Segmentation of Acoustic Events Using Time Encoded Speech (TES) Descriptors

T A Moore, J Holbeche and R A King

Abstract

A preliminary investigation into the suitability of Time Encoded Speech (TES) descriptors for the segmentation of speech into consistent acoustic events is reported upon. Segmentation is accomplished by a system of parallel seekers, each one optimised to detect a particular class of acoustic event. By this method A-Matrices, that is to say second order TES symbol distribution descriptors, may be compiled for each segment and utterances compared using these as a basis. Word comparisons are achieved by a simple dynamic programming algorithm applied to the small number of segments produced.

Introduction

Time Encoded Speech (TES): TES is a form of speech waveform coding. The speech waveform is broken into time intervals (epochs) between successive real zeros. For each epoch of the waveform the code consists of a single digital word. This word is derived from two parameters of the epoch, its quantised time duration and its shape. The measure of duration is straightforward, and the usually adopted strategy for shape description is to classify epochs on the basis of the number of positive minima or negative maxima occurring therein. For economical coding the number of naturally occurring distinguishable symbols produced by this process may then be mapped in non-linear fashion onto a much smaller number (alphabet) of code descriptors. An algorithm to perform an initial TES coding is described in Ref 1.

TES Isolated Word Recognition (IWR) Systems: IWR systems based upon TES have been reported upon, Ref 2. These show promise in high noise environments such as tactical military applications. The real-time TES based recogniser currently under investigation at RMCS forms whole word archetypes (templates) from A-Matrices (Fig 1). The so-called A-Matrix representation of an utterance or acoustic event is a two dimensional pattern recording the second order TES symbol distribution. Higher dimensional models are also candidates for IWR systems. The data obtained from the entire word are merged into this single matrix. Word comparisons are then attempted between the input word and archetypes created during training. This procedure has many advantages, but is limited in its ability to cope with connected or continuous recognition tasks.

TES Connected Word Recognition (CWR) Systems: An examination of the options available for implementing CWR TES systems indicates that the TES symbol stream is capable of division into regions or segments wherein the code distribution remains reasonably `static'. Such a technique will provide a small number of well-populated “A” or higher order matrices in chronological order (Fig 3), which can be used as the basis for CWR recognition.

Fig 2 shows the results of TES coding a short piece of speech and separating this into 10ms frames. In this description each frame is represented not as an A Matrix, but as a simple first order frequency distribution across the 29 possible code symbols. Symbol 29 is a special symbol reserved for epochs of lower amplitude than a preset threshold, intended to represent silence. It can immediately be seen that in this form the speech divides fairly naturally into segments of broadly similar TES code distribution, some of which are as long as 300ms, from which A-Matrices may be formed.

This paper describes a first attempt to utilise this effect, to segment utterances on a simple, very broad, basis. Word comparison is then achieved using a DTW approach with “A” or higher order matrices formed from these segments. Segmentation procedures have evolved, from a single generalised multiclass classification routine, to a parallel detector (or seeker) configuration, with each individual seeker optimised to detect voiced or unvoiced acoustic events, or silence.

The Segmentation Procedure

The Seeker Concept: The concept of using independant seekers (Fig 4) was proposed to allow flexibility in the choice of processing techniques before reaching decisions on the various classes of acoustic event. For example, the signal may be differentiated before coding for the “unvoiced” events so as to emphasise

---

1 Military Communications Research Group, School of Electrical Engineering and Science, Royal Military College of Science, Shrivenham, Swindon, Wilts SN6 8LA
any high frequencies present, or integrated before coding for the "voiced" seeker to amplify any low frequency components. A range of numerical filtering options post-coding are under investigation as a means of improving the ability of the seekers to classify the incoming signal consistently.

The input to each seeker classification algorithm is a string of TES symbols which is first segmented into fixed length time frames (10-20ms) (e.g. Fig 2). The symbol distribution of each frame is examined to provide a simple dichotomised decision — seeker attributes "present" or "absent".

The recent availability of a TES coder device implemented in LSI CMOS has made the parallel processing option realistic in practise.

The Segmentation Decision: The output from each seeker is thus in a binary format, the result for each frame being true if the frame is considered to be of the type sought. The segmentation algorithm must then take these inputs and decide where to segment the speech in order to form A Matrices. The current algorithm is simple, and works by searching for regions of consistent contiguous descriptors and placing segmentation boundaries between them. This approach seems to have reasonable success if certain restrictions are placed on the minimum length of the segments. An example of segmentation for the words "zero" and "nine" is shown in Fig 5.

Formation of A-Matrices: Simultaneously with the segmentation process a separate coding path, optimised in some way for the final comparison of A-Matrices, is examined. An A-Matrix is compiled for the current segment and, as soon as a segmentation boundary is decided upon, the matrix is closed and data entered into the next. The final result is a series of A-Matrices representing variable time slices throughout the utterance. The duration of each segment is stored, together with its A-Matrix descriptor.

Word Comparison

Having compiled the A-Matrices for each utterance, comparison is carried out between the unknown utterance and the archetypes for the known words. At present these archetypes are formed from single utterances of each word, but matching routines currently under consideration will allow more than one word to be merged.

Our initial comparison routine is basically a simple DTW approach, using a distance score defined as:-

\[ D(A, B) = 100 \times \left\{ 1 - \frac{\sum \sum a_{i,j}b_{i,j}}{\sum \sum a_{i,j}^2 \sum \sum b_{i,j}^2} \right\} \]

where \( A = \{ a_{i,j} \}, B = \{ b_{i,j} \} \) are the A-Matrices to be compared.

The variable length of the segments introduces an unusual complication, which is currently addressed by weighting the cumulative minimum distances calculated across the matching table by the durations of the segments under consideration. Fig 6 shows some examples of comparison matrices.

Results

Table 1 shows some initial results, comparing single utterances by the segmentation routine and the whole word routine. The utterances were spoken by the same male speaker in a computer room environment. The results show that the segmentation algorithm is capable of discriminating between similar words which are confused by a whole word approach. The utterance of "nine" was confused with "niner" by the whole word routine, but correctly classified by the segmentation routine. Similarly the utterance of "fix" shows superior discrimination against "six" and "seven" when compared by the segmentation algorithm.

Results such as these indicate recognition scores with the first immature segmentation and comparison algorithms are of the same order of magnitude as with whole word matching. The distribution of errors is however different.

Conclusions

We are optimistic that the approach described above will allow a CWI TES recogniser to be constructed, and allow improvements in the performance of existing TES IWR algorithms.
There is considerable scope for improvement in both the segmentation and the comparison routines. For example the path constraints on the algorithms have been discovered to be too severe, and this is to be corrected.

These initial results are encouraging, but considerable further work will be required to develop and characterise the procedures.

References


<table>
<thead>
<tr>
<th>'0'</th>
<th>'1'</th>
<th>'2'</th>
<th>'3'</th>
<th>'4'</th>
<th>'5'</th>
<th>'6'</th>
<th>'7'</th>
<th>'8'</th>
<th>'9'</th>
<th>'fix'</th>
</tr>
</thead>
<tbody>
<tr>
<td>09</td>
<td>98</td>
<td>96</td>
<td>85</td>
<td>85</td>
<td>73</td>
<td>79</td>
<td>79</td>
<td>93</td>
<td>46</td>
<td>72</td>
</tr>
<tr>
<td>fix</td>
<td>82</td>
<td>100</td>
<td>90</td>
<td>81</td>
<td>89</td>
<td>72</td>
<td>49</td>
<td>92</td>
<td>49</td>
<td>49</td>
</tr>
</tbody>
</table>

Table 1: Overall Distance Scores For Two Utterance Sets Using The Segmentation And Whole Word Algorithms

\[ \text{Fig 1: Example A-Matrices For The Words "One" and "Three"} \]

\[ \text{Fig 2: Short Term TES Code Distribution For The Word "Six"} \]
**Fig 3:** Whole Word A-Matrix vs Segmentation

**Fig 4:** Segmentation System Utilising Parallel "Seekers"

"Zero"

---

UNVOICED: ..........................

VOICED: ..........................

SILENCE: ..........................

AMPLITUDE: ..........................

SEGMENTS: ..........................

"Nine"

---

UNVOICED: ..........................

VOICED: ..........................

SILENCE: ..........................

AMPLITUDE: ..........................

SEGMENTS: ..........................

**Fig 5:** Segmentation Examples For The Words "Zero" and "Nine"

<table>
<thead>
<tr>
<th></th>
<th>&quot;Nine&quot;</th>
<th></th>
<th>&quot;Nine&quot;</th>
<th></th>
<th>&quot;Zero&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Nine&quot;</td>
<td>94</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>&quot;Nine&quot;</td>
<td>100</td>
<td>32</td>
<td>100</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>&quot;Nine&quot;</td>
<td>98</td>
<td>100</td>
<td>47</td>
<td>98</td>
<td>100</td>
</tr>
</tbody>
</table>

OVERALL SCORE = 46

**Fig 6:** Examples Of Dynamic Programming Distance Scores For Various Utterances

<table>
<thead>
<tr>
<th></th>
<th>&quot;Nine&quot;</th>
<th></th>
<th>&quot;Nine&quot;</th>
<th></th>
<th>&quot;Zero&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Nine&quot;</td>
<td></td>
<td>&quot;Zero&quot;</td>
<td>100</td>
<td>99</td>
<td>90</td>
</tr>
<tr>
<td>&quot;Zero&quot;</td>
<td>100</td>
<td>42</td>
<td>100</td>
<td>100</td>
<td>41</td>
</tr>
</tbody>
</table>

OVERALL SCORE = 96

OVERALL SCORE = 36