TIME DOMAIN ACOUSTIC PARAMETERS FOR SPEECH RECOGNITION

W. Millar *

ABSTRACT

This paper examines some of the reasons for performing speech recognition in the time domain. A short review of typical systems using this technique for isolated word recognition is given. Parameters which may be easily extracted from the speech waveform are discussed and a set of typical measurements of these parameters is presented. Some conclusions about the usefulness of these parameters are made and suggestions for further examination are given.

INTRODUCTION

The majority of current, commercially available speech recognisers transform the incoming speech time waveform to another domain (usually frequency) before any pattern matching is performed. The transformed signal is represented by a vector of coefficient values calculated at regular intervals. The pattern recognition algorithm in the recogniser must then manipulate this matrix of values in such a way as to take account of the inherent variability of human speech and, in some way, interpret the pattern as an utterance to be recognised. As such, the recogniser is not a "speech" recogniser in that it cannot distinguish speech from a non-speech signal. It has been appreciated that true speech recognisers will be more than pattern classifiers and will require information and knowledge from various sources in addition to a transform domain representation of the speech signal. The nature of this information and its embodiment is still, however, the subject of debate and research.

In this paper we consider information which is available in the time domain signal. We do this for a number of reasons. Some information may be useful in augmenting the transform domain feature vectors in current speech recognisers. In addition, some information may only be available in the time domain and may be lost in the smearing effect of any transformation. We begin in the next section by looking at some previously reported analysis techniques. In the following section some results of applying these techniques to speech are presented. Finally some conclusions and suggestions for further work are made.

PREVIOUSLY REPORTED WORK

In this section we examine some typical speech recognition systems employing time domain analysis which have appeared in the literature.

The first truly successful automatic speech recogniser was that produced at Bell Laboratories (ref 1). The incoming signal was both high- and low-pass filtered at 900Hz to produce two signals. A zero crossing analysis was then performed on each channel. These measurements were then compared to previously stored values and the most likely utterance selected. Accuracies of 98% for speaker dependent tests on the digits vocabulary.

[* British Telecom Res. Lab., Martlesham Heath, Ipswich.]
were reported. The major problem with the system was the difficulty in retraining for different speakers.

Extensions of this type of approach have been reported in which the number of filters was increased (ref 2) or alternatively, in which additional parameters were used; for example the zero crossing rate of the first derivative of the speech signal (ref 3). Such systems were (at the time) hampered by the lack of processing power available beyond the front end. An interesting study which attempted to overcome this and other problems, compressed the incoming data stream by segmenting it into regions corresponding to different sound classes and then allocating symbols to each region (ref 4). Each utterance was represented by a string of symbols which could then be compared to previously stored symbol strings for recognition. This system produced 94% accuracy for the 30 talkers in the training set on the digits vocabulary and 91% accuracy for 12 talkers outside the training set.

Analysis techniques employing parameters other than energy/zero crossing measurements have been used. For example a measurement of waveform asymmetry has been suggested (ref 5), primarily as a technique for distinguishing between voiced and unvoiced sounds. When applied to isolated word recognition on a vocabulary of 15 words (digits and control words), an accuracy of 96.8% in 29400 words from 14 speakers was obtained.

By a relatively simple extension of the analysis, the time waveform can be analysed in the autocorrelation domain, as is the case in Linear Prediction (LP) analysis. The basis of the LP technique is that the current signal value can be calculated from a weighted linear summation of the previous n sample values, where n is the order of the prediction. This produces an all-pole model of the signal spectrum and can be used to produce a model of any desired accuracy by adjusting the prediction order. If the analysis is restricted to second order, then a gross spectral characterisation can be obtained as suggested in (ref 6). This form of analysis is relatively simple computationally and produces as a result, two poles which may be either both real or form a complex conjugate pair in the z-plane. The position of the poles can then be used to provide information on the spectral shape. As a byproduct of the analysis, the normalised prediction error may be calculated, which can be regarded as giving a 'goodness of fit' of the model to the waveform. A system incorporating the two pole LPC analysis, together with zero crossing and energy measurements has been described (ref 7). This system produced an average error rate of 2.7% on a speaker independent test of 10 recordings from 10 speakers on the digits vocabulary.

RESULTS

The speech used in the analysis was obtained from a master recording of utterances spoken in a quiet acoustic environment, band pass filtered to between 110 and 9500 Hz, sampled at 20 kHz and linearly quantised to 16 bits accuracy. This speech was then low-pass filtered to 3.4 kHz, down sampled and A-law compressed. Fig 1 shows typical waveforms obtained from one utterance of the word /five/ (fricative/diphthong/voiced fricative). The analysis is performed on 10 ms frames. The parameters have been chosen to be representative of those described and where desirable have been modified to make evaluation easier by the removal of multiplications from the calculations. The energy parameter is
calculated as the sum of the unsigned sample values in each frame. The zero crossing rate is the total number of zero crossings (of both polarity) in each frame. The maximum peak to peak signal is obtained by first finding the peak to peak signal between each peak positive signal and each successive peak negative signal and then finding the maximum of these values. The peak to peak symmetry is obtained, in a similar way, as the arithmetic sum of the peak positive signals and the peak negative signals in each 10 ms frame. The total symmetry is the arithmetic sum of the signal values over each frame. The LP parameters are calculated from a second order LP analysis as described previously (no windowing has been used). All plots are shown with the time axis horizontal.

Although only one utterance is shown, the results are typical. It can be seen that the maximum peak to peak signal is closely related to the energy contour. Some spurious values can occur and these could be removed, if required, by taking an average over several values. The main reason for using the maximum peak to peak signal is that it removes the need for calculating a summation of terms and any inherent overhead in that procedure. The maximum value of this parameter is known a priori and the actual value can be easily calculated as the signal is acquired, without the need for any normalisation. The LPC prediction error and the zero crossing rate are closely related. This might be expected since those regions which have a high zero crossing rate correspond to the fricative regions of the utterances and it is in these regions where the LP model is most likely to produce a high error. The LP frequency is not as conclusive as other reports suggest. The initial fricative segment corresponding to the /f/ sound would be expected to have a much higher frequency spectral pole than that shown. The central region corresponding to the diphthong shows a spectral shape representative of two vowels with the spectrum corresponding to the vowel /a/ changing to that of the vowel /i/. Since the vowel /i/ is dominated by a large amplitude first formant, this will be reflected in the spectral shape. The voiced fricative region is similar to that of the unvoiced fricative and there is little information available to distinguish the two sounds. The two symmetry measures show the difficulty in using this measure. The results are inconsistent throughout the utterance and exhibit significant variations within regions. This in turn makes it difficult to set thresholds and place a value of significance on the measurements.

We can summarise these observations by saying that we find the symmetry parameters to be relatively ineffective, the LP error signal is similar to the zero crossing rate signal in most cases and the maximum peak to peak signal provides the same information as the energy measurement.

**CONCLUSIONS**

We have looked at some techniques which have been employed for speech recognition in the time domain. A measurement of the energy in the waveform and the zero crossing rate have been the two most commonly used parameters. Of the other parameters which have been suggested, the two pole LP frequency parameter is likely to contribute most information.

This study has only considered analysis techniques which segment the utterance into fixed length time segments. We have not as yet examined the duration of segments which are important in speech perception. Further work is required in this area.
References


Acknowledgement is made to the Director of Research, British Telecommunications plc, for permission to publish this paper.

Fig 1. Typical analysis results.