AUTOMATIC SEGMENTATION: DATA-DRIVEN UNITS OF SPEECH

S. W. Beet and L. Baghai-Ravary

Aculab plc
Lakeside, Bramley Road, Mount Farm,
Milton Keynes, MK1 1PT, UK.

Tel. +44 1908 273961; Fax. +44 1908 273801
Steve.Beet@aculab.com; Ladan.Ravary@aculab.com

ABSTRACT

An algorithm is presented which allows non-parametric representations of speech to be automatically segmented into units of comparable duration and character to manually-defined phonemes. The consistency of this segmentation across speakers, and across telephone channels, is investigated and the implications of adopting such forms of data for automatic speech recognition are discussed.

1. INTRODUCTION

Throughout the history of speech recognition, there has been a long-standing argument between advocates of the rule-based expert system approach, based on expert phonetic knowledge, and those who would state that 'data is all', relying on blind statistics to provide discrimination between sounds. For many years now, the 'data-driven' approach has proved demonstrably superior, but it is the authors' contention that the knowledge gleaned by generations of phoneticians is itself 'data-driven' and capable of yielding significant insights into the shortcomings of current speech recognition systems.

The symbolic representation of speech devised by phoneticians has been defined by a combination of acoustic and visual perception. The speech units they have devised have been based on patterns observed in real acoustic waveforms and spectrograms. As such they include many patterns which are non-stationary. These periods of changing acoustic characteristics are an inherent part of natural speech, but are not well accounted for in the so-called 'data-driven' formalisms such as hidden Markov models (HMMs). To improve on these techniques, it is necessary to move away from the concept of a finite-state machine (FSM) towards a system which allows for both continuous and abrupt changes.

2. DYNAMIC SPEECH UNITS

One phonetic concept which could contribute to the demise of the FSM approach to speech recognition, is that of articulatory targets. By considering each segment of the speech signal as a transition between two targets, many of the effects of coarticulation become irrelevant. The only problem which remains is the identification of the instants at which each new target comes into view.

Multi-step Adaptive Flux Interpolation (MAFI) [1] provides an algorithm for describing extended segments of the speech signal in terms of an initial parameter vector, a target and a duration. Thus recognition based on these parameters could avoid many of the problems of current systems, provided the values they take are consistent whilst simultaneously including sufficient discriminatory information. The blocks defined during the MAFI analysis are, however, purely data-driven and the units of speech which they describe are not constrained to be directly related to specific phonemes. Thus they succeed in reducing the overall dimensionality of the recognition task, but stop short of implementing hard quantisation.

2.1. Multi-step Adaptive Flux Interpolation

MAFI was developed from the interpolation technique, Adaptive Flux Interpolation (AFI) [2], to model the temporal changes in power spectra of speech and remove redundancy by omitting those frames which can be accurately reconstructed by interpolation between the retained frames. It is essentially a variable frame-rate (VFR) system [3], but the reconstruction of the missing frames is performed using an interpolation which allows for the migration of features from one element of a vector to another within each encoded block. Thus much longer blocks can be encoded than would be possible with a more conventional VFR system.

MAFI seeks to make a flux linkage between every observation vector from N_1 to N_2 explicit, given only the data in the two terminal vectors (Figure 1). The likelihood of each pair of elements being linked, can be assumed to be given by a zero-mean Gaussian distribution of the change in data value from one end of the link to the other. The most likely set of non-crossing links are then found by a simple form of dynamic programming.

Consider the link from element *i* of frame *N-1* to element *j* of frame *N*, as shown here (Figure 1). To find the links between frame *N*-1 and frame *N*, we require a local distance matrix, Γ, containing the likelihood of every potential link. The elements of Γ can be expressed in a normalised log-likelihood form:

$$\gamma_{i,j} = \left| R_{N_1,u} - R_{N_2,v} \right|^2$$

where $R_{N_1,u}$ is the value in element *u* of the known frame N_1 at the point of incidence of the link, linearly extrapolated to reach that frame. $R_{N_2,v}$ corresponds to the other (extrapolated) end of the link, where it intercepts frame N_2.

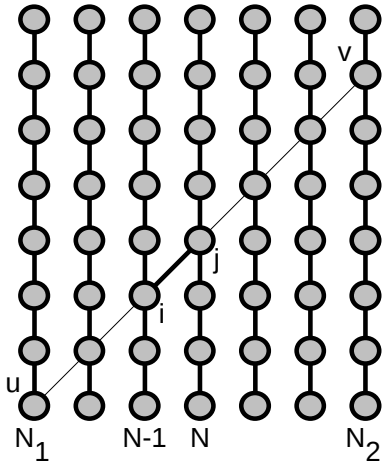

Figure 1: Relationship between linked elements of observation vectors *N* and *N-1*, extended to reach terminal vectors, N_1 and N_2.

Dynamic programming is then used to find the set of links which give the minimum total log likelihood over all selected links. Various constraints can be imposed during the dynamic programming. In particular, links which would be extrapolated to pass beyond the ends of any frames, are disallowed, as are links which correspond to unrealistically rapid frequency transitions.

To segment the data, N_1 is initialised to index the first vector and N_2 is set to N_1+2. N_2 is then repeatedly incremented until the discrepancy between the true vector sequence, and the values found by linear interpolation along the links described above, exceeds some threshold. Once this occurs, the segment is taken to extend from N_1 to N_2-1, N_1 is set to the end of that segment, and the whole process repeated.

2.2. Utilisation

Being able to divide a speech utterance into segments is one thing; gaining any advantage from that process is quite another. For the segments to be useful in speech recognition, their statistical behaviour must be considered. Ideally, the parameters describing the segments should be consistent from one utterance of a word to another. Obviously, this issue is counterbalanced by the need for at least some of the segments to be clearly distinct when different words are used.

The segments identified by MAFI are characterised solely by their duration and their terminal observation vectors. These vectors can take the form of mel-frequency cepstral coefficients (MFCCs), power spectral densities (PSDs), or even auditory representations. Thus the issues of variability are essentially the same as those in most current speech recognition systems. However, for any recognition system to make full use of such a representation, account must be taken of the variation due to the absence of individual boundaries in some cases, and the presence of extraneous ones on other occasions.

3. EXPERIMENTS & RESULTS

This paper uses data collected with a range of channel characteristics (the NTIMIT database [4]) to demonstrate the consistency of MAFI parameters for analysis and representation of speech for recognition. The natural boundaries of the acoustic units identified by MAFI have been compared with the conventional phonetic labels provided with the database. The phonetic segmentation of the example data presented here, is shown in Figure 2.

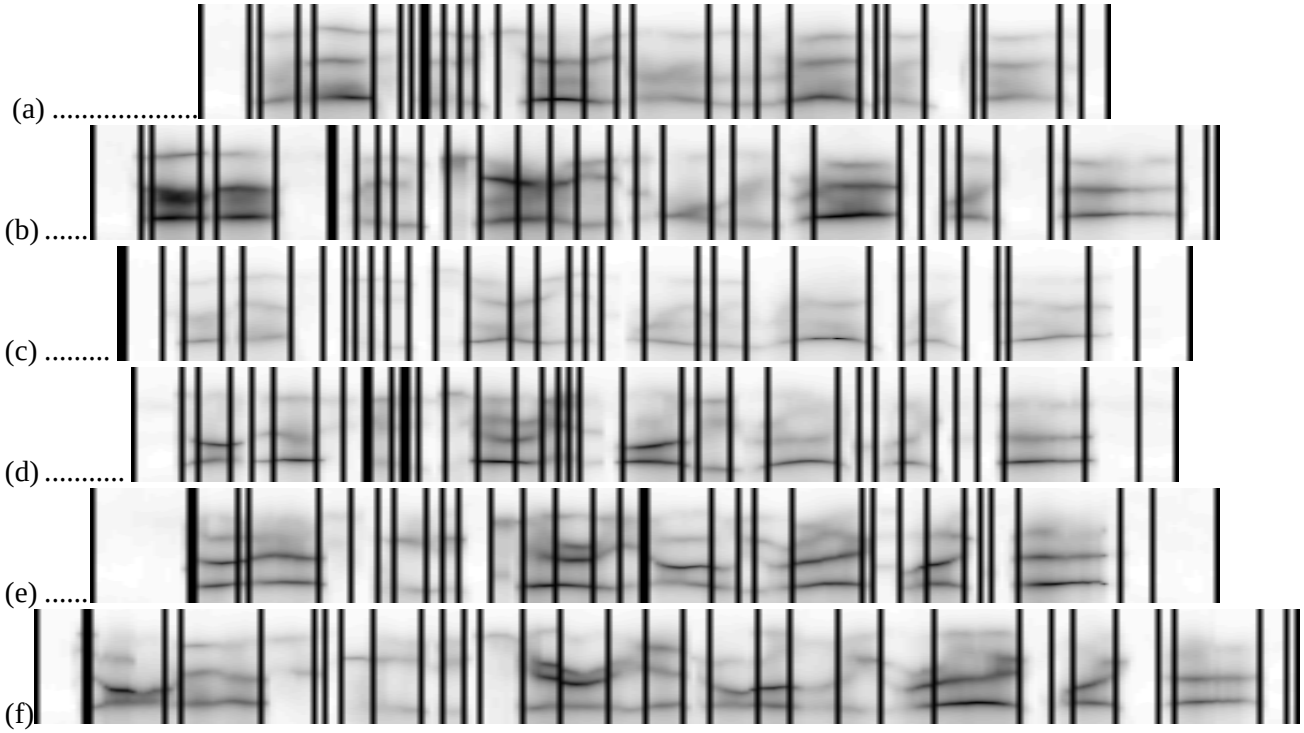

Figure 2: Phonetic segmentation of six sample utterances of the same sentence by different speakers: (a) - (c) female, (d) - (f) male

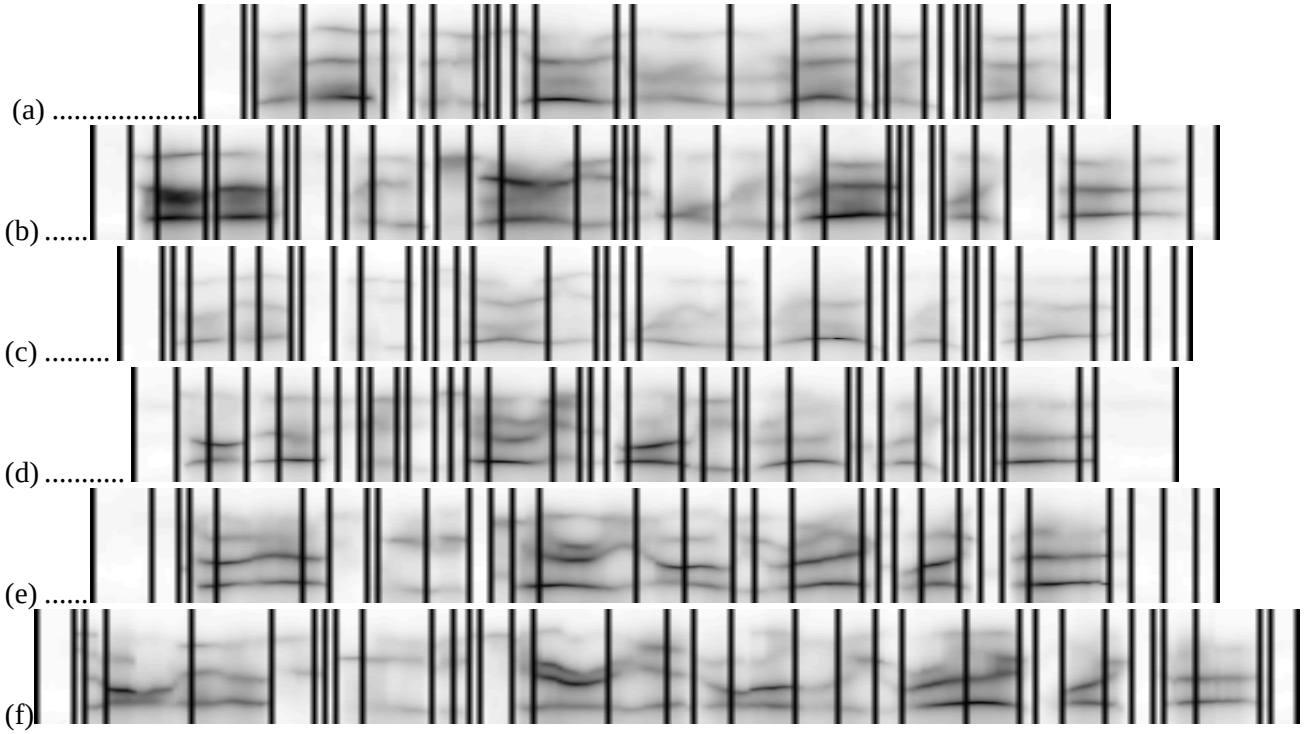

Figure 3: Data-derived segment boundaries for the same utterances shown in Figure 2.

The results of the MAFI segmentation are impractical to fully describe with simple statistics, so some graphical examples, both of data which shows consistent segmentation, and of data with spurious and missing segment boundaries, are shown in Figure 3.

Manual alignment of the various utterances in Figure 3 has identified the cumulative distribution of segment boundaries common to different numbers of utterances of the same sentence (Figure 4).

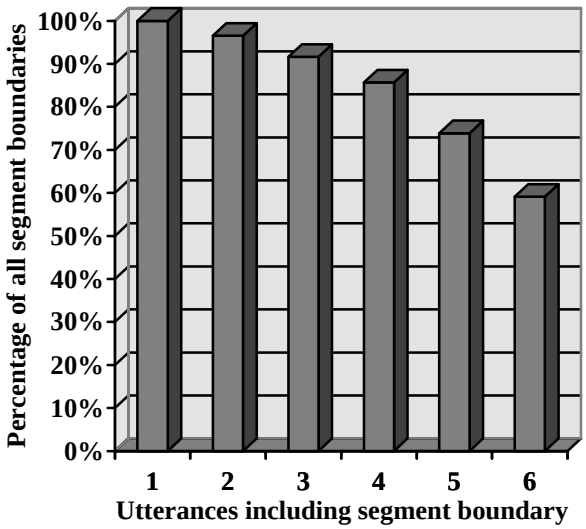

Figure 4: Numbers of data-derived segment boundaries in common between multiple utterances of the same sentence

Figure 4 shows that although over 95% of the detected segment boundaries are common to more than one utterance, it is only possible to rely on 60% of the boundaries (at most) being present in every case. Thus it is essential that any recogniser using this form of data should base its decisions on those boundaries which are common to virtually all instances of the respective words. The recogniser must be able to cope with a large proportion of extraneous boundaries, presumably treating them as 'noise'.

However, it is worth noting that many of the extraneous boundaries actually have a very similar form to their immediate neighbours, so even a recogniser based on conventional HMMs would be able to handle this problem quite well (repeated similar segment boundaries would simply result in extended occupancy of the respective HMM state).

4.CONCLUSIONS

An appropriate choice of MAFI parameters (error threshold and maximum rate of change of position within the vector) can yield segment boundaries which have been found to be similar to the manually labelled phoneme boundaries. However, the targets yielded by automatic segmentation do not always have a one-to-one mapping to conventional phonetic labels. Recognition algorithms which can allow for this feature of the data representation are currently under development.

5. REFERENCES

[1] L. Baghai-Ravary and S. W. Beet, "Multi-step coding of speech parameters for compression". Under review: *IEEE Transactions on Speech and Audio Processing*.

[2] L. Baghai-Ravary, S. W. Beet and M. O. Tokhi, "Adaptive flux interpolation, flow-based prediction, delta or delta-delta coefficients: which is best?", *Proc. Eurospeech '95*, Madrid, pp. 1037-1040, September 1995.

[3] J. N. Holmes, "A variable-frame-rate coding scheme for speech analysis-synthesis systems", *Electronics Letters*, vol. 10, pp. 101-2, 1974.

[4] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database", *Proc. ICASSP-90*, Albuquerque, April 1990.