ACOUSTIC SEGMENTATION USING VARIOUS NEURAL ADAPTATION MODELS

E. Jones, University College Galway, Ireland.
E. Ambikairajah, Regional Technical College, Athlone, Ireland.

ABSTRACT

This paper describes initial results of a comparison of several published models for the inner hair cell/auditory-nerve synapse, for the task of speech segmentation. In each case, the hair cell/synapse model is combined with a model for basilar membrane filtering, and a segmentation algorithm is applied to the neural firing rate in order to emphasize the acoustic boundaries in the speech. The models are tested using utterances from the TIMIT acoustic-phonetic corpus. Performance of each model is assessed by comparing the segmentation it produces with the phonetic transcription provided with the TIMIT database.

1. INTRODUCTION

Segmentation is an important step in the mapping of an acoustic speech signal to a lexicon of word or sub-word units. This is useful in a number of applications, including computer recognition of continuous speech and transcription of speech corpora.

Several researchers have combined the operations of phonetic segmentation and classification into one step. For example, [1] describes a hidden Markov model-based system which classifies a set of 48 phones from the TIMIT database. This system uses a priori knowledge of the phonetic or orthographic content of the utterance to provide phone boundaries. A similar system for phone classification is described in [2].

Another technique which has been used for segmentation is the extraction of a 'boundary function' from a spectral or parametric representation of the utterance. This function encapsulates information about the acoustic boundaries, and can be used either to assist in classification-based segmentation by providing additional temporal information, or as a starting point for a separate stage of classification. For example, in [1], a 'Spectral Variation Function' was extracted from a mel-cepstral representation of the speech to provide additional information about acoustic transitions. Alternatively, in [3], an 'association strength' provided initial potential boundaries, which were subsequently used to generate a multi-level segmentation (a 'dendrogram'), from which final phonetic boundaries were derived using a search procedure.

The boundary function used in [3] was derived from a spectral representation provided by an auditory model [4]. This was found to give better performance, in terms of the least number of boundary insertion and deletion errors, than other representations including discrete Fourier transform and linear predictive coding. One of the reasons for this better performance is the fact that the auditory model used ([5]) included a model for the inner hair cell/auditory-nerve synapse, which exhibits adaptation and recovery, i.e. it enhances sudden onsets and offsets.

While the work described in [4] compared an auditory representation to non-auditory based representations, little work has been carried out to assess the segmentation performance of some other published models for neural adaptation. This issue is addressed in this paper where several adaptation models are combined with a computational model for basilar membrane (BM) mechanics, and used to process continuous speech utterances from the TIMIT database. The representation produced by each model is further processed to produce a time function which contains markers corresponding to potential acoustic boundaries. Performance evaluation is carried out by comparing the auditory-derived boundary markers with those provided with the TIMIT database.

2. THE AUDITORY MODELS

2.1 BM Model

The model for BM mechanics used in this study is based on the transmission line model described in [6] and [7], and consists of a cascade of 128 digital filters covering the frequency range from 70 Hz to 3.4 kHz. The sampling frequency is 8 kHz. The output of each filter in the cascade, corresponding to BM displacement, is used as input to each IHC/synapse model.

2.2 Adaptation Models

The adaptation models chosen for this study were:

- the IHC/synapse model of Seneff [5];
- the Schroeder-Hall reservoir model, as implemented by Cohen [9];
- the adaptation model of Meddis [8];
- an alternative model, based on Meddis's model (Jones at al. [7]).

Since an auditory model could be a useful component of a practical continuous speech recognition system, special emphasis was placed on the adaptation models' relative computational complexity. To this end, it was decided to modify the adaptation models to operate at the same sampling frequency as the BM model, 8 kHz. Both Cohen's model and the alternative model of [7] have particularly simple structures which could be a useful advantage from the point of view of computational load.

As the models operated at 8 kHz, the speech utterances from the TIMIT database, which have a sampling frequency of 16 kHz were decimated by 2. The downsampled speech was processed by each composite BM/adaptation model, with a single neural firing rate vector produced every 5 ms.

3. SEGMENTATION ALGORITHM

The segmentation algorithm applied to the sequence of neural firing rate vectors is based on that described in [4], and operates on the assumption that speech segments can be distinguished from each other by measuring the differences between their spectral representations. The algorithm starts at the beginning of the sequence of spectral vectors and 'associates' each frame with either its past or its future, based on the cumulative distance between that frame and its immediate backward and forward neighbours, over a certain observation range. This observation range was set equal to 50 ms on either side of the current frame [4]. Each frame, n, accumulates forward and backward distances $D_f(n,k)$ and $D_b(n,-k)$, between itself and neighbouring frames k, where $D_f(n,k)$ is defined as follows:

$$D_f(n,k) = \sum_{j=0}^{k} d(n,j)$$

where $d(n,j)$ denotes the Euclidean distance between frame n and frame $n+j$. $D_b(n,-k)$ is defined in a similar manner.

Forward and backward distance measures are accumulated in parallel until either the difference between them exceeds a certain minimum distance, D_{min}, or the observation range is exceeded. An 'association strength', which is the maximum difference between D_f and D_b over the association range, is assigned to each frame. The association strength contour is smoothed using a Gaussian window with a variance of 5 ms [4].

An example of the association strength contour for a portion of the TIMIT utterance "she had your dark suit in greasy wash water all year" is shown in Fig. 1(a). This figure gives the contour derived from Seneff's adaptation model (for clarity, the contours are displayed only as far as the word "suit"). The positive-to-negative crossings of the contours indicate the location of potential acoustic boundaries; this information is converted into a series of pulses, with the height and width of the pulses corresponding to the strength and abruptness of the acoustic change (Fig. 1(b)). A threshold is applied to the pulse sequence, such that pulses below this threshold are set equal to zero. Thus, small pulses which may correspond to false acoustic boundaries are eliminated [4].

The final stage of the segmentation process is the conversion of the pulse sequences of Fig. 1(b) into binary sequences where a '1' indicates the presence of an acoustic boundary at a particular frame, and a '0' indicates no boundary. This allows a straightforward comparison between the boundaries produced by the adaptation models and those obtained from the TIMIT transcriptions. Figure 2 displays such binary sequences for all four adaptation models, where the boundaries provided by the TIMIT transcription are indicated by the pulse train in part (e) of the figure.

4. PERFORMANCE EVALUATION

Initial segmentation parameter choice and performance evaluation was carried out using a subset of the TIMIT database consisting of 40 sentences, with over 1400 boundaries. The binary pulse train produced by each BM/adaptation model combination was compared with the pulse train extracted from the TIMIT transcription and the number of alignments, deletions and insertions was noted. Clearly, the performance of the system as a whole depends as much on the choice of parameters for the segmentation algorithm as it does on the representation produced by the auditory front-end. Since some parameters affect the number of insertions and deletions in different ways, an 'equal error' criterion was used for choosing certain parameters, i.e. the value which gave equal numbers of insertions and deletions was chosen. The numbers of alignments and errors were then used as a measure of relative performance.

Table 1 summarises the relative performance of the models examined, where a tolerance of 25 ms was used for boundary alignment. While the equal error criterion is useful in the current study, the absolute performance of any given model could be improved upon, by considering additional strategies for reducing the number of insertions or deletions, e.g. in a further classification stage it might be possible to recover some deleted boundaries [4], therefore at the segmentation stage a smaller number of insertions can be achieved at the expense of a larger number of deletions.

Table 1. Summary of segmentation performance of all four adaptation models (total errors = sum of insertions and deletions).

Model	Total alignments	Total errors	% Correct
Seneff	1045	842	72.4
Cohen	1112	667	77.1
Meddis	970	962	67.2
Jones et al.	1086	699	75.3

5. DISCUSSION AND CONCLUSIONS

This paper has presented initial results from a comparison between various neural adaptation models, applied to the task of determining acoustic boundaries in continuous speech. The segmentation method used does not make use of any a priori knowledge of the phonetic sequence, it relies solely on the information extracted from the speech by the auditory front-end processor. From experiments with a subset of sentences from the TIMIT database, it would appear that Cohen's model provides slightly better performance than the model of [7], with the other two models a few percent behind. It is interesting to note that the two models which present the lightest computational load give the best performance. Future work will involve validation of these results using a larger database, as well more detailed analysis of the nature of the segmentation errors which occur.

REFERENCES

[1] Brugnara, F., Falavigna, D. & Omologo, M. (1993), "Automatic segmentation and labeling of speech based on hidden Markov models", *Speech Comm.* 12, pp. 357-370.
[2] Ljolje, A. & Riley, M. (1991), "Automatic segmentation and labeling of speech", *Proc. ICASSP*, pp. 473-476.
[3] Glass, J. & Zue, V. (1988), "Multi-level acoustic segmentation of continuous speech", *Proc. ICASSP*, pp. 429-432.
[4] Glass, J. & Zue, V. (1986), "Signal representation for acoustic segmentation", *Proc. SST-86*, pp. 124-129.
[5] Seneff, S. (1988), "A joint synchrony/mean-rate model of auditory speech processing", *J. of Phonetics* 16, pp. 55-76.
[6] Ambikairajah, E. Black, N. & Linggard, R. (1989), "Digital filter simulation of the basilar membrane", *Comp. Speech & Lang.* 3, pp. 105-118.
[7] Jones, E. & Ambikairajah, E. (1993), "Comparison of various adaptation mechanisms in an auditory model for the purpose of speech processing", *Proc. Eurospeech '93*, pp. 717-720.
[8] Meddis, R. (1986), "Simulation of mechanical to neural transduction in the auditory receptor", *J. Acoust. Soc. Am.* 79, pp. 702-711.
[9] Cohen, J. R. (1989), "Application of an auditory model to speech recognition", *J. Acoust. Soc. Am.* 85, pp. 2623-2629.

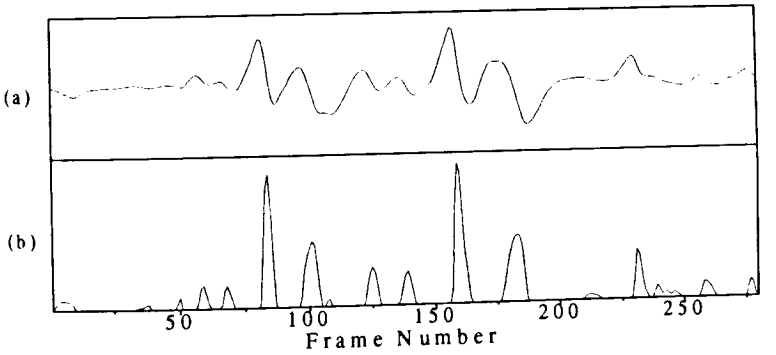

Figure 1. (a) Association strength contour, and (b) boundary pulse sequence for the TIMIT utterance fragment "She had you dark suit" obtained using Seneff's adaptation model.

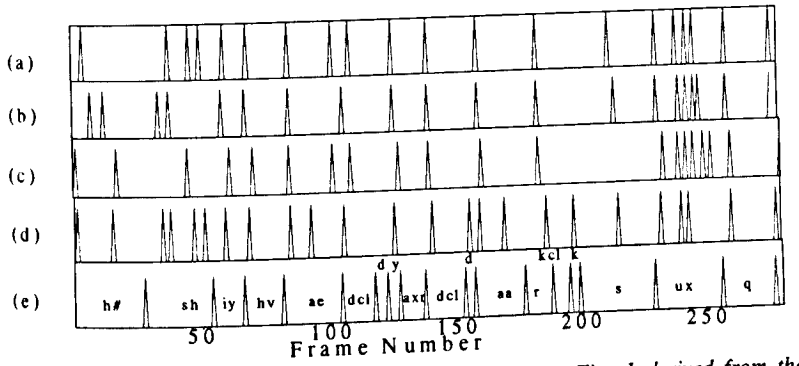

Figure 2. Boundary markers for the same utterance as in Fig. 1 derived from the following adaptation models: (a) Seneff (b) Cohen (c) Meddis, and (d) Jones et al. [7]. The actual boundaries derived from the TIMIT transcription are given in part (e), with phonetic labels.