Proceedings of The Institute of Acoustics

TOWARDS AN EARLY SYMBOLIC REPRESENTATION OF SPEECH BASED ON AUDITORY MODELLING

M. P. Cooke

Department of Computer Science, The University of Sheffield,
Sheffield S10 2TN

INTRODUCTION

Recent years have seen increasing interest in attempts to model the mammalian auditory periphery. In the automatic speech recognition community, this interest is stimulated by the hope that auditory modelling will provide the clarity and completeness of signal analysis necessary to support linguistically based recognition strategies. The physiological results on which AMs are based have been around for some time, and reported models have many processing stages in common. What is needed now is an examination of how best to represent the 'neural-analogue' data derived from such models within a speech recognition scheme.

AM results are most commonly presented in a form that is compatible with the results of conventional speech analysis techniques: in uniformly quantised time-frequency intervals. For instance, Seneff's pseudo-spectrogram [1] and Lyon's VQ Cochleagram [2], are motivated by a desire to compare results with those gained by Fourier analysis and LPC in order to allow an early test of AM performance. This seems to us to be an unnatural way to proceed since it constitutes a form of smoothing and, further, requires that a complete description of activity across the whole spectrum at each time quantum is carried forward. Such a representation contains much redundancy.

Alternatively, Kohonen [3] proposes a connectionist schema which could be adapted so that AM outputs provide the continuously varying input activation levels. While this avoids the problems inherent in trying to find a succinct representation of the auditory data, it limits the application of knowledge to the descriptive process.

In contrast, we suggest here an approach based on intermediate representations derived using auditory grouping principles. The next section expands on this theme.

A REPRESENTATIONAL FRAMEWORK FOR AUDITORY PROCESSING

Although little is known about the detailed way in which central auditory processes organise and utilise data emanating from the PAS, work by Bregman [5] and others on auditory grouping forces suggests the form of processing performed by the higher levels of the auditory system. A currently popular view is that perception of acoustic sources involves the creation of auditory streams, and that the ongoing interpretation of the signal is made with reference to these. Bregman suggests that such streams correspond to objects in vision.

So far, the factors which affect streaming have been studied with relatively simple stimuli. The underlying grouping forces which such studies reveal will of course apply equally to the perception of natural speech signals. An exciting possibility is that auditory streams represent a meeting point between phonetic interpretation and acoustic evidence, just as object-centred representations have a critical role in vision.

Our long term goal is to provide a computational test of the usefulness of these ideas in ASR. However, it is clear that several processing stages intervene between auditory-nerve data and auditory streams. We need a good early description of the auditory scene, preferably in a symbolic (*1) form. This suggests the use of a representational framework, as an auditory counterpart to the 'speech-sketch' proposed by Green & Wood [6].

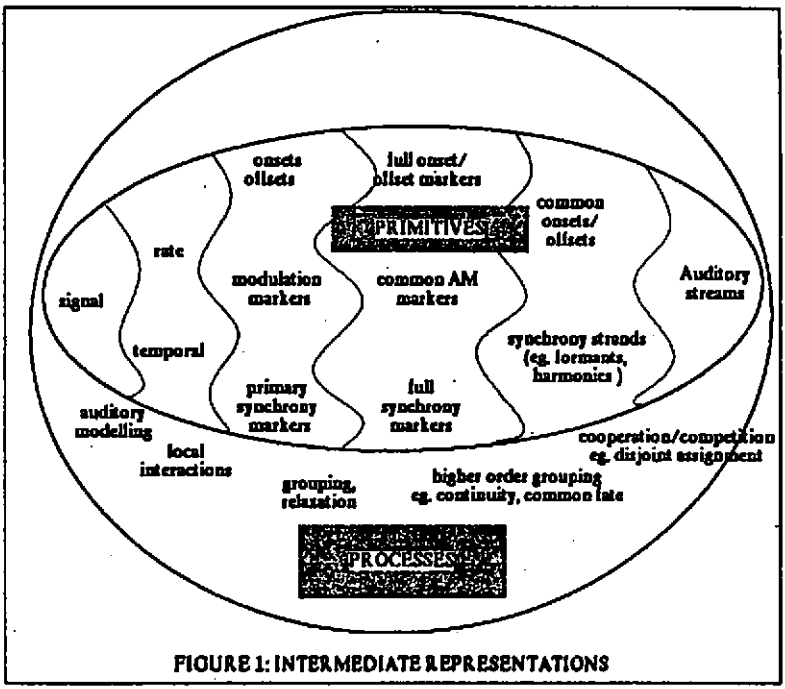

FIGURE 1: INTERMEDIATE REPRESENTATIONS

Figure 1 is a tentative first step at defining what the intermediate levels of representation should be. It is too early to be sure what primitives will be involved at each stage, but some suggestions can be made based on psychoacoustic studies (Hall & Haggard [7]; Pastore [8]). Working from the

(*1) We use the term 'symbolic' in the sense employed in vision research rather than in the more restricted sense of segmentation and labelling [23].

left of the figure, auditory modelling provides a dual rate/temporal characterisation of the signal. Various low-level descriptions can be obtained from this representation. We envisage these including: onsets/offsets in individual channels; some indication of phase-locking; a description of the amplitude modulation seen in HF channels. These primitives can be obtained via local operations on the rate/temporal data. The next two stages involve the detection of properties which extend across a range of frequencies and exist for larger time intervals. Examples are shown in the figure. Progressively higher order grouping forces produce, for instance, regions of synchrony continuity; these in turn combine via common fate, creating harmonic sets. Finally, rules of competition operate to provide a coherent, ongoing interpretation of the evidence - the auditory stream. The right hand part of this figure owes much to ideas contained in Bregman [5].

The following sections describe our progress in implementing this model up to and including the first truly symbolic stage, represented by the central part of the figure.

A MODEL OF THE AUDITORY PERIPHERY

This model is fully described in Cooke [9] and is briefly reviewed here. The overall architecture of the model comes from a consideration of cochlear filtering and detection at places along the basilar membrane (BM). The model consists of a number (usually 61) of independent channels spaced equally along a bark scale - 0.25 bark intervals from 1 bark to 16 bark provides a good coverage of the frequency range 100-5000 Hz.

The first stage of analysis is the bandpass-nonlinear (BPNL) model of BM filtering action and two-tone suppression, originally suggested by Pfeiffer [10]. This scheme consists of a static nonlinearity sandwiched between two bandpass filters. It should be stressed that such a model is no longer a valid physiological description of cochlear filtering, but serves to produce similar gross mechanical and suppressive effects to those observed by physiologists. Very recent research has demonstrated clearly that BM tuning curves are as sharp as those observed in auditory-nerve fibres and that BM action is highly nonlinear. Furthermore, the true role of other structures in the cochlear partition (particularly outer hair cells) is only just becoming clear. For a review of these radical findings, see [11]. It is too early to say how these developments will be reflected in auditory models.

The next process in the computer model is designed to reflect the hair-cell to nerve fibre signal transformation. It essentially consists of a three-partition model of the dynamic firing rate characteristics of auditory-nerve fibres. Besides the usual properties of increased firing rate at stimulus onset, decay of the rate with sustained stimulation and reduction to below the spontaneous rate at stimulus offset (all of which can be produced by simple models of neurotransmitter release, eg. Schroeder & Hall, [12]), the model is able to produce additive adaptation effects (Smith & Zwislocki, [13]) without over-complex processing.

RATE & TEMPORAL CHARACTERISATIONS OF MODEL OUTPUTS

The separation of model outputs into rate and temporal measures reflects a belief in the parsimony of central auditory processing. Rate information comes directly from the adaptation stage. We have yet to utilise this measure in the manner suggested by the representational framework of Figure 1, but an eventual goal is to generate onset/offset markers which would include the attributes of onset discharge rate, range of channels in which the onset occurred, and rise time. Furthermore, the attempt to characterise common amplitude modulation in the HF range will undoubtedly rest upon a suitable transformation of rate information.

Temporal information is extracted from the signal arising from the BPNL stage. Phase-locking and hair-cell rectification indicate that timing information could be extracted from the extrema of BPNL outputs. Other possibilities exist. For instance, an instantaneous frequency measure provides a continuous estimate of the dominant component to which a particular channel is responding. An approximation to instantaneous frequency may be obtained by inverting the interval between successive zero-crossing times of the BPNL stage output, and multiplying by half the sampling frequency.

The rationale for using a frequency estimate based on zero-crossings is as follows.

1. This measure, though not continuous, is more compatible with the detection mechanisms available to the auditory system, given that fibres generally discharge at most once during a single cycle of a stimulus component.

2. Since auditory-nerve fibres show a distinct preference to fire at a particular phase of the stimulating waveform, it is of value to identify and localise a single point per cycle. Zero-crossings are both easy to find and their position may be more accurately determined than peaks.

3. Characterising zero-crossing intervals rather than locations bypasses the potential problems of local comparisons caused by absolute phase differences between neighbouring channels.

4. Work by Schofield [14] indicates that the frequency estimates derived by this method lie very close to the continuous instantaneous frequency measure.

Carlson & Granstrom [15] suggested the use of such a measure in their DOMIN method of auditory temporal analysis, and a similar form of processing has been used by Neiderjohn & Lahat [16] for formant analysis in noisy signals.

Figure 2 shows the frequency estimates derived from zero-crossing intervals for

part of the phrase "hello operator" uttered by a male speaker (#2). A single dot is plotted for each frequency estimate. Here, phase-locked responses are visible due to the clustering of similar estimates derived from a number of channels. Some characteristic properties of auditory responses (eg. Delgutte and Kiang [17]) have correlates in this representation. Note 1) synchronisation of low frequency channels to harmonics of the fundamental; 2) synchronisation to dominant stimulus components (formants) in the mid frequency regions, and 3) loss of synchrony at high frequencies. Further, this figure is devoid of rate information - silent intervals will show up as a random dot texture.

DERIVING A SYMBOLIC REPRESENTATION

The goal of the processes operating at this stage is to derive a representation in which certain intrinsic properties of the data are made explicit. Though no definite proposals have been made to date, psychoacoustic considerations suggest that the important features of the auditory response to speech include synchronisation to dominant components (Carlson, Fant & Granstrom [18]), onsets and offsets within broad frequency regions (Summerfield & Assmann [19]), and common amplitude modulation (Hall & Haggard [7]).

Making synchrony explicit

The goal of this process is to describe the patterns of phase-locking present in neighbouring groups of channels, and thereby retrieve an indication of the stimulus component to which the responses are synchronised. The approach outlined can be contrasted with those proposed by Seneff [20] and Ghitza [21]. We see this as a two-stage algorithm, generating in turn primary and full synchrony markers.

Primary synchrony markers (PSMs)

The first stage is an attempt to disentangle the excitation effects present in the initial temporal characterisation of the auditory model outputs. At a fine level of detail, the temporal response within a single channel shows effects which may be attributed to the open and closed phases of the larynx cycle. In other words, our initial characterisation of phase-locking is subject to periodic disruptions. However, we can exploit the fact that whole groups of channels will be synchronised to the same component, and furthermore, that most within channel estimates will be consistent. In this case, local support is sought in the 3-neighbourhood of frequency estimates. For channel i, the 3-neighbourhood consists of previous estimates in channels i-1, i and i+1 (with obvious modifications for the first and last channels). Each frequency estimate is compared with those in its 3-neighbourhood, and a primary synchrony marker is produced only if the estimates are sufficiently close. Thus, primary synchrony markers are frequency estimates for which there is local support.

(#2) This data is part of the Stockholm fragment as used by the Alvey Speech Club

Figure 3 shows the PSMs generated by applying quite a tight criterion of similarity to within 0.15 Bark, which is a little less than typical formant bandwidths. Clearly, the PSMs generated correspond largely to synchronised regions. Note that very little information above F3 is preserved due to loss of synchrony.

Full synchrony markers

Primary synchrony markers provide locally consistent estimates of the component to which the response in individual channels is phase-locked. However, this response synchrony is exhibited over extended time intervals, and across several channels. The intention of the second stage is to describe these regions.

To start with, we take the view that processing at this early level should remain local. At the same time, the algorithm for determining the range of synchronised responses must be able to operate across a number of channels. These criteria suggest an iterative relaxation process [22], in which decisions regarding synchrony are propagated amongst neighbouring channels on each iteration.

There are a number of reasons why a direct application of relaxation is not straightforward. The major one concerns when to apply the iteration. Since HF channels will generate far more PSMs than LF ones, it may be inappropriate to perform relaxation labelling at uniform time intervals. Similarly, it would be computationally infeasible to apply full relaxation whenever a new PSM arrives. In any case, a new PSM from a HF channel will not affect the labelling at mid and low frequency channels.

Our solution is to let time take the place of iteration in the relaxation labelling process. In essence, the algorithm determines an assignment of labels (Full Synchrony Markers) to processing channels such that channels with similar synchrony gets the same label. Obviously, channels will form dynamic coalitions. For instance, if a formant rises through a particular region, the set of channels phase-locked to its frequency will be different at the start and ends of the transition. Whenever a new PSM arrives, it has the chance to influence the labelling in its neighbourhood. As PSMs become available in a range of channels, their influence propagates across the channels.

The results of applying this process are shown in Figure 4. This represents a significant data reduction (eg. a single FSM can represent over 100 PSMs) over the previous data, and we see it as a viable candidate component of the speech sketch.

ACKNOWLEDGEMENT

This work was supported by ALVEY grant MMI 052.

REFERENCES

[1] S. Seneff, 'Characterising formants through straight-line approximations without explicit tracking', Proc. ICA Symposium on Speech Recognition, Montreal, (1986).

[2] R. Lyon, 'Speech recognition experiments with a cochlea model', Proc. ICA Symposium on Speech Recognition, Montreal, (1986).
[3] T. Kohonen, Self-Organisation and Associative Memory, Springer-Verlag, (1984).
[4] D. Marr, Vision, W. H. Freeman, (1982).
[5] A. S. Bregman, 'Auditory scene analysis', Proc. 7th. Int. Conf. Pattern Recognition, Montreal, 168-175, (1984).
[6] P. D. Green and A. R. Wood 'A representational approach to knowledge-based acoustic-phonetic processing in speech recognition', Proc. ICASSP, 23.4, (1986).
[7] J. W. Hall and M. Haggard, 'Co-modulation - A principle for auditory pattern analysis in speech', Proc. 11th ICA, 69-71, (1983).
[8] R. E. Pastore, 'Possible psychoacoustic factors in speech perception', in: Perspectives on the study of speech, P. D. Eimas and J. L. Miller (eds), LEA, (1981).
[9] M. P. Cooke, 'A computer model of peripheral auditory processing incorporating phase-locking, suppression and adaptation effects', Speech Communication, 5, (1986).
[10] R. R. Pfeiffer, 'A model for two-tone inhibition of single cochlear nerve fibres', JASA, Vol. 48, No. 6, 1373-1378, (1970).
[11] Hearing Research, Vol. 22, (1986).
[12] M. R. Schroeder and J. L. Hall, 'Model for mechanical to neural transduction in the auditory receptor', JASA, Vol. 55, No. 5, 1055-1060, (1974).
[13] R. L. Smith and J. J. Zwislocki, 'Short-term adaptation and incremental responses of single auditory-nerve fibres', Biol. Cybern., vol. 17, 169-182, (1975).
[14] D. Schofield, 'Visualisations of speech based on a model of the peripheral auditory system', NPL Report 62/85, HMSO, (1985).
[15] R. Carlson and B. Granstrom, 'Towards an auditory spectrograph', in: The representation of speech in the peripheral auditory system, R. Carlson and B. Granstrom (eds), Elsevier, (1982).
[16] R. J. Niederjohn and M. Lahat, 'A zero crossing consistency method for formant tracking of voiced speech in high noise levels', IEEE Trans. Acoust., Speech and Sig. Proc., Vol. ASSP-33, No. 2, 349-355, (1985).
[17] B. Delgutte and N. Y. S. Kiang, 'Speech coding in the auditory-nerve: I. Vowel-like sounds', JASA, Vol. 75, No. 3, 866-878, (1984).
[18] R. Carlson, G. Fant and B. Granstrom, 'Two-formant models, pitch and vowel perception', in: Auditory analysis and perception of speech, G. Fant and M. A. A. Tatham (reds), Ac. Press, (1975).
[19] A. Q. Summerfield and P. Assman, 'Auditory enhancement in speech perception', Nato workshop: 'The psychophysics of speech perception', Utrecht, (1986).
[20] S. Seneff, 'Pitch and spectral estimation of speech based on auditory synchrony model', Proc. ICASSP, 36.2, (1984).
[21] O. Ghitza, 'Speech analysis/synthesis based on matching the synthesised and the original representation in the auditory nerve level', Proc. ICASSP, 37.11, (1986).
[22] L. S. Davis and A. Rosenfeld, 'Cooperating processes for low-level vision: A survey', Artificial Intelligence, Vol. 17, 245-263, (1981).
[23] J. S. Bridle and R. K. Moore, 'Boltzmann machines for speech pattern processing', Proc. IOA, (1984).

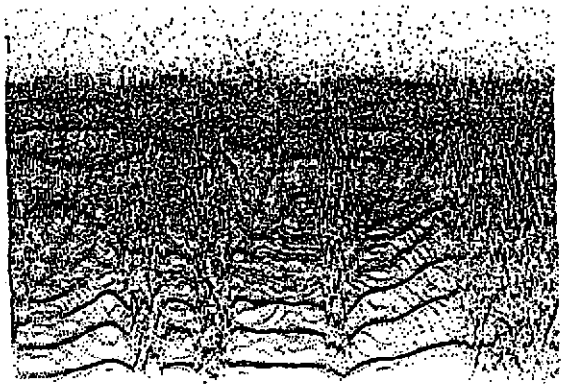

FIGURE 2: Zero crossing frequency estimates. A single dot is plotted for each estimate. Data: part of the utterance "hello operator". Horiz. extent: 1 second. Vert. extent: 20 Barks.

FIGURE 3: Primary synchrony markers generated from the data of fig. 2. A neighbourhood similarity criterion of 0.15 Barks was used.

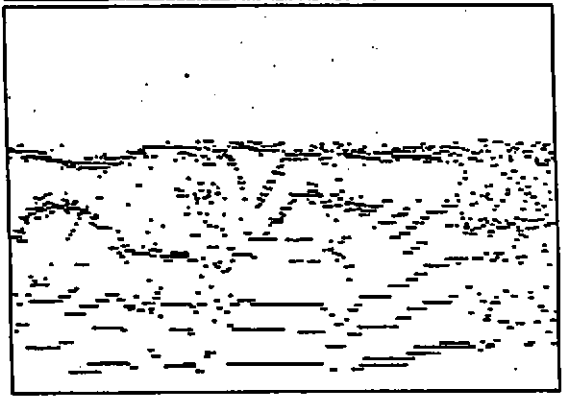

FIGURE 4: Full synchrony markers derived via a discrete relaxation process using the primary synchrony markers of fig. 3.