition system. It is shown that the modified DTW algorithm works better without imposing any slope constraint on the warping function. Also, the modified algorithm requires less computation time and gives better recognition accuracy.

REFERENCES

(1) H. Sakoe and S. Chiba, "Dynamic programming optimization for spoken word recognition", IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-26, No.1, Feb. 1978, pp. 43-49.

(2) K.K. Paliwal, A. Agarwal and S.S. Sinha, "Automatic recognition of spoken (Hindi) digits using DP time warping", Indian Journal of Technology, 1982 (in press).

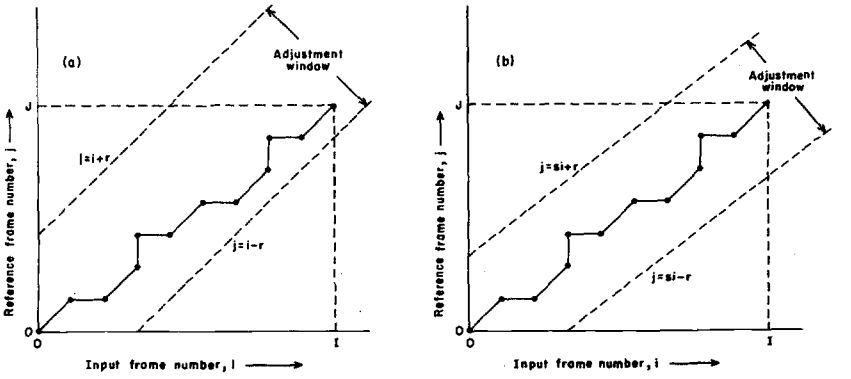

Fig. 1 Illustration of adjustment window for (a) Sakoe and Chiba's DTW algorithm and (b) the modified DTW algorithm.

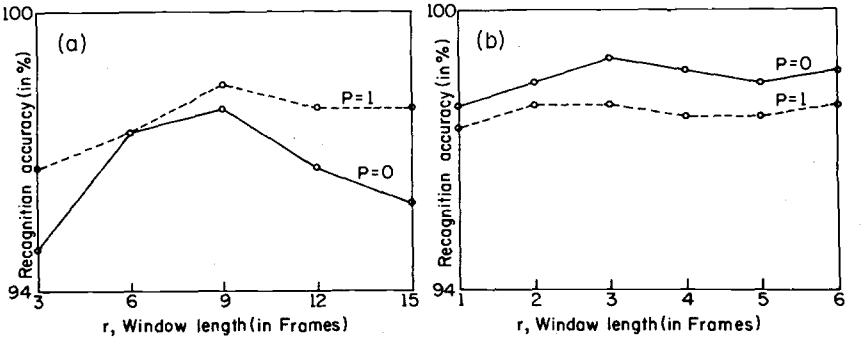

Fig. 2 Recognition accuracy as a function of window length for (a) Sakoe and Chiba's DTW algorithm and (b) the modified DTW algorithm.