Cross-words Reference Template for DTW-based Speech Recognition Systems

Waleed H. Abdulla, David Chow, and Gary Sin
Electrical and Electronic Engineering Department
The University of Auckland, Auckland, New Zealand
Email: w.abdulla@auckland.ac.nz , http://www.ele.auckland.ac.nz/~wabd002

Abstract—One of the main problems in dynamic time-warping (DTW) based speech recognition systems are the preparation of reliable reference templates for the set of words to be recognised. This paper presents a simple novel technique for preparing reliable reference templates to improve the recognition rate score. The developed technique produces templates called crosswords reference templates (CWRTs). It extracts the reference template from a set of examples rather than one example. This technique can be adapted to any DTW-based speech recognition systems to improve its performance. The speaker-dependent recognition rate, as tested on the English digits, is improved from 85.3%, using the traditional technique, to 99%, using the developed technique.

1. INTRODUCTION

Searching for the best path that matches two time-series signals is the main task for many researchers, because of its importance in these applications. DTW is one of the prominent techniques to accomplish this task, especially in speech recognition systems [13]. DTW is a cost minimisation matching technique, in which a test signal is stretched or compressed according to a reference template.

Although there are other advanced techniques in speech recognition such as the hidden Markov modelling (HMM) and artificial neural network (ANN) techniques [14], the DTW is widely used in the small-scale embedded-speech recognition systems such as those embedded in cell phones. The reason for this is owing to the simplicity of the hardware implementation of the DTW engine, which makes it suitable for many mobile devices. Additionally, the training procedure in DTW is very simple and fast, as compared with the HMM and ANN rivals.

The accuracy of the DTW-based speech recognition systems greatly relies on the quality of the prepared reference templates. The normal procedure in selecting the reference templates is to select one example then test its recognition rate. If the recognition rate is high then this reference is kept, otherwise another template has to be selected. A common way to improve the recognition performance is to use several templates for each word. This procedure is computationally inefficient because it increases the number of templates. Vector quantisation (VQ) is another solution to prepare reliable templates for the DTW-based speech recognition systems [10]. However, it requires many training examples to prepare a reliable codebook. The developed technique is very suitable to prepare reference templates for small vocabulary speech recognition systems. It can easily be adapted to any hardware- or software-implemented speech recognition system based on DTW technique. It is reliable and computationally efficient, as it doesn't change the number of reference. It does not require many examples as it is in the VQ case.

The remainder of this paper is organised as follows. Section 2 gives a description of the speech features used. It also introduces the DTW technique. Section 3 describes the developed template preparation technique. Section 4 depicts its performance as compared with the traditional technique. Section 5 derives the main conclusions from this work.

2. FEATURE EXTRACTION AND DTW

The feature vectors used are the mel frequency cepstral coefficients (MFCC) [4, 6]. The temporal structure of the speech signal is represented by the first and the second derivatives of the short time spectra [5, 7]. The distance d between a frame r and a frame t of a speech signal can be represented by:

$$d^2(r,t) = \sum_{n=1}^{p} w_{1n}(c_n^r - c_n^t)^2 + \sum_{n=1}^{m} w_{2n}(Dc_n^r - Dc_n^t)^2 + \sum_{n=1}^{s} w_{3n}(DDc_n^r - DDc_n^t)^2 \quad (1)$$

where c_n^r is the cepstral coefficient n of the frame r, Δ and $\Delta\Delta$ represent the first and the second derivatives of the coefficients. P, m, and s are the orders of each feature stream. w_{1n}, w_{2n} and w_{3n} are the weighting functions which are taken (in our experiments) to be the inverse of the variance of each particular feature [9, 10]. Twenty

0-7803-7651-X/03/$17.00 © 2003 IEEE

eight features are selected (p = 11, m = 10, s = 7) [1]. A necessary condition in the training and recognition stages is that the speech signal has to be extracted accurately from the background before the alignment procedure starts. Many techniques are available to accomplish this condition. The simplest one is based on the zero-crossings rate and the energy level of the speech signal, which is suitable for a clean environment [10]. A more sophisticated technique based on the HMM can be used for speech signal detection in a noisy environment [2, 3].

The idea of the DTW technique is to match a test input represented by a multi-dimensional feature vector $T = [t_1, t_2, ..., t_I]$ with a reference template $R = [r_1, r_2, ..., r_J]$ [11, 12]. Figure (1) depicts graphically the idea of the DTW. The co-ordinate (i,j) is the location of the local distance between frame i to frame j. The aim of dynamic time warping is to find the function w(i) such that it gives the least cumulative difference between the compared signals.

3. TEMPLATE PREPARATION

The traditional way of preparing the reference templates is by judiciously selecting one example for each word (needed to be recognised) and considering it as a reference template for that word. The disadvantage of using a single reference template is that it is not robust to the speech signal variability. That is because it is almost impossible for a person to repetitively speak a word exactly in the same way. The speech signal produced would vary according to many factors. Therefore, if the template created is bad, the user would have to change it until he finds a suitable template for the tested word. To overcome this problem without incurring more computations in the recognition phase, a technique is developed to prepare more robust templates, called crosswords reference templates (CWRTs). Using these templates has greatly improved the recognition accuracy, as it is prepared from multiple examples rather than just one example.

A few examples (4 examples is normally sufficient) for each word have to be prepared beforehand. Then, the average length of the extracted templates is calculated. Next, the template with the length nearest to the average length is chosen to be the best template. This later template is considered as the initial reference template. Then the other templates are time aligned by the DTW process such that their lengths will be equal to the chosen initial template. Finally, the final reference template will be created by averaging the time-aligned templates across each frame.

The physical meaning of the average template is to create a template whose local magnitude spectrum is optimised over time and frequency dimensions.

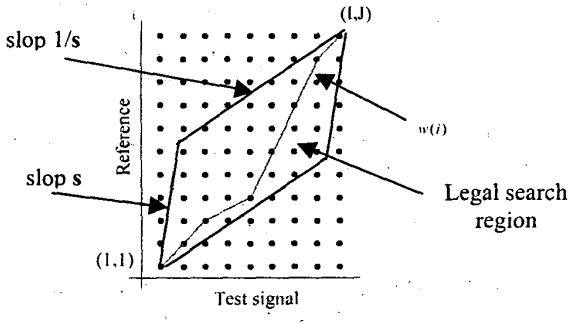

Figure (1) Global path constraint region and the optimum alignment path *w(i)*.

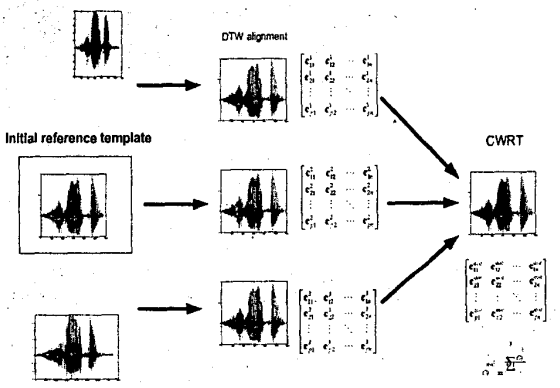

Figure (2) CWRT as extracted from p^{th} dimensional features of three examples.

The compress/expand algorithm to the initially selected reference template can be summarised by the following steps:

1- Align the first template with the initial reference template and find the optimum warping function $w(i)$, $1 \le i \le I$.
2- Trace backward from the last frame to the first one, looking for the slope of the sub-paths for every frame of the speech signal along the alignment path w(i).
3- There are three slop possibilities to deal with:
   a) Slope is 1: Nothing is needed to be changed
   b) Slope is 2: The frame of the speech signal is replicated (expand), i.e., w(i-1) gets the identical frame to w(i).
   c) Slope is 0.5: An average frame is calculated from the two consecutive frames (compress), i.e., w(i) is merged with w(i-1) and gets their average value.

4- Repeat steps 1 and 2 for all the other templates of the available examples to get a set of equal length templates.
5- Average the aligned templates across each frame to get the final reference template.

Figure (2) depicts graphically the CWRT preparation as derived from three examples.

4. RESULTS

Two experiments have been conducted by using the traditional and the developed methods for template preparation. Then, these templates have presented to the same DTW-based speech recognition engine to measure their performances. The first experiment has used the traditional single-template method. Several recognition experiments have been made to select the best template from several training candidates. The second experiment has used the CWRT based method. In this respect, four examples from each word in the dataset have been selected to prepare its corresponding template. The first and second experiments are speaker-dependent experiments, as the training and testing examples are collected from the same speaker.

The dataset used for testing the templates' performance is the 10 English digits, with each digit spoken 40 times by a single speaker. Training examples are excluded from the recognition tests in all recognition tests. Tables I and II show the confusion matrices for the two experiments.

5. CONCLUSIONS

This paper presents a novel and simple training technique to prepare the reference templates (called CWRTs) for DTW-based speech recognition systems. Significant improvements have been attained with this training technique. The reference template prepared by the developed technique is simply derived from a few examples of the words to be recognised rather than selecting just one example as a reference. It differs from the VQ technique in a sense that it requires fewer examples, and it doesn't incur any quantisation errors. In fact the VQ technique cannot work satisfactorily when only few training examples are available. The speaker-dependent recognition rate for the 10 English digits has been improved from 85.3% (using traditional technique) to 99% using the developed technique.

This technique is not exclusive to the ASR applications, but it can easily be extended to any applications that use DWT as part of their functions.

	0	1	2	3	4	5	6	7	8	9
0	33				3	2	2			
1		31								9
2		2	35	2		1				
3	1		1	36					2	
4			1		34	4			1	
5			4	2		34				
6							38	1	1	
7	5				2	2		30		1
8		1			2		2		35	
9		3	1						1	35

Table I – The confusion matrix produced by a single speaker using the traditional single-template method.

	0	1	2	3	4	5	6	7	8	9
0	40									
1		39								1
2			40							
3				40						
4					39				1	
5					1	39				
6							40			
7	1							39		
8									40	
9										40

Table II – The confusion matrix produced by a single speaker using CWRT-based technique.

6. ACKNOWLEDGEMENT

The authors would like to thank Professor Garry Tee for his constructive feedback on this paper. This work is supported by The University of Auckland NSRF grant 3602239/9273.

7. REFERENCES

[1] Abdulla, W. H. (2002a), 'Signal Processing and Acoustic Modelling of Speech Signals for Speech Recognition Systems,' PhD. Thesis, University of Otago, New Zealand.

[2] Abdulla, W. H. (2002b) HMM-based techniques for speech segments extraction, Accepted for publication in Journal of Scientific Programming.

[3] Abdulla, W. H. and N. K. Kasabov (1999b). Two pass hidden Markov model for speech recognition systems. Proc. ICICS'99, Singapore.

[4] Furui, S. (1981) 'Cepstral analysis technique for automatic speaker verification,' IEEE Trans. ASSP-29, 2, pp. 254-272.

[5] Furui, S. (1986) 'Speaker-independent isolated word recognition using dynamic features of spectrum,' IEEE-ASSP-34, no.1, pp 52-59.

[6] Furui, S. (2000) 'Digital Speech Processing, Synthesis, and Recognition,' Marcel Dekker, Inc., New York.

[7] Hanson, B. A. Applebaum, T. H. and Junqua, J. C. (1996) 'Spectral dynamics for speech recognition under adverse conditions,' Automatic Speech and Speaker Recognition Advanced Topics, (Eds) C. H. Lee, F. K.

[8] Soong and K. K. Paliwal, Kluwer Academic Publishers.

[9] Paliwal, K. K. (1992). 'Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer,' Digital Signal Processing 2: 157-173.

[10]Rabiner, L. and Juang, B. H. (1993) 'Fundamentals of Speech Recognition ,' Prentice-Hall, Englewood Cliffs, New Jersey, USA.

[11]Sakoe, H. and Chiba, S. (1971) 'Recognition of continuously spoken words based on time-normalization by dynamic programming,' J. Acoust. Soc. Jap., vol 27, no. 9, pp. 483-500.

[12]Sakoe, H. and Chiba, S. (1978) 'Dynamic programming algorithm optimization for spoken word recognition,' IEEE Trans. ASSP, vol.26, no. 1, 43-49.

[13]Silverman, H. F. and Morgan, D. P. (1990) ' The application of dynamic programming to connected speech recognition,' IEEE ASSP Magazine, pp. 7-25, July.

[14]Trentin, E. (2001) 'Robust combination of neural networks and hidden Markov models for speech recognition,' PhD. Thesis, University of Florence (Italy).