PROSODICALLY GUIDED METHODS FOR NEAREST NEIGHBOUR CLASSIFICATION OF SYLLABLES

Frantz CLERMONT * and Simon J. BUTLER **

* Computer Sciences Laboratory
Research School of Physical Sciences
The Australian National University
Canberra, ACT 2601, Australia

** Speech, Hearing and Language Research Centre
Macquarie University
Sydney, New South Wales 2109, Australia

ABSTRACT : An approach to Nearest Neighbour (NN) classification of syllables in continuous speech is described. Acoustic prosodic segmentation of speech is used to guide the conventional Dynamic Time Warping (DTW) distance measure. The acoustic prosodic analysis robustly determines the nuclei of syllables, and establishes neighbouring intervals which include the syllable boundaries. The limits of these intervals are then used to define the global constraints, which serve to restrict the DTW warping paths within an allowable region. Reliable syllable boundaries are therefore determined implicitly in the matching process. Furthermore, when the proposed method is used in NN-classification of a small database of Australian English diphthongs embedded in continuous speech, the accuracy is comparable to that achieved by current DTW-based systems for isolated word recognition.

INTRODUCTION

One of the most difficult problems in continuous speech recognition is finding the boundaries of speech segments (such as word boundaries). This may be contrasted with isolated word recognition where the word end points can be reliably detected and for which Nearest Neighbour (NN) classification techniques such as Dynamic Time Warping have been successfully applied for many years. If word end points could be reliably established in continuous speech, then, in principle, these well established pattern matching techniques would be immediately applicable to continuous speech.

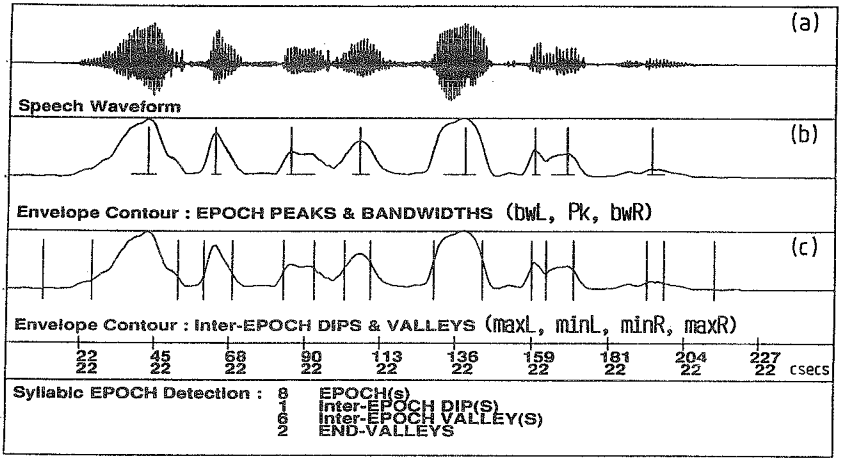
A number of approaches have been suggested to apply DTW when reliable end points are not available. Christiansen and Rushforth (1977) describe a word spotting system which "slides" input speech passed the reference word templates, applying DTW at each time point until a suitable match is found. The computational cost of this approach, however, is substantial, requiring a complete DTW distance calculation for each frame of continuous speech for each reference template. A second approach incorporates the word level DTW (Sakoe, 1979; Myers and Rabiner, 1981) into a global dynamic programming calculation for the entire sentence. In this way, the word boundaries are defined implicitly as part of the process of recognising the entire sentence. The use of syntactic constraints such as these algorithms imply is known to improve recognition performance in some cases. However, these systems are not applicable whenever unrestricted syntax is used or some elements in an utterance cannot be reliably recognised.

One approach which has been proposed to circumvent the segmentation problem is to focus on the "islands of reliability" in speech rather than the segment boundaries themselves (Lea, 1975,1980). These islands of reliability are claimed to be the stressed syllable centres which can be robustly identified using acoustic prosodic analysis (Lea, 1980; Clermont, 1982; Lea and Clermont, 1984). Apparently, however, very few attempts have been made to implement this recognition philosophy.

In this study, a conventional Dynamic Time Warping approach is used to calculate distances. Since this approach to NN-classification is both well understood and can flexibly accomodate various global constraints and boundary conditions, this study endeavours to combine NN-speech recognition algorithms with the acoustic prosodic analysis.

ACOUSTIC PROSODIC ANALYSIS

One of the better developed aspects of acoustic prosodic analysis systems is the process of syllable detection in continuous speech. Syllable detection methods locate areas of high signal energy in the speech waveform which usually represent the vocalic portions of syllables (Lea, 1975, 1980; Mermelstein, 1975; Clermont, 1982; Lea and Clermont, 1984). While the syllable boundaries are difficult to locate precisely, the onsets and offsets of syllabic nuclei can be robustly determined using these techniques. Since syllable boundaries always occur between syllabic nuclei, the offset and onset of adjacent nuclei define intervals which contain the syllable boundaries. It is the limits of these intervals which can be exploited in place of accurate segment boundaries to perform speech recognition.



DMA1:[FRANTZ2.IM9WC1]IM210 FLD1=IM ID=<WC>HIGH 210, SF= 10000 (14-OCT-88)
N= 256, M=14, PR= 98, HAM=Y, SDFR= 250, STFR= 2, NFR= 246, CTX=100

Figure 1: Acoustic Prosodic Processing (Syllable Detection Module)

The processing of the acoustic prosodic system (syllable detection module) is illustrated in Figure 1. The speech waveform of an utterance containing the diphthong /hai/ in the carrier phrase "Now, I see a [...] collection" is shown in Figure 1(a). An envelope detector followed by moving-averaging yields the smooth envelope contours shown in Figures 1(b,c). The envelope contour is then partitioned by a number of thresholding levels determined between a percentage value below the overall maximum and a percentage value above the overall minimum. At any given level, the whole contour is searched for left and right crossings, which signal the existence of peaks above that level. As the envelope is scanned from a lower to a higher level, any peaks previously detected are ignored while new ones are retained. This process yields an array of peak locations which are further inspected for a correct time sequence. Furthermore, a region neighbouring every syllable peak is delineated by its bandwidth, which is a percentage value below the peak. This is the first stage of the syllable detection process. Subsequently, a forward and backward search from every peak towards its successor and predecessor determines whether the separating intervals are sufficiently long to be declared valleys, or else to be declared dips.

From the envelope contour of the signal, seven prosodic markers are therefore determined for each syllable analysed. They include (Figure 1(b)) the syllable centres and bandwidths, and, in Figure 1(c), the onset and offset of the syllabic nucleus (minL,minR), and the offset of the preceding syllable nucleus (maxL) and the onset of the following syllable nucleus (maxR). The acoustic prosodic system generally extracts these markers reliably, although a systematic investigation of difficult cases where the syllables are not well defined has not been undertaken.

Finally, once syllabification is completed, a measure of stress is computed as the "integral" under the curve of every syllable nucleus marked between (minL) and (minR). Lea (1975) has argued that this measure is more desirable in light of previous studies which have shown that duration is a better cue to stress than is intensity. Furthermore, large "integral" values combined with local increases in fundamental frequency are known to be amongst the best acoustic correlates of stress. In this preliminary study, no attempt is made to extract fundamental frequencies, and, therefore, the "integral" measure alone is insufficient for assigning relative stress levels. However, the "integral" measure has proven powerful for automatically detecting target syllables as either the most or second most stressed in a carrier phrase.

RECOGNITION METHODOLOGY AND EXPERIMENTS

Prosodically Guided Dynamic Time Warping

One commonly used variation of the DTW technique for isolated word recognition uses global constraints with relaxed end points (Sakoe and Chiba, 1978). The advantage of this approach is that modest imperfections in end point detection do not adversely affect the recognition performance. The end points are found implicitly in finding the best warping function when matching templates. Given a syllable template with prosodic markers defining a region enclosing the syllable boundaries, it is natural to apply this same implicit process of DTW boundary location in matching syllables. However, whereas Sakoe and Chiba's approach allows a fixed relaxation of end point constraints, the acoustic prosodic analysis provides bounds for relaxation, which are sensibly determined from the syllable itself and therefore adapt to the requirements of the data. The global constraints are therefore a variable function of the prosodic markers (maxL,minL,minR,maxR). Figure 2 schematises the special case under which the end points of the reference syllable are known and those of the test are not. Figure 3 reproduces a real example of this special case, where the end points of the TEST template are shown to be implicitly defined in the DTW-matching process.

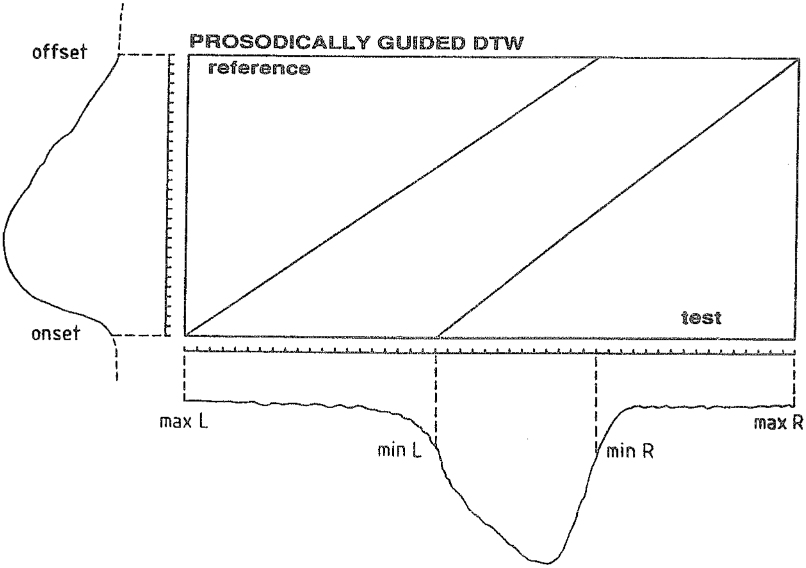


Figure 2: Prosodically Guided DTW-Global Constraints : REFERENCE in isolation versus TEST extracted from continuous speech

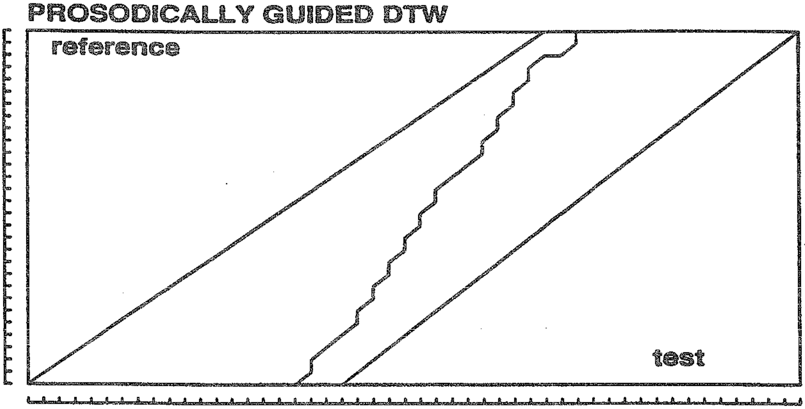


Figure 3: Prosodically Guided DTW-Global Constraints : Real Example of Special Case shown in Figure 2

Experiments

The prosodically guided recognition method can generally be used with both reference and test syllables that are extracted from continuous speech using the acoustic prosodic analysis. Training (i.e., acquiring reference data) from continuous speech, however, obviously requires some supervision to select the appropriate reference syllables. A straightforward approach to training is to use isolated syllables for training data so that conventional end point detection can precisely locate their boundaries. An investigation was therefore undertaken to determine whether monosyllables, trained in isolation, can be accurately recognised in continuous speech. In other words, are syllables in continuous speech sufficiently similar to their isolated counterparts to permit matching within the framework of DTW ?

An Australian English diphthong database was used and consists of both diphthongs in /hVV/ context and semivowels in /CV/ context totalling 16 syllables. These monosyllables, presented five times in a random order, were spoken both in isolation and in a carrier phrase by one adult male speaker. A software system was developed for NN-speech recognition using DTW with prosodically guided global constraints. This system, called RNN (for Robust Nearest Neighbour), performs K-nearest neighbour recognition and allows a number of different LPC based features and distance measures. A specification of the training and test data sets is generated externally to the system providing maximum flexibility in designing recognition experiments. Furthermore, several alternative DTW global constraints may be used which are currently under experimental investigation.

For the experiment reported here, one (K=1) nearest neighbour is used with one reference template per class. Figure 4 shows the recognition accuracy achieved for a family of Linear Prediction (Order 8) parameters and distance measures. Highest performance is obtained with the index-weighted cepstral distance (NDPS) and ITAKURA's ratio of residual energies, while an expected degradation in performance is recorded for the euclidean distance based on autocorrelation and reflection coefficients (AC_EUC and RC_EUC).

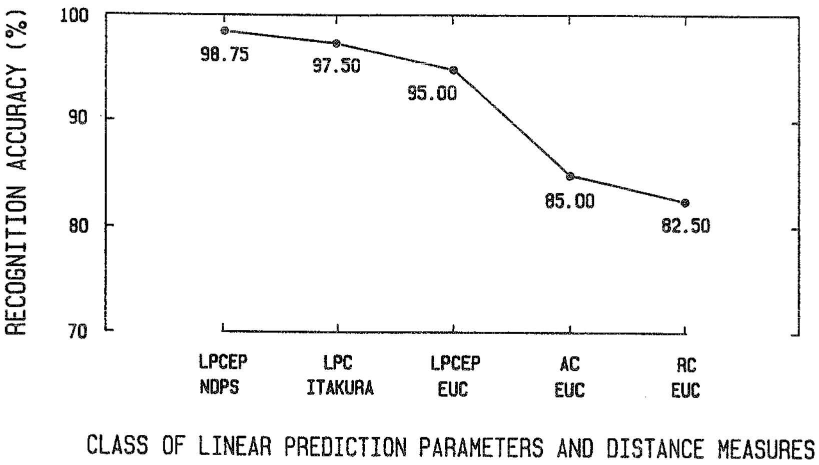


Figure 4: NN-Classification Performance of Prosodically Guided DTW Distances

These results are comparable to those achieved by current DTW-based systems for isolated word recognition. The prosodically guided method seems robust and sufficiently flexible to adapt to recognition of syllables in continuous speech. While this does not represent a difficult vocabulary from an isolated word recognition point of view, there is a certain degree of complexity in classifying, in continuous speech, diphthongs in syllable final context. Because of coarticulation with the following syllable, diphthongs may not reach their second vowel targets, thus increasing the confusability of the test set. Although vocabulary selection is, of course, crucial to recognition success even in isolated word DTW, the purpose of this study was not to demonstrate that any vocabulary when trained in isolation can be successfully recognised in continuous speech. Rather, the aim was to illustrate the validity of the proposed method when applied to the carefully designed vocabularies available for this experiment.

DISCUSSION

The prosodically guided DTW is a general method for matching syllables extracted from continuous speech. Isolated syllables are just one special case which has proven useful as training data in the described experiments. These experiments have also been confined to monosyllabic words for simplicity. The higher level strategies for integrating syllable recognition into a more general speech recognition system have not been discussed. The approach described here, however, is very flexible as it can support anything from word spotting (Christiansen and Rushforth, 1977) to the use of rigid syntactic constraints (Myers and Rabiner, 1981). In this sense, prosodically guided dynamic time warping is proposed as the nucleus of a more general system, around which may be devised a variety of higher level strategies.

It may be noted, however, that the recognition performance of this approach can only be as good as the underlying acoustic prosodic and DTW components will allow. Whilst acoustic prosodic analysis is relatively reliable, difficult cases undoubtedly exist which may not be consistent with the definitions of prosodic markers that have been used. Furthermore, experience with isolated word DTW systems indicates that vocabulary selection is a critical factor in performance. As with current recognition systems, the proposed approach can achieve similar performance under carefully controlled conditions with the added capability of recognition in continuous speech.

REFERENCES

Christiansen, R.W. and Rushforth, C.K. (1977) "Detecting and locating key words in continuous speech using linear predictive coding", IEEE Trans. Acoust. Speech and Sig. Proc., 25, 361-367.

Clermont, F. (1982) "Syllabic epoch detection by multi-level search of the envelope contour", Unpublished research report.

Lea, W.A., Medress, M.F., Skinner, T.E. (1975) "A prosodically guided speech understanding strategy", IEEE Trans. Acoust. Speech, and Sig. Proc., 23, 30-38.

Lea, W.A. (1980) "Prosodic aids to speech recognition", *Trends in Speech Recognition* (W.A. Lea, Ed.), Englewood Cliffs, NJ:Prentice-Hall, Ch. 8.

Lea, W.A. and Clermont, F. (1984) "Algorithms for acoustic prosodic analysis", IEEE Int. Conf. Acoust. Speech and Sig. Proc., Conf. Rec., 2471-2474.

Mermelstein, P. (1975) "Automatic segmentation of speech into syllabic units",J. Acoust. Soc. Am., 58, 880-883.

Myers, C.S. and Rabiner, L.R. (1981) "A level building dynamic time warping algorithm for connected word recognition", IEEE Trans. Acoust., Speech and Sig. Proc., 29, 284-297.

Sakoe, H. (1979) "Two-level DP-matching: a dynamic programming based pattern matching algorithm for connected word recognition", IEEE Trans. Acoust. Speech and Sig. Proc., 27, 588-595.

Sakoe, H. and Chiba, S. (1978) "Dynamic programming algorithm optimization for spoken word recognition", IEEE Trans. Acoust. Speech and Sig. Proc., 26, 43-49.